BEV^{2}3D: BEV-BASED EVENT-GUIDED 3D DETECTION FOR LIDAR BLIND INTERVALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern 3D object detectors for autonomous driving often operate at the rate of their slowest sensor, leaving 50–100 ms LiDAR blind intervals without updated detections. To update detections within these intervals, Ev-3DOD introduced event-based blind-time perception through an RoI-based formulation that updates individual detections (Cho et al., 2025). We instead introduce BEV^{2}3D, a framework named after the two "EV"s in B**EV** and **EV**ent, which reformulates blind-interval perception as event-guided BEV updating. BEV^{2}3D couples a frozen LiDAR reference detector with an event-driven branch that updates a sparse bird's-eye-view (BEV) representation and re-estimates detections throughout each LiDAR blind interval. Within this event-driven branch, Frustum Event Lifting (FEL) lifts event features with an inference-time LiDAR depth prior, while Support-Gated Temporal Fusion (SGTF) combines fresh event evidence with cached LiDAR features. On DSEC-3DOD and Ev-Waymo, BEV^{2}3D outperforms Ev-3DOD in every evaluated category by up to 8.6 and 11.3 AP, respectively, while relying only on LiDAR and events, without the RGB modality used by Ev-3DOD. Our code is included in the supplementary material and will be publicly released upon acceptance.

1 INTRODUCTION

Modern 3D detectors for autonomous driving typically rely on temporally aligned inputs from conventional sensors, such as LiDAR and cameras, and often update at the rate of their slowest sensor. For pipelines operating at a LiDAR frame rate of 10–20 Hz, the gap between consecutive sweeps forms a 50–100 ms *blind interval* without updated detections, as illustrated in Figure 1(a). This interval is long enough for a vehicle traveling at 30 m/s to move by up to 3 m. Such blind intervals motivate complementary high-rate sensing.

Event cameras are well matched to this gap. Each pixel asynchronously reports log-intensity changes that exceed a contrast threshold, providing a continuous stream of visual evidence with microsecond temporal resolution (Lichtsteiner et al., 2008; Gallego et al., 2020). Ev-3DOD (Cho et al., 2025) recently established that events captured during a blind interval can support 3D detection. It estimates a rigid motion for each object detected at the last active frame and propagates that object's box forward. This was the first demonstration that the blind interval need not be silent.

Ev-3DOD formulates blind-time perception as RoI-based *object motion estimation*. For each active-frame proposal, it pools event and LiDAR features within the RoI, predicts a 4-DoF motion comprising 3D translation and yaw change, and applies this motion to the corresponding box. Its blind-time predictions therefore remain tied to the active-frame proposals, as shown in Figure 1(b).

Our reformulation. We instead retain event evidence in a shared BEV representation and fuse it with cached LiDAR features. At each intermediate query timestamp, a detection head re-estimates object locations, geometry, and confidence from the fused BEV, as illustrated in Figure 1(c).

Our BEV^{2}3D instantiates this formulation with two detectors that run at different rates, as shown in Figure 2. The frozen *reference detector* is a standard LiDAR detector based on VoxelNeXt (Chen et al., 2023). It runs at the LiDAR frame rate and caches voxel features and detections at each active timestamp. A trainable *asynchronous detector* runs at every intermediate timestamp of the blind interval. It encodes accumulated events and updates the cached BEV in two stages.

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.

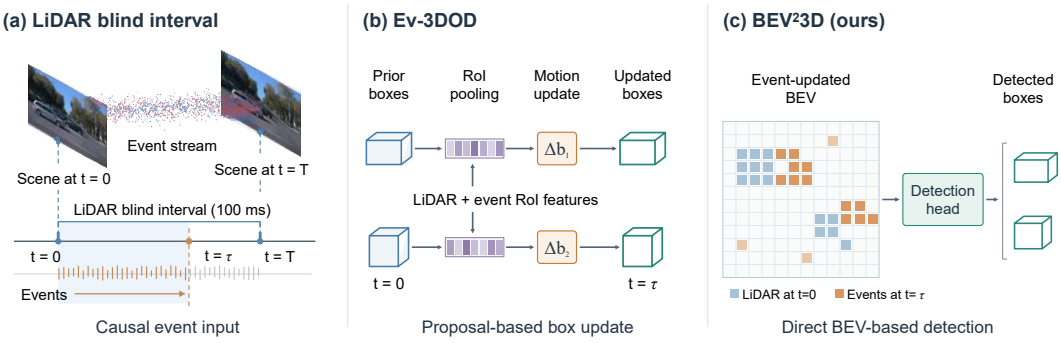


Figure 1: **Blind-time perception, two formulations.** **(a)** A blind interval lies between synchronized LiDAR sweeps, and each query uses only events observed since the latest LiDAR frame. **(b)** Ev-3DOD (Cho et al., 2025) updates prior boxes through per-box motion estimation. **(c)** BEV^{2}3D lifts events into a spatial BEV map and runs a BEV detection head at intermediate timestamps.

The first stage addresses the temporal mismatch between cached LiDAR geometry and current events. Because object motion can make the cached depth inconsistent with current event evidence, Frustum Event Lifting (FEL) treats LiDAR depth as a prior rather than a fixed assignment, allowing event evidence to revise the depth weighting before splatting the resulting features into a sparse bird's-eye-view (BEV) map. The second stage, Support-Gated Temporal Fusion (SGTF), fuses the event BEV with the cached LiDAR features, using keyframe box supports and query time to control the event contribution. Finally, a detection head operates on the fused BEV to produce updated detections at the query timestamp.

During training, only the asynchronous detector is optimized. It uses cached features from the frozen reference detector and a depth prior obtained by projecting the latest LiDAR point cloud onto the event camera's image plane. Experiments on DSEC-3DOD (Gehrig et al., 2021; Cho et al., 2025) and Ev-Waymo (Cho et al., 2025; Sun et al., 2020) show that BEV^{2}3D substantially and consistently outperforms Ev-3DOD in every category.

Contributions. We summarize our contributions as follows.

- We reformulate blind-time perception by retaining a spatial BEV until prediction rather than reducing event evidence to per-object motion updates (§3).
- We propose BEV^{2}3D, a dual-rate architecture that couples a frozen LiDAR reference detector with a trainable asynchronous event detector to produce updated detections throughout each LiDAR blind interval.
- We introduce FEL and SGTF. FEL lifts event features through a depth-weighted frustum into BEV, guided by an inference-time LiDAR depth prior, while SGTF fuses the resulting event BEV with cached LiDAR BEV features.
- On DSEC-3DOD (Gehrig et al., 2021) and Ev-Waymo, BEV^{2}3D improves over Ev-3DOD across all evaluated categories by up to 8.6 AP and 11.3 AP, respectively, while using LiDAR and events without Ev-3DOD's RGB modality (§4).

2 Related Work

Event-based blind-time detection. Event cameras report asynchronous brightness changes (Lichtsteiner et al., 2008; Gallego et al., 2020) and have supported optical flow, reconstruction, and 2D detection (Zhu et al., 2018; Rebecq et al., 2019; Gehrig & Scaramuzza, 2023). In blind-time 3D detection, Ev-3DOD (Cho et al., 2025) uses events to estimate a 4-DoF translation-and-yaw update for each box detected at the preceding LiDAR keyframe. Kang et al. (2025) instead develop an event-only stereo system for continuous-time 3D detection, removing the LiDAR keyframe altogether. BEV^{2}3D addresses the complementary keyframe-prior regime. It retains reliable LiDAR geometry from the latest keyframe and uses events causally within the ensuing interval to update a BEV representation

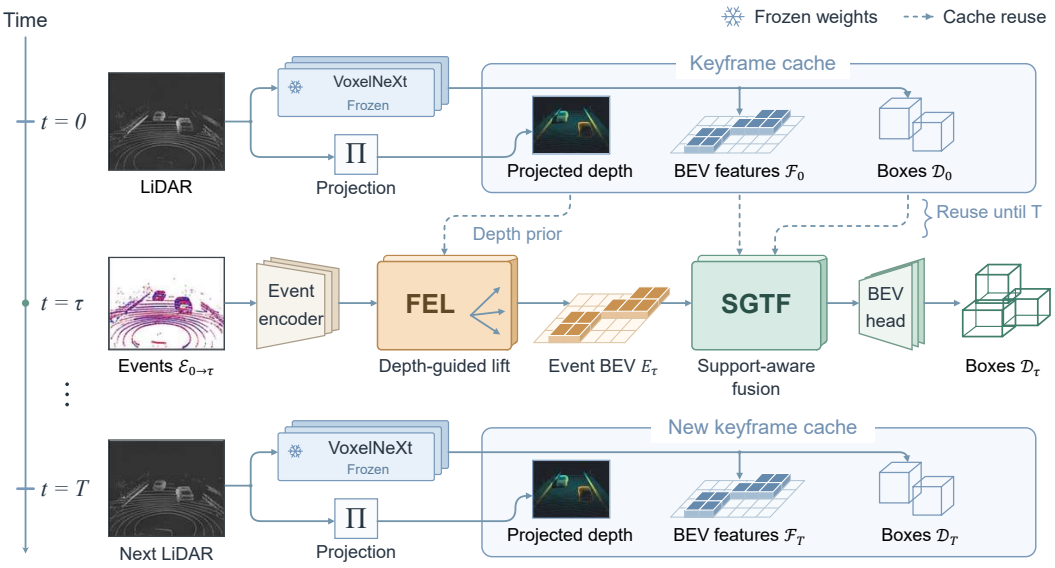


Figure 2: **Overview of BEV^{2}3D.** At the latest active frame (t=0), the frozen VoxelNeXt reference detector extracts LiDAR BEV features $\mathcal{F}_0$ and detections $\mathcal{D}_0$. The active LiDAR scan is also projected into the event camera to provide the depth prior used by FEL. At each query time τ, the trainable asynchronous detector encodes the accumulated event stream $\mathcal{E}_{0\to\tau}$. FEL lifts it into a sparse event BEV map E_τ, and SGTF fuses E_τ with the cached LiDAR BEV features before the BEV head predicts $\mathcal{D}_\tau$. The next LiDAR frame at t=T refreshes the cache for the following blind interval.

before detection. This also differs from streaming and tracking-by-detection methods that propagate outputs or state directly (Li et al., 2020).

BEV-based 3D object detection. Bird's-eye-view (BEV) representations provide a common spatial interface for 3D detection. LiDAR detectors construct BEV features from sparse voxels (Yan et al., 2018; Chen et al., 2023; 2022; Zhang et al., 2023a) or point pillars (Lang et al., 2019; Yin et al., 2021). Camera-based detectors exploit BEV representations through explicit view transformation (Philion & Fidler, 2020; Huang et al., 2021), depth-supervised lifting (Li et al., 2023b), or query-based modeling in BEV space (Li et al., 2024; Yang et al., 2023; Liu et al., 2023a). We use BEV as the spatial interface for bringing event evidence into cached LiDAR features between keyframes. FEL forms an event BEV representation from a depth-weighted frustum using an inference-time soft depth prior derived from cached LiDAR points.

Multi-modal 3D object detection. LiDAR–camera detectors fuse complementary geometry and appearance through unified BEV representations (Liu et al., 2023b) or transformer-based fusion (Bai et al., 2022; Li et al., 2023a). These methods rely on temporally aligned sensor inputs and typically produce detections at LiDAR keyframes. In contrast, BEV^{2}3D uses LiDAR only at keyframes and adds an event-driven branch between keyframes. It thus retains a LiDAR-derived spatial prior while producing causal intermediate-time detections without an RGB input at inference.

3 Method

3.1 The Blind-Time Detection Problem

Let t=0 denote the latest active frame and τ an intermediate query time within the ensuing blind interval. At time τ, no synchronized LiDAR frame is available. The asynchronous detector can use only information from the latest active frame and events accumulated up to τ. Its goal is to produce fresh detections $\mathcal{D}_\tau$ at each query time using only these causal inputs.

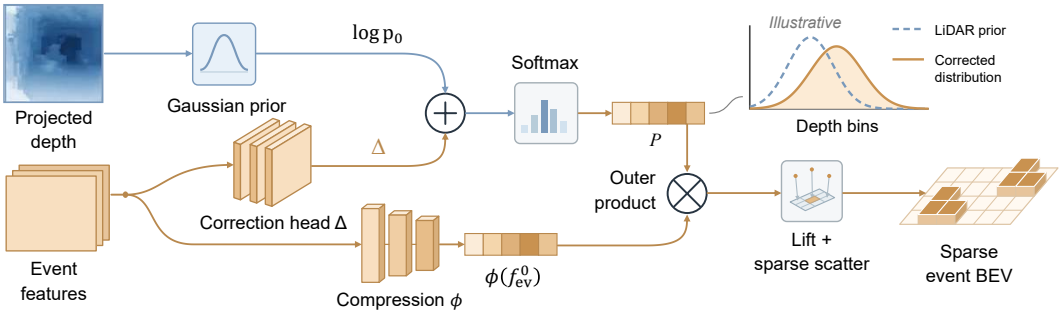


Figure 3: **Frustum Event Lifting (FEL).** FEL uses projected LiDAR depth to provide a soft prior for placing event features in 3D. The event-conditioned correction Δ adjusts this prior over depth bins, so cached geometry guides rather than fixes the event-to-BEV lift. The corrected depth distribution and compressed event features are combined along camera rays and sparsely scattered into E_τ. The depth distributions are illustrative.

3.2 Reference Detector: Frozen LiDAR Backbone

The reference detector is a pre-trained VoxelNeXt (Chen et al., 2023) model that remains frozen during training. Given the active-frame LiDAR point cloud $\mathcal{P}_0$, it produces

$$\mathcal{F}_0,\ \mathcal{D}_0 = g_{\text{ref}}(\mathcal{P}_0), \tag{1}$$

where $\mathcal{F}_0$ denotes the sparse LiDAR BEV features immediately before the detection head and $\mathcal{D}_0$ the final 3D detections. Both are cached for use by the asynchronous detector. Freezing g_{ref} preserves its active-frame behavior and confines optimization to the asynchronous branch.

3.3 Asynchronous Detector

The asynchronous detector f_{async} (Figure 2) encodes the event stream $\mathcal{E}_{0\to\tau}$ accumulated up to query time τ with a 2D backbone (§3.3.1). The resulting event features are processed by Frustum Event Lifting (FEL; §3.3.2) and then fused with the cached LiDAR BEV features by Support-Gated Temporal Fusion (SGTF; §3.3.3), followed by a detection head that predicts $\mathcal{D}_\tau$ (§3.3.4).

3.3.1 Event Representation and Backbone

Following Ev-3DOD (Cho et al., 2025), we voxelize the accumulated events using the event-volume representation (Zhu et al., 2019). Each event contributes its signed polarity to neighboring spatio-temporal bins through trilinear interpolation, yielding an event volume with five temporal bins. This preserves the spatio-temporal structure of the event stream while remaining compatible with standard 2D backbones. We use a Swin-T backbone (Liu et al., 2021) with an FPN (Lin et al., 2017a) to extract multi-scale event features,

$$\{f_{\text{ev}}^{\ell}\}_{\ell=0}^{L-1} = \Phi(\text{Vox}(\mathcal{E}_{0\to\tau}))\,, \tag{2}$$

Here, Vox denotes event voxelization and Φ the Swin-T backbone with FPN. We use $L{=}4$, and $f_{\text{ev}}^{\ell} \in \mathbb{R}^{C_\ell \times H_\ell \times W_\ell}$ denotes the event feature at pyramid level ℓ.

3.3.2 Frustum Event Lifting (FEL)

FEL constructs a sparse event BEV map E_τ from the highest-resolution FPN feature map f_{ev}^0 and the cached LiDAR point cloud $\mathcal{P}_0$. Because the cached LiDAR geometry and current events correspond to different timestamps, object motion can make the projected LiDAR depth inconsistent with current event evidence. FEL therefore uses the projected LiDAR depth as a soft prior and lets event features revise the depth distribution along each camera ray, as summarized in Figure 3.

Soft depth prior. For each pixel with a projected LiDAR return, we construct a categorical distribution over D fixed depth bins using a Gaussian centered at the projected depth $d_{\text{lidar}}(u,v)$:

$$p_0(u,v,d) \ \propto\ \exp\!\Big(-\tfrac{(d_{\text{bin}}(d)-d_{\text{lidar}}(u,v))^2}{2\sigma^2}\Big), \quad \textstyle\sum_d p_0(u,v,d) = 1, \tag{3}$$

where $d_{\text{bin}}(d)$ is the center of bin d and σ is a kernel width. Pixels without a projected return receive a uniform categorical prior. The Gaussian assigns nonzero mass to neighboring bins, so the LiDAR-projected depth serves as a soft prior rather than a fixed depth assignment.

Prior-conditioned depth assignment. FEL combines the LiDAR depth prior with event evidence through a Bayesian-inspired update. A learned event-dependent correction Δ is added to the log-prior before normalization:

$$\begin{aligned} z(u,v,d \,|\, \mathcal{E}_{0\to\tau}) &= \underbrace{\log p_0(u,v,d)}_{\text{LiDAR log-prior}} + \underbrace{\Delta\big(f^0_{\text{ev}}(u,v)\big)_d}_{\text{learned correction}}, \\ P(u,v,\cdot \,|\, \mathcal{E}_{0\to\tau}) &= \text{softmax}_d\big(z(u,v,\cdot \,|\, \mathcal{E}_{0\to\tau})\big), \end{aligned} \tag{4}$$

Equivalently, the depth probabilities are proportional to $p_0 \exp(\Delta)$, with the learned correction reweighting the LiDAR prior. The final layer of Δ is zero-initialized, so the initial distribution reproduces the LiDAR prior exactly. Training then learns an event-conditioned correction that can revise stale keyframe depth assignments.

Frustum scatter to BEV. Using the predicted depth distribution and event-camera calibration, each pixel feature is projected along its camera ray across all depth bins to form a frustum volume. We scatter the ray–bin candidates into BEV cells and sum their weighted features:

$$E_\tau(y,x) \;=\; \sum_{u,v} \sum_{d} P(u,v,d)\, \phi\big(f^0_{\text{ev}}(u,v)\big)\; \mathbb{I}\big[(u,v,d) \to (y,x)\big], \tag{5}$$

where ϕ is a feature-compression layer and $\mathbb{I}[(u,v,d) \to (y,x)]$ indicates whether the ray–bin candidate (u,v,d) projects to BEV cell (y,x) under the same calibration. FEL retains ray–bin candidates whose projected (x,y) positions lie within the BEV range and quantizes them to grid cells. The resulting unique BEV coordinates define the indices of the 2D sparse convolution tensor E_τ, which is refined by two sparse 2D convolution layers before fusion.

The property that distinguishes E_τ from a per-object motion update is its spatial structure. Each activated BEV location carries a feature vector summarizing local event evidence, rather than contributing only to a proposal-wise feature used to regress a single box update.

3.3.3 Support-Gated Temporal Fusion (SGTF)

Once E_τ is built, SGTF fuses it with the cached LiDAR BEV features $\mathcal{F}_0$. It uses a keyframe box-support mask to modulate the additional event contribution at each sparse BEV location.

Let Ω_0 and Ω^E_τ denote the active coordinate sets of the cached LiDAR BEV tensor and E_τ, respectively. SGTF operates on their union,

$$\Omega_\tau = \Omega_0 \cup \Omega^E_\tau. \tag{6}$$

For every $i \in \Omega_\tau$, let $\mathbf{s}_i$ and $\mathbf{e}_i$ denote the cached LiDAR and event features, respectively. They are zero-extended outside their original supports, so $\mathbf{s}_i = \mathbf{0}$ for $i \notin \Omega_0$ and $\mathbf{e}_i = \mathbf{0}$ for $i \notin \Omega^E_\tau$. The fused feature is

$$\mathbf{f}_i \;=\; \mathbf{s}_i \;+\; \gamma_i\, \mathbf{e}_i, \qquad \gamma_i \;=\; \gamma_{\text{bg}} \;+\; m_i\, \alpha(\rho), \tag{7}$$

where $m_i \in \{0,1\}$ is the box-support mask and $\alpha(\rho)$ is the support-conditioned gain. A LiDAR-only coordinate retains $\mathbf{s}_i$. An event-only coordinate enters the fused tensor as $\gamma_i \mathbf{e}_i$ and remains available to the detection head.

Box-support mask. For each active-frame support box with BEV extent (c_x, c_y, w, h), we define an expanded support region with half-extents

$$\text{half}_w = \frac{w}{2\, v_x\, s}\, r + r_{\min}, \qquad \text{half}_h = \frac{h}{2\, v_y\, s}\, r + r_{\min}, \tag{8}$$

where v_x and v_y are voxel sizes, s is the feature-map stride, r is an expansion ratio, and $r_{\min}$ is a minimum radius. A BEV grid point is in the support set, with m_i=1, if it falls inside any expanded box. During training, we construct the mask from the annotated keyframe boxes at t=0. At inference, we construct the mask from the frozen reference detector's predictions $\mathcal{D}_0$. Although Table 5 shows that using frozen-reference predictions during training yields comparable performance, we retain annotated training supports to align with Ev-3DOD's training convention. The mask only modulates the additional in-support event gain, while event features outside the support, including event-only BEV coordinates, remain available to the detection head.

Temporal schedule. The event representation changes as the accumulation window grows, while the detection head is shared across query times. We therefore make the additional in-support event gain time-dependent. Let $\rho \in [0, 1]$ denote normalized progress over the ordered intermediate query times, with ρ=0 for the first query and ρ=1 for the last. We use a linear decay,

$$\alpha(\rho) \;=\; \alpha_{\text{start}} \;+\; \rho\,(\alpha_{\text{end}} - \alpha_{\text{start}}), \tag{9}$$

where $\alpha_{\text{end}} < \alpha_{\text{start}}$. Inside a support region, where m_i=1, the event gain is $\gamma_{\text{bg}} + \alpha(\rho)$. Outside the support, where m_i=0, the gain remains γ_{bg}.

This schedule gradually reduces the extra emphasis on the fixed keyframe supports, while leaving the LiDAR contribution unchanged. In the DSEC-3DOD ablation (Table 5), it outperforms uniform fusion and the constant-gain alternative.

3.3.4 Center-Based Detection Head

The fused sparse BEV tensor is processed by a center-based detection head (Yin et al., 2021; Chen et al., 2023), which predicts a center heatmap and per-center box regression across the fused BEV sites. At every τ, the head jointly re-estimates target location, geometry, and confidence from the fused BEV tensor.

3.4 Training Objective

Only the asynchronous detector is optimized. We use the standard VoxelNeXt loss, with focal loss (Lin et al., 2017b) for classification and ℓ_1 regression for box parameters. The loss is averaged over the intermediate query times $\mathcal{Q}$ in each blind interval:

$$\mathcal{L} \;=\; \frac{1}{|\mathcal{Q}|} \sum_{\tau \in \mathcal{Q}} \mathcal{L}_{\text{det}}\big(\, \mathcal{D}_\tau,\; \mathcal{D}_\tau^* \,\big), \tag{10}$$

where $\mathcal{Q}$ is the set of intermediate query times and $\mathcal{D}_\tau^*$ is the ground-truth annotation at time τ. The reference detector remains frozen, and its cached features $\mathcal{F}_0$ are treated as fixed inputs. Following Ev-3DOD, the asynchronous detector is trained against the released dense intermediate-time annotations.

4 Experiments

4.1 Experimental Setup

4.1.1 Datasets

We evaluate BEV^{2}3D on DSEC-3DOD (Gehrig et al., 2021; Cho et al., 2025) and Ev-Waymo (Cho et al., 2025; Sun et al., 2020). We use the dataset splits and 100 Hz intermediate-time annotations released by Ev-3DOD and follow its reporting convention (Cho et al., 2025). On DSEC-3DOD, we report Vehicle and Pedestrian results at LEVEL_1. On Ev-Waymo, we report Vehicle, Pedestrian, and Cyclist results at both LEVEL_1 and LEVEL_2. Further details on the datasets and intermediate-time annotations are provided in Appendix A.

4.1.2 Evaluation Protocol and Metrics

We follow the evaluation protocol of Ev-3DOD (Cho et al., 2025). Each interval contains the active timestamp t=0 and nine blind-time query timestamps. At t=0, our method uses the reference detections $\mathcal{D}_0$. At each query time, it uses the asynchronous detections $\mathcal{D}_\tau$. We pool detections and annotations from all ten timestamps before computing AP and APH for each class. In the tables, L, I, and E denote LiDAR, RGB images, and events, respectively.

Online methods use only observations available by the evaluated timestamp. Offline interpolation baselines may also use observations from the next active frame. They therefore provide non-causal references. N-PCI denotes NeuralPCI (Zheng et al., 2023), which interpolates LiDAR point clouds. EMA (Zhang et al., 2023b) and CBM (Kim et al., 2023) are frame-based and event-based video interpolation methods, respectively. Unless otherwise noted, the baseline results and method combinations in the main comparison tables are taken from Ev-3DOD (Cho et al., 2025) under this protocol. Implementation details and hyperparameters are reported in Appendix B.

Table 2: Ev-Waymo results (AP/APH) under the Ev-3DOD evaluation protocol, aggregating the active timestamp and nine blind-time query timestamps. Baseline results are taken from Ev-3DOD. Best and second-best results are bold and underlined within each offline or online setting, respectively.

	Interp. Method	Method	Mod.	VEH		PED		CYC	
				L1	L2	L1	L2	L1	L2
Offline	N-PCI	VoxelNeXt	L	62.96/62.51	59.10/58.67	**68.43/60.77**	**62.85/55.69**	**44.55/43.15**	**38.88/37.65**
		HEDNet	L	59.57/59.09	55.90/55.45	63.28/55.32	58.03/50.62	43.14/41.80	37.63/36.46
	N-PCI+EMA	Focals Conv	L+I	55.49/55.02	54.45/53.99	48.07/43.06	46.54/41.69	27.70/26.60	26.59/25.54
		LoGoNet	L+I	63.48/63.05	60.15/59.73	67.85/60.28	62.43/55.37	42.40/41.37	37.29/36.38
	N-PCI+CBM	Focals Conv	L+I	55.48/55.01	54.44/53.98	48.07/43.06	46.54/41.68	27.67/26.57	26.56/25.50
		LoGoNet	L+I	**63.60/63.16**	**60.27/59.85**	68.14/60.53	62.77/55.67	42.85/41.81	37.66/36.74
Online	N/A	VoxelNeXt	L	44.40/44.10	41.78/41.49	40.52/36.23	36.93/32.96	24.40/23.73	21.24/20.66
		HEDNet	L	42.03/41.71	39.32/39.02	38.86/34.43	35.67/31.53	22.64/21.99	19.72/19.14
		Focals Conv	L+I	37.31/37.01	36.60/36.31	29.20/26.16	28.41/25.44	14.30/13.78	13.79/13.29
		LoGoNet	L+I	44.14/43.87	41.73/41.47	39.98/35.84	36.48/32.67	24.71/24.15	21.59/21.10
		Ev-3DOD	L+I+E	60.30/59.95	59.19/58.85	57.40/50.78	55.30/48.93	31.08/30.38	29.69/29.03
		Ours	L+E	**71.48/70.99**	**70.48/69.99**	**66.53/60.00**	**65.35/58.76**	**41.34/40.42**	**40.55/39.54**
			Δ	+11.18/+11.04	+11.29/+11.14	+9.13/+9.22	+10.05/+9.83	+10.26/+10.04	+10.86/+10.51

4.2 Main Results

Tables 1 and 2 report results on DSEC-3DOD and Ev-Waymo, respectively. On DSEC-3DOD, BEV^{2}3D uses LiDAR and events to reach 50.20/49.46 AP/APH on Vehicle and 25.24/18.15 on Pedestrian at LEVEL_1. It improves over Ev-3DOD, which uses RGB, LiDAR, and events, by +8.55/+8.26 and +4.56/+6.27, respectively. On Ev-Waymo, BEV^{2}3D improves over Ev-3DOD by +11.18/+11.29 AP on Vehicle at LEVEL_1/LEVEL_2. The corresponding gains are +9.13/+10.05 AP on Pedestrian and +10.26/+10.86 AP on Cyclist. These gains are consistent across datasets, object categories, and difficulty levels. They demonstrate the effectiveness of BEV^{2}3D for asynchronous 3D detection during LiDAR blind intervals. We also compare the two methods at individual query times. As shown in Appendix E, BEV^{2}3D outperforms Ev-3DOD at every evaluated intermediate timestamp in overall AP on DSEC-3DOD and in AP for all three classes on Ev-Waymo.

Table 1: Results on DSEC-3DOD under the Ev-3DOD evaluation protocol, aggregating the active timestamp and nine blind-time query timestamps. Baseline results are taken from Ev-3DOD. Best and second-best results are bold and underlined within each offline or online setting, respectively.

	Interp. Method	Method	Mod.	VEH	PED
Offline	N-PCI	VoxelNeXt	L	**45.35/44.93**	27.23/19.93
		HEDNet	L	41.82/41.35	**33.64/20.10**
	N-PCI + EMA	Focals Conv	L+I	35.52/35.04	25.05/16.65
		LoGoNet	L+I	44.64/44.06	29.96/18.72
	N-PCI + CBM	Focals Conv	L+I	35.56/35.08	24.99/16.64
		LoGoNet	L+I	44.76/44.17	29.94/18.73
Online	N/A	VoxelNeXt	L	21.42/21.23	6.46/4.57
		HEDNet	L	18.48/18.29	9.02/5.07
		Focals Conv	L+I	15.74/15.55	5.88/3.67
		LoGoNet	L+I	19.50/19.32	7.07/4.49
		Ev-3DOD	L+I+E	41.65/41.20	20.68/11.88
		Ours	L+E	**50.20/49.46**	**25.24/18.15**
			Δ	+8.55/+8.26	+4.56/+6.27

Notably, BEV^{2}3D outperforms several offline interpolation baselines on DSEC-3DOD and most offline interpolation results on Ev-Waymo. These non-causal baselines access the next active frame. This shows that causal blind-time detection can remain competitive with non-causal interpolation.

To better contextualize these gains, we compare the reference detectors used by the two methods: VoxelNeXt for BEV^{2}3D and LoGoNet for Ev-3DOD. The online baseline results show comparable performance for VoxelNeXt (LiDAR) and LoGoNet (LiDAR+RGB). On Ev-Waymo, their results differ by at most 0.54 LEVEL_1 AP across the three classes. By contrast, the gap between BEV^{2}3D and Ev-3DOD reaches 9–11 AP in blind-time detection on Ev-Waymo. This disparity suggests that the gains are unlikely to arise solely from differences in reference-detector performance, and are instead associated with how blind-time evidence is incorporated during the asynchronous update.

Table 3: DSEC-3DOD support-constrained ablation (LEVEL_1 AP/APH). The constraint retains the indicated fraction of the BEV grid for the fused detection tensor. Center coverage is measured on intermediate ground-truth boxes.

Fused-BEV availability	BEV retained	GT centers covered	Vehicle AP / APH	Pedestrian AP / APH
Keyframe support (r=1.0)	2.14%	92.91%	40.03 / 39.45	18.95 / 14.05
Keyframe support (r=2.0)	5.06%	94.05%	45.43 / 44.75	23.03 / 16.71
Full BEV	100%	100%	**50.20 / 49.46**	**25.24 / 18.15**

Table 4: Comparison of four event-to-BEV lifting configurations (LEVEL_1 AP/APH).

Configuration	DSEC-3DOD		Ev-Waymo		
	Vehicle	Pedestrian	Vehicle	Pedestrian	Cyclist
soft prior	**50.20 / 49.46**	25.24 / 18.15	71.48 / 70.99	66.53 / 60.00	**41.34 / 40.42**
hard one-hot prior	49.71 / 48.99	24.16 / 17.74	70.85 / 70.38	64.93 / 58.25	40.84 / 39.87
LSS (no LiDAR prior)	47.73 / 47.08	**25.88 / 18.52**	68.27 / 67.83	60.73 / 54.80	37.85 / 36.99
Depth-cond. (LiDAR input channel)	46.41 / 45.73	22.57 / 16.50	**71.52 / 71.08**	**67.66 / 60.98**	40.85 / 39.95

Inference latency. On DSEC-3DOD, BEV^{2}3D processes each event-time update in 11.8 ms, compared with 12.7 ms for Ev-3DOD. Appendix C provides the measurement details.

4.3 Fused-BEV Context Beyond Keyframe Supports

We assess whether blind-time detection benefits from fused-BEV context beyond the regions surrounding active-frame detections. Ev-3DOD updates individual RoIs by aggregating point-wise LiDAR features within each proposal. BEV^{2}3D instead predicts from a fused BEV tensor built on cached sparse voxel features. Exchanging these prediction heads would change the detector parameterization, feature representation, and capacity rather than isolate the spatial evidence available for prediction. We therefore retain the same BEV detection head and restrict its fused-BEV input to the union of expanded keyframe support regions. This yields a controlled comparison of full-BEV and proposal-supported spatial access. During training, the supports are derived from t=0 ground-truth boxes. At inference, they are derived from frozen reference predictions. All other components and training settings are identical to those of the full model.

Table 3 shows that restricting the detector to expanded keyframe supports substantially reduces detection performance. With r=2.0, the constrained model retains only 5.06% of the BEV grid and still covers 94.05% of intermediate ground-truth centers. It loses 4.77 Vehicle AP and 2.21 Pedestrian AP relative to the full-BEV model. The tighter r=1.0 support reduces both the retained area and accuracy further. These results show that high coverage of intermediate object centers alone is insufficient to preserve full-BEV performance. More broadly, relying primarily on local object-centric features to update motion cues may miss useful blind-time evidence beyond active-frame detection regions.

4.4 Ablations

We ablate the two core modules: FEL on both benchmarks and SGTF on DSEC-3DOD. Table 4 compares alternatives for the LiDAR-conditioned event lift, and Table 5 isolates the support-gated temporal fusion design. Endpoint sensitivity of the SGTF schedule is reported in Appendix D.

Table 4 shows that FEL conditions the event-to-BEV lift on cached LiDAR geometry. On Ev-Waymo, LSS reduces all classes by 3.2–5.8 AP relative to the soft prior, and the hard prior also yields lower performance. Depth conditioning is competitive on Ev-Waymo but regresses substantially on DSEC. The soft prior provides the most consistent performance across the two benchmarks and serves as our shared configuration.

As shown in Table 5, uniform fusion removes both support conditioning and temporal variation, reducing Vehicle AP by 6.5 points. Retaining support conditioning but replacing the time-varying

Table 5: SGTF ablation on DSEC-3DOD (LEVEL_1 AP/APH, Vehicle / Pedestrian).

Configuration	Vehicle (AP/APH)	Pedestrian (AP/APH)
BEV^{2}3D (full; GT supports during training)	**50.20 / 49.46**	25.24 / 18.15
w/o Support-Gated Temporal Fusion (uniform γ=1.0)	43.67 / 43.16	19.68 / 14.30
constant support gain ($\alpha(\rho) \equiv 0.6$; matched mean)	38.96 / 38.36	16.32 / 11.81
BEV^{2}3D (PRED supports during training)	49.10 / 48.52	**26.14 / 18.69**

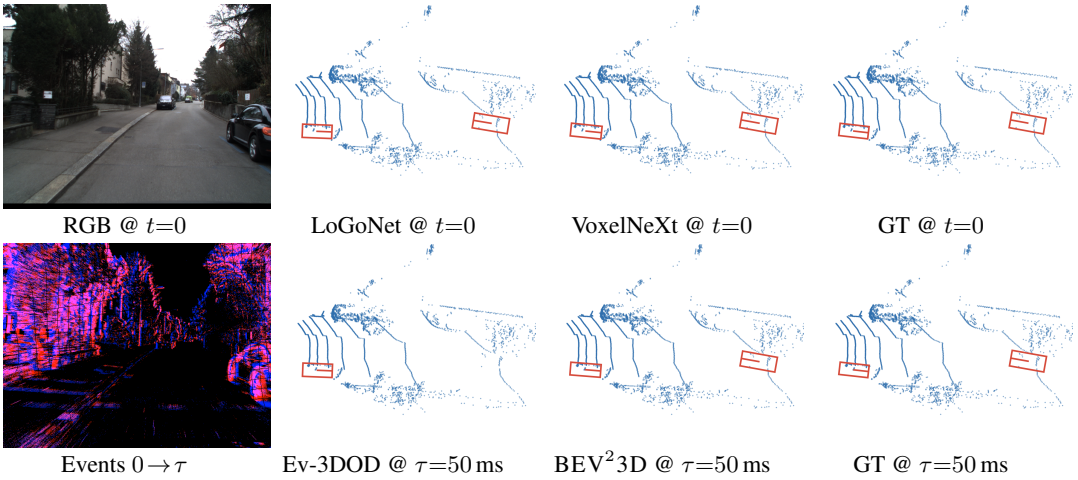


Figure 4: **Qualitative comparison on DSEC-3DOD.** The top row shows the active-frame reference detectors, LoGoNet for Ev-3DOD and VoxelNeXt for BEV^{2}3D, which both align with the two ground-truth vehicles. The bottom row compares blind-time detections at τ=50 ms, where Ev-3DOD misses one vehicle whereas BEV^{2}3D retains both detections. To compare them clearly, we plot all detections and ground truth in the t=0 LiDAR BEV frame.

support gain with a constant $\alpha(\rho) \equiv 0.6$ preserves the same mean in-support event gain ($\gamma_{\text{bg}} + \bar{\alpha} = 1.6$), yet reduces Vehicle AP by 11.2 points. Thus, the fixed-gain alternative performs substantially worse than the time-varying schedule in this configuration. Finally, replacing annotated keyframe supports with frozen-reference predictions during training yields comparable results. Across Vehicle and Pedestrian, AP and APH differ by at most 1.1 and 0.94 points, respectively, with neither support source consistently outperforming the other. This suggests that the training-time support source has little effect on performance in this setting.

4.5 Qualitative Results

Figure 4 compares active-frame detections and their blind-time updates on a DSEC-3DOD scene. At t=0, both reference detectors produce detections that are well aligned with the two ground-truth vehicles. By τ=50 ms, the predictions have diverged: Ev-3DOD retains only one vehicle, while BEV^{2}3D maintains detections for both vehicles with consistent spatial alignment. This example qualitatively illustrates the more complete blind-time predictions of BEV^{2}3D, consistent with the quantitative results in Appendix E. Appendix F provides complete qualitative results across the full blind interval on DSEC-3DOD and Ev-Waymo.

5 Conclusion

We present BEV^{2}3D, which updates a spatial BEV representation with events between LiDAR frames. FEL lifts accumulated event evidence with cached LiDAR geometry as a soft depth prior. SGTF then integrates the event BEV with cached LiDAR features over the blind interval. Across DSEC-3DOD and Ev-Waymo, BEV^{2}3D improves asynchronous detection over Ev-3DOD while using LiDAR and events without RGB. We discuss the limitations of BEV^{2}3D in Appendix G.

AI Assistance Statement. We used generative AI tools to aid literature discovery and to draft and polish portions of this manuscript. All AI-assisted text and cited sources were reviewed and revised or verified by the authors. We also used AI-generated visual assets for selected conceptual and architectural illustrations, which were subsequently curated and assembled by the authors. These assets were used only for illustrative purposes and not for experimental visualizations. The authors take full responsibility for the final content of this work.

Reproducibility Statement. We provide anonymous code in the supplementary material, with README instructions for reproducing our experiments. The code will be made publicly available upon acceptance.

REFERENCES

Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1090–1099, 2022.

Yukang Chen, Yanwei Li, Xiangyu Zhang, Jian Sun, and Jiaya Jia. Focal sparse convolutional networks for 3d object detection. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 5418–5427. IEEE, 2022.

Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21674–21683. IEEE, 2023.

Hoonhee Cho, Jae-young Kang, Youngho Kim, and Kuk-Jin Yoon. Ev-3dod: Pushing the temporal boundaries of 3d object detection with event cameras. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27197–27210. IEEE, 2025.

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

Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Recycling video datasets for event cameras. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3583–3592. IEEE, 2020.

Mathias Gehrig and Davide Scaramuzza. Recurrent vision transformers for object detection with event cameras. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13884–13893. IEEE, 2023.

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

Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021.

Jae-Young Kang, Hoonhee Cho, and Kuk-Jin Yoon. Unleashing the temporal potential of stereo event cameras for continuous-time 3d object detection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6869–6881. IEEE, 2025.

Taewoo Kim, Yujeong Chae, Hyun-Kurl Jang, and Kuk-Jin Yoon. Event-based video frame interpolation with cross-modal asymmetric bidirectional motion fields. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18032–18042. IEEE, 2023.

Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 12689–12697. IEEE, 2019.

Mengtian Li, Yu-Xiong Wang, and Deva Ramanan. Towards streaming perception. In *European conference on computer vision*, pp. 473–488. Springer, 2020.

Xin Li, Tao Ma, Yuenan Hou, Botian Shi, Yuchen Yang, Youquan Liu, Xingjiao Wu, Qin Chen, Yikang Li, Yu Qiao, et al. Logonet: Towards accurate 3d object detection with local-to-global cross-modal fusion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17524–17534. IEEE, 2023a.

Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 1477–1485, 2023b.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):2020–2036, 2024.

Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120 db 15 μs latency asynchronous temporal contrast vision sensor. *IEEE journal of solid-state circuits*, 43(2):566–576, 2008.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 936–944. Ieee, 2017a.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017b.

Haisong Liu, Yao Teng, Tao Lu, Haiguang Wang, and Limin Wang. Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18534–18544. IEEE, 2023a.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9992–10002. Ieee, 2021.

Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, pp. 2774–2781. ieee, 2023b.

Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European conference on computer vision*, pp. 194–210. Springer, 2020.

Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):1964–1980, 2019.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2443–2451. IEEE, 2020.

Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18 (10):3337, 2018.

Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17830–17839. IEEE, 2023.

Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11784–11793, 2021.

Gang Zhang, Chen Junnan, Guohuan Gao, Jianmin Li, and Xiaolin Hu. Hednet: A hierarchical encoder-decoder network for 3d object detection in point clouds. *Advances in Neural Information Processing Systems*, 36:53076–53089, 2023a.

Guozhen Zhang, Yuhan Zhu, Haonan Wang, Youxin Chen, Gangshan Wu, and Limin Wang. Extracting motion and appearance via inter-frame attention for efficient video frame interpolation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5682–5692. IEEE, 2023b.

Zehan Zheng, Danni Wu, Ruisi Lu, Fan Lu, Guang Chen, and Changjun Jiang. Neuralpci: Spatio-temporal neural field for 3d point cloud multi-frame non-linear interpolation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 909–918. IEEE, 2023.

Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Ev-flownet: Self-supervised optical flow estimation for event-based cameras. *arXiv preprint arXiv:1802.06898*, 2018.

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

A Detailed Datasets and Annotation Protocol

DSEC-3DOD. Built on the DSEC stereo event dataset (Gehrig et al., 2021), DSEC-3DOD provides stereo Prophesee Gen3.1M event cameras and a Velodyne VLP-16 LiDAR at 10 Hz. We use the left event stream and, following Ev-3DOD's main DSEC evaluation setting (Cho et al., 2025), accumulated LiDAR sweeps augmented with disparity-derived 3D points. We adopt Ev-3DOD's released 100 Hz annotations and split (Cho et al., 2025). For blind-time labels, Ev-3DOD linearly interpolates the active-frame boxes, generates intermediate sensor data through event-based video and point-cloud interpolation, and has annotation experts refine the interpolated boxes.

Ev-Waymo. We use the benchmark released by Ev-3DOD (Cho et al., 2025). It is derived from the Waymo Open Dataset (Sun et al., 2020), which provides synchronized multi-camera RGB and LiDAR at 10 Hz with high-quality 3D bounding boxes. Ev-3DOD synthesizes its event streams from RGB frames with Vid2E (Gehrig et al., 2020) and provides 100 Hz blind-time annotations using the interpolation-and-refinement procedure above. We report the three official classes (Vehicle, Pedestrian, Cyclist).

Splits. For both DSEC-3DOD and Ev-Waymo, we follow the train/validation splits released by Ev-3DOD (Cho et al., 2025).

B Key Hyperparameters and Training Recipe

Table 6 reports the key architectural and training hyperparameters used for the asynchronous detector.

Table 6: Hyperparameters of the asynchronous detector and its training.

Component	Value
Reference detector	VoxelNeXt (Chen et al., 2023) (frozen)
Event backbone	Swin-T (Liu et al., 2021) + FPN (Lin et al., 2017a), width 128, L=4
Sparse BEV refinement	2 sparse 2D conv layers
Detection head	center-based (Yin et al., 2021)
Event input size (DSEC / Ev-Waymo)	240×320 / 480×960
Event image transform	fixed 0.5× resize; no crop, rotation, or flip
Event voxelization bins	5
Point-cloud range (DSEC)	$[0, 57.6]\times[-28.8, 28.8]\times[-3, 3]$ m
Point-cloud range (Ev-Waymo)	$[0, 75.2]\times[-37.6, 37.6]\times[-2, 4]$ m
Voxel size	$(0.1, 0.1, 0.15)$ m
FEL depth bins D (DSEC / Ev-Waymo)	64 / 96
FEL depth range (DSEC / Ev-Waymo)	$[1.0, 57.6]$ / $[1.0, 75.2]$ m
FEL Gaussian width σ (DSEC / Ev-Waymo)	one bin width (≈ 0.88 / 0.77 m)
FEL correction head Δ	zero-initialized final layer
SGTF γ_{bg}	1.0
SGTF α_{start}, α_{end}	1.0, 0.2
SGTF expansion ratio r, $r_{\min}$	1.25, 1.0
Optimizer	AdamW (β_1=0.9, β_2=0.99, wd 0.01)
Learning rate	3×10^{-3}
Batch size (DSEC / Ev-Waymo)	4 / 2 per GPU
Epochs	20
Random seed	666
LR schedule	one-cycle (pct_start 0.4, div_factor 10, moms $[0.95, 0.85]$)
GPUs	2× NVIDIA GeForce RTX 4090 D

C Latency and Efficiency

Table 7 compares the per-update latency at each event timestamp. By reusing cached LiDAR features, BEV^{2}3D produces a BEV-based detection update in 11.8 ms ($\sim$85 Hz), slightly faster than the 12.7 ms (79 Hz) of Ev-3DOD. It should be acknowledged that the current implementation does not yet sustain the full 100 Hz query rate. Further implementation-level optimizations, such as kernel fusion, could reduce the update latency.

Table 7: Latency to process one event-time input on DSEC-3DOD val, RTX 4090 D, batch size 1, measured with CUDA events after inputs are on GPU.

Method	Latency per event time (ms)	Rate (Hz)
Ev-3DOD	12.7	79
BEV^{2}3D	11.8	$\sim$85

D Extended Ablations

Table 8 reports the SGTF endpoint sensitivity study. We vary α_{end} across both datasets to assess the robustness of the shared configuration.

Table 8: SGTF endpoint sensitivity (LEVEL_1 AP/APH). Best and second-best results are bold and underlined. The shared default is $\alpha_{\text{end}} = 0.2$; DSEC results are mixed across classes, whereas Ev-Waymo favors the shared default.

Configuration	DSEC V	DSEC P	Ev-Waymo V	Ev-Waymo P	Ev-Waymo C
$\alpha_{\text{end}} = 0.0$	<u>50.12 / 49.41</u>	**25.88 / 18.68**	70.28 / 69.84	64.45 / 58.02	<u>40.69 / 39.78</u>
$\alpha_{\text{end}} = 0.2$ (base)	**50.20 / 49.46**	<u>25.24</u> / 18.15	**71.48 / 70.99**	**66.53 / 60.00**	**41.34 / 40.42**
$\alpha_{\text{end}} = 0.4$	49.22 / 48.50	24.54 / <u>18.38</u>	<u>70.60 / 70.13</u>	<u>64.99 / 58.48</u>	38.86 / 37.94
$\alpha_{\text{end}} = 0.6$	46.02 / 45.39	20.20 / 15.10	70.13 / 69.70	63.61 / 57.55	39.84 / 38.98
$\alpha_{\text{end}} = 0.8$	44.90 / 44.20	20.21 / 14.99	69.88 / 69.42	63.32 / 57.22	40.07 / 39.23
$\alpha_{\text{end}} = 1.0$	38.51 / 37.96	16.43 / 11.97	70.07 / 69.63	62.69 / 56.68	39.35 / 38.52

On DSEC-3DOD, the shared default α_{end}=0.2 gives the best Vehicle AP, while 0.0 gives the best Pedestrian AP; relative to the shared default, 0.0 changes Vehicle and Pedestrian AP by -0.08 and $+0.64$, respectively. On Ev-Waymo, however, reducing the endpoint to 0.0 decreases AP by 1.20, 2.08, and 0.65 for Vehicle, Pedestrian, and Cyclist. Across both datasets, larger endpoints generally degrade performance, supporting the use of a reduced support-conditioned gain toward the end of the blind interval. We therefore use α_{end}=0.2 as a shared configuration across both datasets rather than tuning the endpoint separately for each benchmark.

E Blind-Interval Diagnostics

We report time-resolved AP and localization diagnostics across the blind interval in Figures 5 and 6. On DSEC-3DOD, BEV^{2}3D remains ahead of Ev-3DOD in overall AP at every intermediate timestamp. The common-TP curves show a lower center distance throughout the interval and comparable 3D IoU, with small crossings at individual timestamps. On Ev-Waymo, BEV^{2}3D improves AP for Vehicle, Pedestrian, and Cyclist at every timestamp, and the advantage persists from the early to the late part of the blind interval.

F Extended Qualitative Results

This appendix provides complete blind-interval qualitative comparisons on DSEC-3DOD and Ev-Waymo. Figure 7 extends the DSEC-3DOD example in Figure 4 to all nine intermediate query

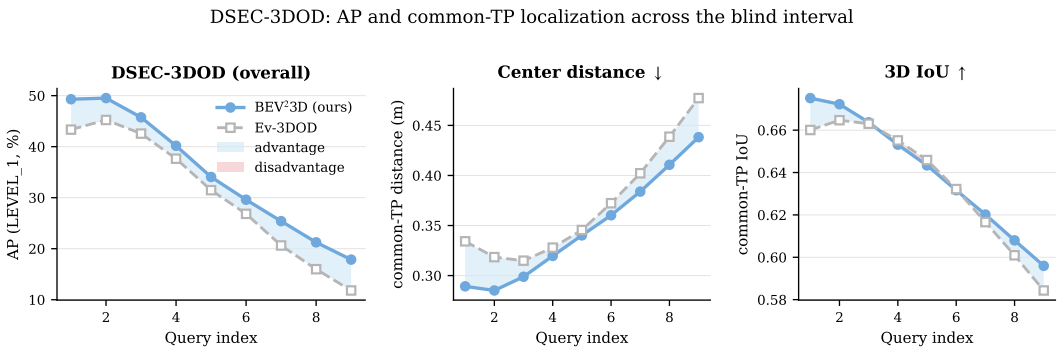


Figure 5: **Per-timestamp analysis on DSEC-3DOD.** The three panels show AP at each blind-time timestamp (LEVEL_1), center distance, and 3D IoU. The localization metrics are computed on the same GT instances matched by both methods at IoU$\geq$0.1, which avoids method-specific TP selection bias. Arrows indicate the preferred direction for each metric, and the shaded regions mark which method is better. BEV^{2}3D maintains the higher AP, with lower center distance and comparable 3D IoU on this common subset.

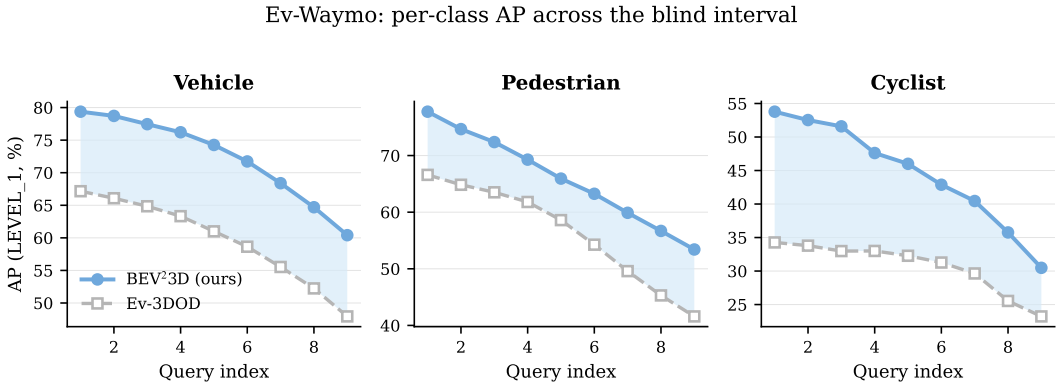


Figure 6: **Per-timestamp AP on Ev-Waymo.** LEVEL_1 AP for Vehicle, Pedestrian, and Cyclist at each intermediate timestamp in the blind interval. The shaded regions mark the per-timestamp advantage of BEV^{2}3D over Ev-3DOD.

times. Figure 8 shows a full blind-interval sequence from Ev-Waymo. For visual clarity, the same confidence threshold is applied to both methods in each comparison. On DSEC-3DOD, BEV^{2}3D maintains vehicle detections more consistently across the blind interval than Ev-3DOD. On Ev-Waymo, BEV^{2}3D recovers an additional vehicle for most intermediate queries, with fewer missed detections than Ev-3DOD.

G LIMITATIONS

Events provide limited corrective evidence when neither ego motion nor object motion produces substantial brightness changes. In such scenes, the event stream is sparse, resulting in a weak E_τ and predictions that remain close to the cached reference. At the other extreme, very fast nearby motion can produce dense and overlapping event traces within the discretized event volume, increasing temporal and spatial ambiguity during event-to-BEV lifting. These cases remain challenging for the current formulation.

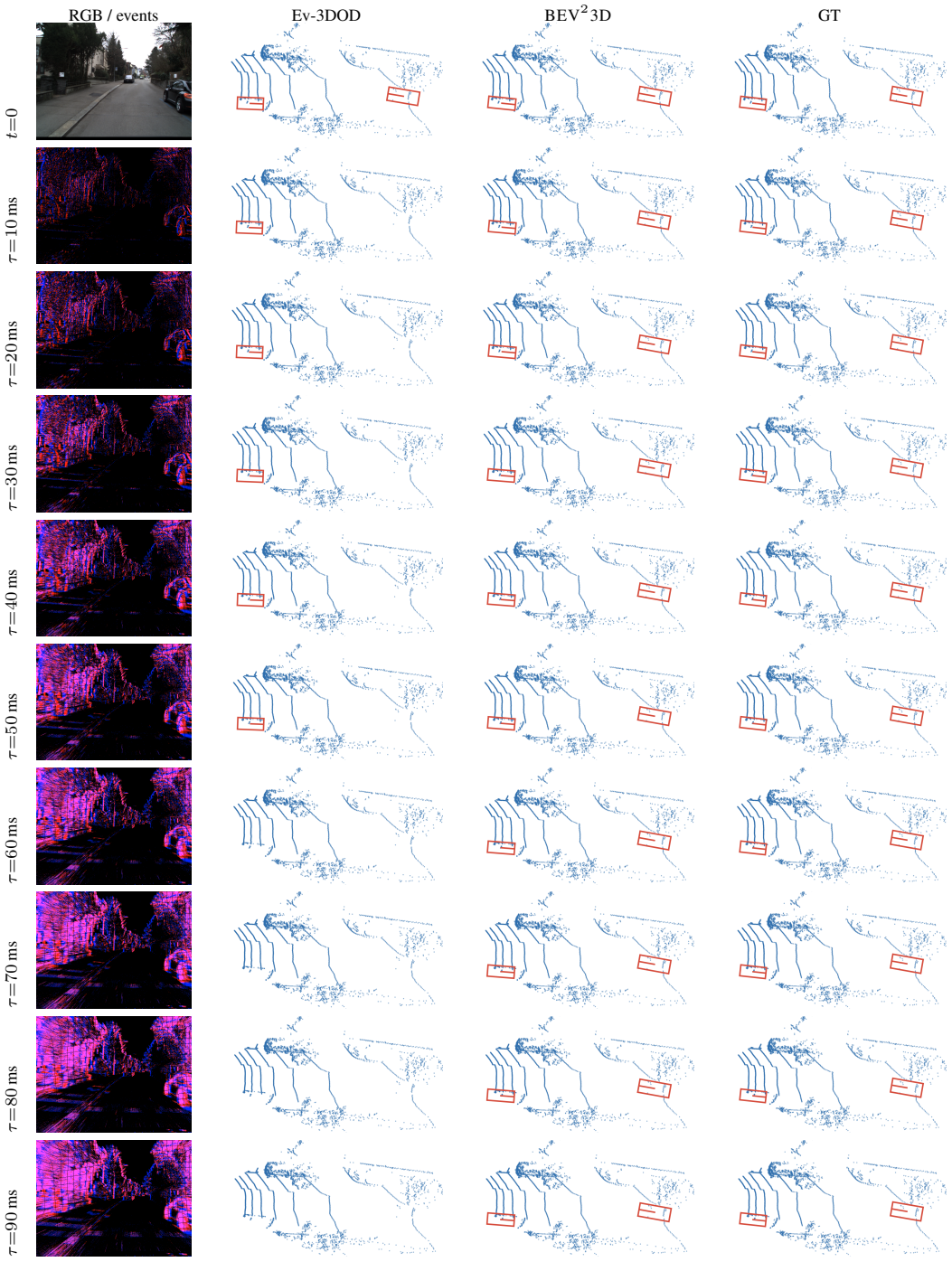


Figure 7: **Extended qualitative comparison on DSEC-3DOD.** The rows span the active frame and all nine blind-time queries. The first image column shows the active RGB frame followed by cumulative event images. The remaining columns show the detections of Ev-3DOD and BEV^{2}3D and the ground truth. All BEV results are plotted in the t=0 LiDAR coordinate frame.

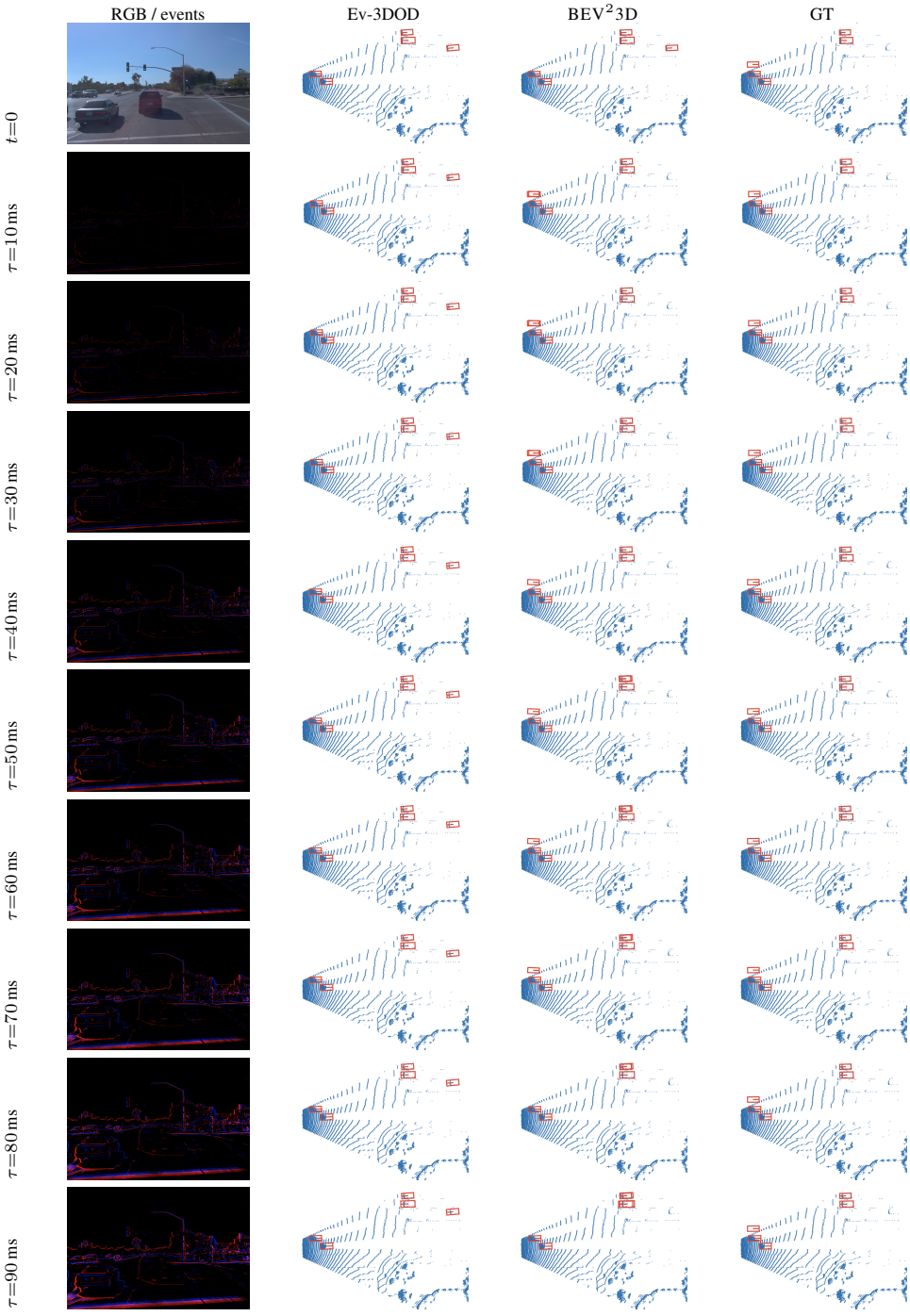


Figure 8: **Extended qualitative comparison on Ev-Waymo.** The rows span the active frame and all nine blind-time queries. The first image column shows the active RGB frame followed by cumulative event images. The remaining columns show the detections of Ev-3DOD and BEV^{2}3D and the ground truth. All BEV results are plotted in the t=0 LiDAR coordinate frame.