EVENTGEOPOSE: POSE-REFERENCED EVENT ENCODING AND EVENT-GEOMETRY CONSISTENCY FOR SPIKING 6D POSE ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Event cameras have shown considerable potential for 6D object pose estimation under fast motion and challenging illumination. However, current event-based approaches still rely heavily on dense ANN computation, fixed-duration temporal discretization, and conventional pose supervision, which respectively limit energy-efficient event processing, representation of highly nonuniform event dynamics, and explicit exploitation of event-derived geometric evidence. To address these issues, we present EventGeoPose, a directly trained SNN framework for 6D object pose estimation from event streams. We construct the spiking backbone with I-LIF neurons to improve the numerical fidelity of continuous regression while retaining spike-driven inference. We propose Pose-Referenced Adaptive Event Encoding, which combines activity-adaptive temporal discretization with duration-normalized local event-rate encoding to mitigate motion-state mixing and uneven observation density. We further introduce Event-Geometry Consistency Regularization, which couples predicted 3D object geometry with observed event structures through differentiable image-space reprojection, providing observation-aware geometric supervision for pose prediction. Additionally, we construct two complementary event-based datasets, EVA-6D-Syn and EVA-6D-Real, covering diverse motion, illumination, and real-sensor conditions. Extensive experiments demonstrate consistent improvements over representative methods while maintaining low theoretical inference energy, highlighting the potential of event-driven SNNs for robust and energy-efficient 6D pose estimation.

1 INTRODUCTION

The six-degree-of-freedom (6D) pose of a rigid object describes its three-dimensional rotation and translation relative to the camera, providing essential spatial information for robotic manipulation Liu et al. (2023a); Wen et al. (2022), VR/AR Aguirrezabal et al. (2026); Manawadu et al. (2024); Zhao et al. (2023), and human-machine interaction Xia et al. (2025); Fan et al. (2022); Liu et al. (2023b). Accurate and robust 6D pose estimation is therefore fundamental to visual systems that interact with dynamic physical environments.

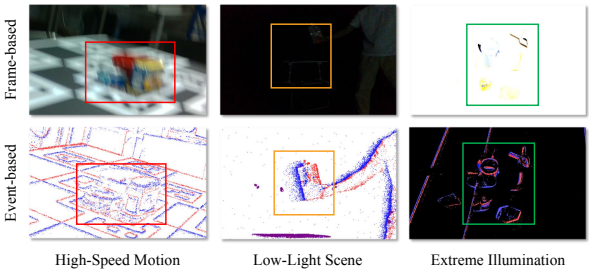


Figure 1: **Comparison of frame- and event-based observations under challenging conditions.** Conventional frames suffer from motion blur and severe exposure degradation, whereas event cameras retain informative object contours and motion cues, providing more reliable visual evidence for 6D pose estimation.

Most existing methods, however, rely on conventional frame-based cameras, whose fixed-rate exposure is susceptible to motion blur under rapid motion and severe degradation under challenging illumination Kar et al. (2026); Chakravarthi et al. (2024); Gallego et al. (2020), as illustrated in Fig. 1. Event cameras provide an alternative sensing paradigm by

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.

asynchronously reporting pixel-wise brightness changes, offering microsecond-level temporal resolution, a dynamic range exceeding 100 dB, and low-power operation. These properties allow event cameras to preserve informative motion and structural cues under fast motion and large illumination variations, making them particularly attractive for 6D object pose perception in highly dynamic scenes Glover et al. (2024); Li et al. (2023); Kang et al. (2026).

Despite these advantages, fully realizing the potential of event cameras for 6D object pose estimation requires addressing three key issues. ***First***, existing event-based approaches Zhang et al. (2025); Liu et al. (2025a); Wang et al. (2026); Yang et al. (2025) usually convert asynchronous events into dense representations and process them with conventional artificial neural networks (ANNs). Such dense, clock-driven computation incurs substantial computational and energy costs, limiting the low-power benefits of event sensing, particularly for continuously operating edge platforms. ***Second***, event streams exhibit highly nonuniform temporal activity, whereas commonly used fixed-duration discretization allocates temporal resolution independently of scene dynamics. During rapid motion, a single bin may span multiple pose states, causing motion-state mixing, while low-activity intervals may provide insufficient observations. This imbalance compromises both fine-grained motion information and structural evidence required for pose estimation. ***Third***, existing pose objectives Maji et al. (2024); Ousalah et al. (2026); Çetiner & Ekenel (2026) mainly constrain errors in pose parameters or transformed 3D model points, but rarely enforce consistency between predicted geometry and the event structures observed in the image plane. Since events primarily arise around moving edges and contours, they provide direct geometric evidence for evaluating pose consistency. The lack of such event-aware supervision leaves a gap between pose prediction and observed visual evidence.

To address these challenges, we propose EventGeoPose, a directly trained spiking neural network (SNN) Maass (1997) framework for 6D pose estimation from event streams. ***At the computational level***, EventGeoPose adopts spike-driven computation to exploit the sparse nature of event sensing. To mitigate the quantization error of binary spikes in continuous pose regression, we construct the spiking backbone with Integer-valued Leaky Integrate-and-Fire (I-LIF) neurons Luo et al. (2024), which improve numerical fidelity through integer-valued training while retaining binary spike-driven inference. ***At the representation level***, we introduce Pose-Referenced Adaptive Event Encoding (PRAEE), which combines activity-adaptive temporal discretization with duration-normalized local event-rate cues. PRAEE balances event evidence across timesteps, accounts for variable bin durations, and reduces motion-state mixing to better capture pose-relevant event dynamics. ***At the supervision level***, we propose Event-Geometry Consistency Regularization (EGCR), which projects predicted 3D object geometry into the image plane and enforces consistency with event-derived geometric evidence. EGCR is applied only during training and introduces no additional inference overhead. In addition, we construct two complementary event-based 6D pose datasets, EVA-6D-Syn and EVA-6D-Real, covering diverse motion patterns, illumination conditions, and real sensor observations. Experiments on the proposed datasets and a public benchmark demonstrate consistent improvements over representative methods while maintaining low theoretical inference energy.

2 Related Work

Frame-based Object Pose Estimation. Existing frame-based 6D object pose estimation methods can be broadly categorized into feature matching, template matching, and direct regression. Feature-matching methods Lin et al. (2024); Liu et al. (2025b); Yu et al. (2025); Moon et al. (2025) establish correspondences between observations and object models followed by geometric pose recovery, but typically rely on stable texture or shape cues that are difficult to obtain from sparse event data. Template-matching methods Wen et al. (2024); Chen et al. (2026); Labbé et al. (2022); Nguyen et al. (2024); Lee et al. (2025) estimate pose by comparing observations with rendered templates, reducing the dependence on local correspondences at the cost of additional rendering, retrieval, and hypothesis evaluation. Direct-regression methods Maji et al. (2024); Liu et al. (2025c); Ousalah et al. (2026); Çetiner & Ekenel (2026) learn an end-to-end mapping from observations to pose parameters, avoiding explicit correspondence construction and template retrieval. This paradigm is more amenable to efficient multi-object inference, but provides weaker explicit geometric verification, motivating additional geometry-aware supervision.

Event-Based Vision and Spiking Perception. Event-based object-level 6D pose estimation remains relatively underexplored. Recent datasets, including YCB-Ev Rojtberg & Pöllabauer (2024), YCB-

Ev SD Rojtberg & Kühn (2025), and E-POSE Hay et al. (2025a), have begun to provide event observations and pose annotations for this task. Existing methods mainly rely on explicit correspondences, template matching, motion propagation, or auxiliary geometric information Liu et al. (2024); Wang et al. (2026); Hay et al. (2025b); Li et al. (2026); Kang et al. (2026), and predominantly employ conventional ANNs or explicit geometric optimization. In parallel, advances in surrogate-gradient learning have extended directly trained SNNs to dense visual tasks such as object detection Su et al. (2023); Yao et al. (2024); Fan et al. (2024). SpikeYOLO Luo et al. (2024) further introduced the I-LIF neuron to improve the numerical fidelity of continuous regression while retaining spike-driven inference. Existing SNN-based pose studies remain limited to specialized single-object settings or hybrid pipelines with explicit geometric solvers Courtois et al. (2025); Nguyen et al. (2026). Consequently, directly trained SNN frameworks for joint multi-object detection and direct 6D pose regression from event streams remain largely unexplored.

3 METHOD

3.1 OVERVIEW

Given an observation window $[t_s, t_r]$, where t_r denotes the pose reference time Rojtberg & Pöllabauer (2024); Rojtberg & Kühn (2025), an event stream is represented as

$$\mathcal{E} = \{e_i\}_{i=1}^{N}, \qquad e_i = (x_i, y_i, t_i, p_i), \tag{1}$$

where (x_i, y_i), t_i, and $p_i \in \{-1, +1\}$ denote the pixel location, timestamp, and polarity of the i-th event, respectively, with $t_s \le t_1 \le \cdots \le t_N \le t_r$. Given $\mathcal{E}$ and the camera intrinsics $\mathbf{K}$, our goal is to jointly detect multiple objects and estimate their 6D poses. For the m-th predicted instance, the output is represented as $\left\{\hat{c}_m, \hat{\mathbf{b}}_m, \hat{\mathbf{R}}_m, \hat{\mathbf{t}}_m\right\}$, where $\hat{c}_m$, and $\hat{\mathbf{b}}_m$ denote the class label and bounding box, while $\hat{\mathbf{R}}_m \in \mathrm{SO}(3)$ and $\hat{\mathbf{t}}_m \in \mathbb{R}^3$ denote the predicted rotation and translation.

As illustrated in Fig. 2, EventGeoPose consists of an event representation module, a spiking network, decoupled detection and pose heads, and a training-time geometric regularizer. First, Pose-Referenced Adaptive Event Encoding (PRAEE) converts the asynchronous event stream into a fixed-shape representation $\mathbf{X} \in \mathbb{R}^{T\times 3\times H\times W}$ using activity-adaptive temporal discretization and duration-normalized event-rate cues. The encoded representation is then processed by a spiking backbone and decoupled heads to jointly predict object categories, bounding boxes, rotations, and translations. During training, Event-Geometry Consistency Regularization (EGCR) further constrains the predicted object geometry using event-derived spatial evidence. EGCR is removed at inference and therefore introduces no additional inference branch or post-processing.

3.2 SPIKING NEURONS

Compared with conventional ANNs, SNNs enable sparse spike-driven computation and provide a more energy-efficient computing paradigm for event-based perception. However, standard Leaky Integrate-and-Fire (LIF) neurons Maass (1997) quantize membrane potentials into binary spikes, which can cause substantial information loss in continuous regression. We therefore construct our spiking backbone Sapkota et al. (2025) using the I-LIF neuron introduced in SpikeYOLO Luo et al. (2024), which employs bounded integer-valued activations during training to improve numerical fidelity while retaining binary spike-driven inference.

At timestep k, the membrane potential $\mathbf{u}_k$ and integer activation $\mathbf{s}_k$ are updated as

$$\mathbf{u}_0 = \mathbf{x}_0, \tag{2}$$

$$\mathbf{u}_k = \lambda\left(\mathbf{u}_{k-1} - \mathrm{sg}(\mathbf{s}_{k-1})\right) + \mathbf{x}_k, \qquad k > 0, \tag{3}$$

$$\mathbf{s}_k = \mathrm{round}(\mathrm{clip}(\mathbf{u}_k, 0, D))\,, \tag{4}$$

where $\mathbf{x}_k$ denotes the synaptic input, λ is the membrane decay factor, and D is the maximum integer activation. The stop-gradient operator $\mathrm{sg}(\cdot)$ acts as the identity in the forward pass while blocking gradients through the reset term. Following Luo et al. (2024), we set $\lambda = 0.25$ and $D = 4$. The resulting $(D + 1)$-level activations provide finer activation granularity than binary LIF neurons, reducing quantization error. Since rounding is non-differentiable, we employ a straight-through

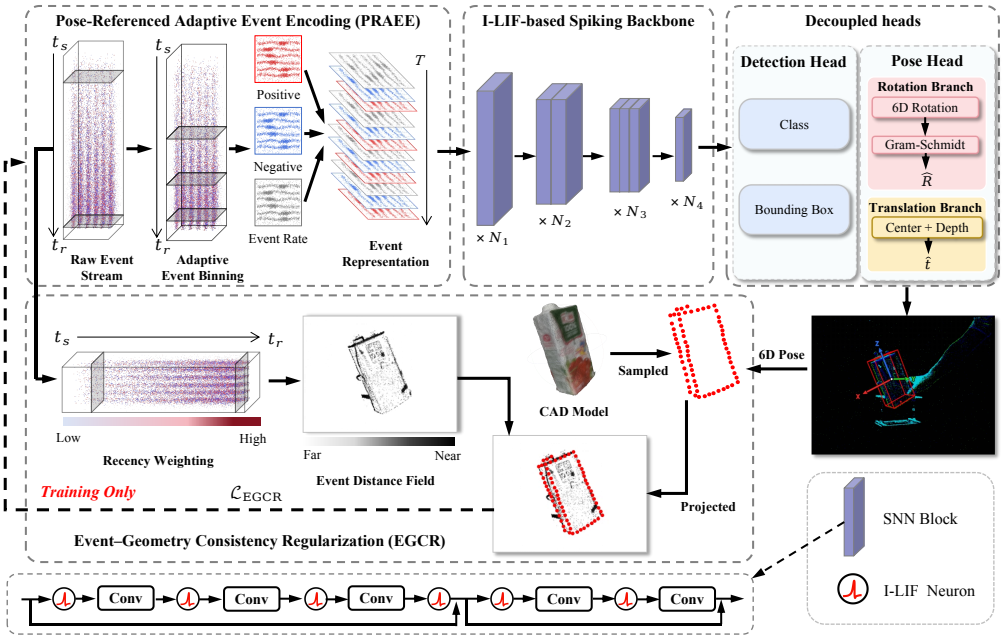


Figure 2: **Overall architecture of EventGeoPose.** PRAEE encodes raw events for an I-LIF-based spiking backbone with decoupled detection and pose heads, while training-only EGCR enforces consistency between projected CAD geometry and event-derived evidence.

estimator (STE) during backpropagation:

$$\frac{\partial s_{k,i}}{\partial u_{k,i}} \approx \mathbb{I}\left[0 \le u_{k,i} \le D\right]. \tag{5}$$

During inference, each integer activation $\mathbf{s}_k^l$ is decomposed into D binary spikes along a virtual sub-timestep dimension:

$$\mathbf{b}_{k,d}^l = \mathbb{I}\left[d < \mathbf{s}_k^l\right], \qquad d = 0, \ldots, D-1, \tag{6}$$

By construction, $\mathbf{s}_k^l = \sum_{d=0}^{D-1} \mathbf{b}_{k,d}^l$. Therefore, for a linear synaptic operator $\mathcal{W}^l(\cdot)$, $\mathcal{W}^l\left(\mathbf{s}_k^l\right) = \sum_{d=0}^{D-1} \mathcal{W}^l\left(\mathbf{b}_{k,d}^l\right)$. This formulation reduces quantization-induced information loss during training while preserving sparse spike-driven accumulation at inference.

3.3 Pose-Referenced Adaptive Event Encoding (PRAEE)

Although each event packet is collected within a fixed observation window $[t_s, t_r]$, its effective temporal support $[t_1, t_N]$ and activity distribution vary substantially with object motion, scene texture, and illumination. Fixed-duration discretization therefore assigns temporal resolution independently of the underlying event activity. During high-activity periods, a single bin may span multiple pose states and cause motion-state mixing, whereas low-activity periods may provide insufficient structural evidence. To address this imbalance, we propose PRAEE, which combines activity-adaptive temporal discretization with duration-normalized local event-rate cues.

Given N temporally ordered events, PRAEE partitions them into T sequential event bins. When $N \ge T$, we write $N = qT + r, \quad 0 \le r < T$, and define the number of events in each bin as

$$n_1 = q + r, \qquad n_t = q, \qquad t = 2, \ldots, T. \tag{7}$$

The remainder r is assigned to the earliest bin, yielding a pose-referenced partition in which bins closest to t_r maintain a consistent event budget q. Defining the cumulative boundaries as $a_0 = 0, \quad a_t = \sum_{j=1}^{t} n_j$, the index set of the t-th bin is $\mathcal{I}_t = \{a_{t-1} + 1, \ldots, a_t\}$, Because the events are temporally ordered, $\mathcal{I}_T$ contains those closest to t_r. Consequently, high event activity naturally

produces shorter physical intervals, whereas lower activity results in longer intervals, allowing the effective temporal resolution to adapt to scene dynamics. When $0 < N < T$, the events are assigned chronologically to the latest N bins and the preceding bins are left empty, when $N = 0$, the entire representation is set to zero. For each bin, positive and negative events are accumulated separately:

$$C_t^+(x,y) = \frac{1}{\kappa}\min\Big(\kappa, \sum_{i\in\mathcal{I}_t} \mathbb{I}[p_i = +1]\mathbb{I}[x_i = x]\mathbb{I}[y_i = y]\Big), \tag{8}$$

$$C_t^-(x,y) = \frac{1}{\kappa}\min\Big(\kappa, \sum_{i\in\mathcal{I}_t} \mathbb{I}[p_i = -1]\mathbb{I}[x_i = x]\mathbb{I}[y_i = y]\Big), \tag{9}$$

where $\kappa = 255$ is the clipping threshold. The two channels preserve the complementary spatial responses of positive and negative events.

Since activity-adaptive partitioning produces bins with different physical durations, event counts alone cannot distinguish rapid event bursts from slower accumulation over longer intervals. We therefore introduce a duration-normalized local event-rate map. For a non-empty bin, its temporal span is defined as $\Delta_t = \max_{i\in\mathcal{I}_t} t_i - \min_{i\in\mathcal{I}_t} t_i$. Let $A_t(x,y)$ denote the polarity-agnostic raw event count map,

$$A_t(x,y) = \sum_{i\in\mathcal{I}_t} \mathbb{I}[x_i = x]\,\mathbb{I}[y_i = y]. \tag{10}$$

Local event activity is aggregated using a 3×3 mean operator $\mathcal{U}_3$ and normalized by the bin duration:

$$R_t(x,y) = \frac{\mathcal{U}_3(A_t)(x,y)}{\Delta_t + \epsilon}, \tag{11}$$

where $\epsilon = 10^{-6}$ ensures numerical stability. For empty bins, R_t is set to zero. To compress the dynamic range, logarithmic scaling followed by window-wise maximum normalization is applied:

$$Q_t(x,y) = \frac{\log(1 + R_t(x,y))}{\log(1 + R_{\max})},\ R_{\max} = \max_{j,u,v} R_j(u,v). \tag{12}$$

For $R_{\max} = 0$, we set $Q_t(x,y) = 0$, consistent with the zero representation for an empty event window. The resulting rate map incorporates duration-dependent local activity cues that complement the event-count channels under variable bin durations.

Finally, the representation at the t-th timestep is $\mathbf{X}_t = \big[C_t^+, C_t^-, Q_t\big] \in \mathbb{R}^{3\times H\times W}$, and the complete event representation is

$$\mathbf{X} = \{\mathbf{X}_t\}_{t=1}^{T} \in \mathbb{R}^{T\times3\times H\times W}. \tag{13}$$

By jointly adapting temporal resolution to event activity and incorporating duration-normalized local event-rate cues, PRAEE mitigates motion-state mixing and provides a more balanced representation of pose-relevant event dynamics.

3.4 Event-Geometry Consistency Regularization (EGCR)

Direct pose regression Maji et al. (2024); Ousalah et al. (2026); Çetiner & Ekenel (2026) is typically supervised in the pose-parameter or 3D model space, without explicitly exploiting the geometric evidence contained in event observations. Although event streams provide only sparse and incomplete object structures, their responses around moving edges and contours still offer informative geometric cues. We therefore propose EGCR, which aligns the projected object geometry under the predicted pose with an event-derived image-space geometry field, providing complementary observation-aware geometric supervision for pose estimation.

Recency-aware event geometry field. Since the target pose corresponds to the reference time t_r, events closer to t_r provide more relevant geometric evidence. For a non-empty event packet ($N > 0$), we assign each event $e_i \in \mathcal{E}$ a recency weight $w_i = \exp\big(-\frac{t_r - t_i}{\tau}\big)$, where the decay scale is determined by the temporal span from the earliest observed event to the pose reference time $\tau = \max\big(\frac{t_r - t_1}{3}, \epsilon_t\big)$, where ϵ_t is a small constant for numerical stability. Ignoring event polarity, the weighted events are accumulated into an event evidence map

$$A_{\mathrm{E}}(x,y) = \sum_{i=1}^{N} w_i\,\mathbb{I}[x_i = x]\,\mathbb{I}[y_i = y]. \tag{14}$$

After thresholding and morphological dilation, the resulting event-support region is denoted by Ω_{E}. If $\Omega_{\mathrm{E}} \neq \varnothing$, we construct an event distance field $\mathcal{D}_{\mathrm{E}}(\mathbf{u}) = \min_{\mathbf{v}\in\Omega_{\mathrm{E}}} \|\mathbf{u}-\mathbf{v}\|_2$. Which assigns a lower cost to image locations that are better supported by the observed event structures. For an empty event packet or an empty event-support region, EGCR is skipped and $\mathcal{L}_{\mathrm{EGCR}}$ is set to zero.

Event-geometry projection consistency. For an object category c, let $\mathcal{M}_c = \{\mathbf{q}_j \in \mathbb{R}^3\}_{j=1}^{N_c}$ denote its CAD model points in the object coordinate system. During training, visible geometric edges under the ground-truth viewpoint are extracted offline using the ground-truth pose and scene depth. From these edges, a set of J visible model points $\mathcal{Q}_c^{\mathrm{vis}} = \{\mathbf{q}_j\}_{j=1}^{J}$ is sampled. Under the predicted pose $(\hat{\mathbf{R}}, \hat{\mathbf{t}})$, each point is projected onto the image plane as $\hat{\mathbf{u}}_j = \pi\left(\mathbf{K}\left(\hat{\mathbf{R}}\mathbf{q}_j + \hat{\mathbf{t}}\right)\right)$, where $\mathbf{K}$ denotes the camera intrinsic matrix and $\pi(\cdot)$ denotes perspective projection. Only points with positive depth and valid image coordinates are retained, with their index set denoted by $\mathcal{V}$.

The geometric discrepancy of each projected point is obtained by bilinearly sampling the event distance field $d_j = \mathcal{D}_{\mathrm{E}}(\hat{\mathbf{u}}_j)$. To reduce the influence of incomplete event responses and unmatched projections, we employ a truncated robust penalty:

$$\mathcal{L}_{\mathrm{EGCR}} = \frac{1}{\max(1, |\mathcal{V}|)} \sum_{j\in\mathcal{V}} \sqrt{\left(\frac{\min(d_j, d_{\max})}{d_{\max}}\right)^2 + \epsilon_c^2}. \tag{15}$$

Here, $d_{\max}$ truncates large projection errors, while bilinear sampling allows gradients to propagate through the projected coordinates to the predicted rotation and translation. EGCR thus provides observation-level geometric supervision complementary to conventional pose losses.

To avoid unreliable geometric guidance from inaccurate predictions during early training, EGCR is activated after an initial warm-up stage and weighted by $\lambda_{\mathrm{EGCR}}(e)$, which is subsequently increased linearly with the training epoch e. Since EGCR is used only during training, it introduces no additional learnable parameters or inference-time computation.

4 Experiment

4.1 Datasets

We evaluate our method on three event-based 6D object pose datasets, including the public synthetic benchmark YCB-Ev SD Rojtberg & Kühn (2025), and two datasets introduced in this work, EVA-6D-Syn and EVA-6D-Real, with additional details provided in the Appendix B. ***YCB-Ev SD*** is a synthetic dataset derived from the 21 YCB-Video object categories Xiang et al. (2017), with event observations generated from simulated camera motion. We use it as a public benchmark under relatively standard scene conditions. ***EVA-6D-Syn*** is constructed to cover more diverse dynamic and illumination conditions. High-frame-rate sequences are rendered with BlenderProc and Blender Cycles and converted into events using v2e Hu et al. (2021). It contains 10,000 sequences over 21 YCB-Video objects, including camera motion, object motion, coupled motion, low-light scenes, dynamic illumination, and high-speed motion, with synchronized RGB, depth, masks, and pose annotations. ***EVA-6D-Real*** provides real-sensor evaluation using an RGB-D camera, a DVXplorer Micro event camera, and a 240 Hz motion-capture system, as shown in Fig. 3. It contains 6 objects, 12 motion sequences, and 13,128 synchronized samples under normal and low-light conditions. Dense 6D pose annotations are obtained by combining RGB-D keyframe annotation with calibrated motion-capture trajectory propagation.

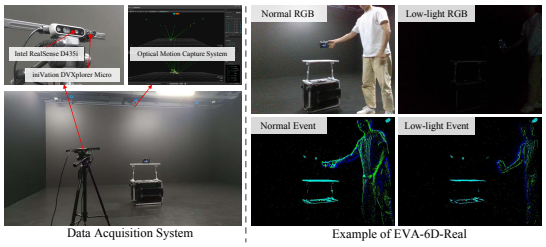


Figure 3: **Overview of the EVA-6D-Real dataset.**

4.2 Implementation Details

All experiments are conducted on a server with an Intel Xeon Gold 6226R CPU and 8 NVIDIA RTX A6000 GPUs. Models are trained for 300 epochs at an input resolution of 640×480 with a

Table 1: **Class-wise quantitative comparison on EVA-6D-Syn (%).** TS denotes Time Surfaces. Best and second-best results are highlighted.

Object	**SAM6D**	**BoxDreamer**	**FoundationPose**	**PoseGAM**	**YOLO6D-Pose**	**GDRNPP**	**Ours-TS**	**Ours**
Venue	CVPR'24	ICCV'25	CVPR'24	CVPR'26	3DV'24	TPAMI'25	–	–
Paradigm	Feature	Feature	Template	Template	Direct	Direct	Direct	Direct
Input	TS + Depth	TS	TS + Depth	TS	TS	TS	TS	Event
master_chef_can	59.0	6.7	17.5	2.9	80.9	65.9	85.4	86.0
cracker_box	67.3	17.5	38.2	12.4	75.3	84.0	79.5	80.2
sugar_box	33.7	10.4	12.5	8.4	39.2	53.0	56.9	67.2
tomato_soup_can	28.1	5.0	10.8	0.7	38.3	47.2	49.2	62.1
mustard_bottle	49.0	4.5	68.3	13.0	41.8	53.7	62.2	67.8
tuna_fish_can	27.6	0.5	35.4	0.0	21.8	26.1	35.0	39.8
pudding_box	50.7	0.1	24.9	1.2	60.7	69.5	65.0	69.7
gelatin_box	24.2	2.6	23.5	0.1	41.2	46.0	44.2	51.2
potted_meat_can	56.8	0.9	13.2	0.4	42.6	46.4	49.5	55.6
banana	58.2	7.6	66.4	2.7	55.0	49.8	61.9	64.6
pitcher_base	74.5	21.7	49.5	29.4	76.9	87.6	82.8	82.0
bleach_cleanser	69.3	16.4	76.3	21.3	57.1	67.5	69.5	72.5
bowl	57.2	1.2	38.6	4.0	84.6	69.3	85.0	87.8
mug	72.5	0.1	47.0	0.4	76.2	64.9	77.6	83.7
power_drill	79.4	5.7	67.6	3.1	73.2	79.4	76.7	77.5
wood_block	47.9	13.4	42.1	11.5	69.8	59.4	77.1	79.2
scissors	69.4	6.2	57.9	0.4	54.3	51.7	57.6	67.6
large_marker	44.8	1.2	37.5	0.0	51.2	34.6	54.2	64.4
large_clamp	69.7	2.3	68.9	0.8	71.6	74.2	74.2	80.1
extra_large_clamp	79.5	3.8	71.9	3.4	77.1	81.5	83.2	83.3
foam_brick	55.7	1.7	2.3	0.0	59.3	31.3	60.4	71.9
Mean	55.9	6.2	41.4	5.5	59.4	59.2	66.1	71.2

total batch size of 64. We use MuSGD with an initial learning rate of 0.01, momentum of 0.9, and weight decay of 5×10^{-4}. After a 3-epoch warm-up, the learning rate is linearly decayed to 1% of its initial value. Image-space augmentations that may disturb the correspondence between event observations and pose annotations, including HSV perturbation, geometric transformations, Mosaic, and Copy-Paste, are disabled. The depth scale is set to $\gamma = 100$, and $J = 256$ visible model-edge points are sampled for EGCR. EGCR is activated after 10 epochs, with λ_{EGCR} linearly increased to 0.1 over the following 10 epochs. Unless otherwise specified, all variants of our method use the same training configuration, while comparison methods follow the recommended settings of their public implementations. Pose accuracy is evaluated using ADD(-S)$_{0.5d}$ Rojtberg & Kühn (2025).

4.3 Comparison with State-of-the-Art Methods

Since there is currently no unified benchmark for event-based 6D pose estimation, we compare EventGeoPose with six representative methods spanning feature matching (SAM6D Lin et al. (2024), BoxDreamer Yu et al. (2025)), template matching (FoundationPose Wen et al. (2024), PoseGAM Chen et al. (2026)), and direct regression (YOLO6D-Pose Maji et al. (2024), GDRNPP Liu et al. (2025c)). As these methods are designed for frame-based inputs, event streams are converted into Time Surfaces, with depth additionally provided to SAM6D and FoundationPose as required. Except for YOLO6D-Pose and our method, the remaining baselines are given ground-truth object masks to minimize localization errors. We also report Ours-TS using the same Time Surface representation to isolate the effect of PRAEE.

Results on EVA-6D-Syn. Table 1 reports the class-wise results on EVA-6D-Syn. Among the baselines, the direct-regression methods perform best, with YOLO6D-Pose and GDRNPP reaching 59.4% and 59.2%, respectively. EventGeoPose achieves 71.2%, outperforming the best baseline by 11.8 percentage points and ranking first on 13 of the 21 categories. With the same Time Surface input as the baselines, Ours-TS reaches 66.1%, exceeding YOLO6D-Pose and GDRNPP by 6.7 and 6.9 points. Replacing Time Surface with PRAEE further improves accuracy by 5.1 points, confirming the benefit of the proposed event representation. Qualitatively, Fig. 4 shows that YOLO6D-Pose suffers from false and missed detections, while GDRNPP, despite using ground-truth masks, exhibits noticeable rotation errors; our method provides more accurate localization and pose alignment.

Results on EVA-6D-Real. Table 2 presents the results on EVA-6D-Real, where all methods degrade substantially due to real sensor noise, complex motion, and illumination variation. YOLO6D-Pose and GDRNPP remain the strongest baselines at 22.2% and 20.7%, respectively. Ours-TS reaches 27.2%, while the full model further improves to 34.7%, outperforming the best baseline by 12.5

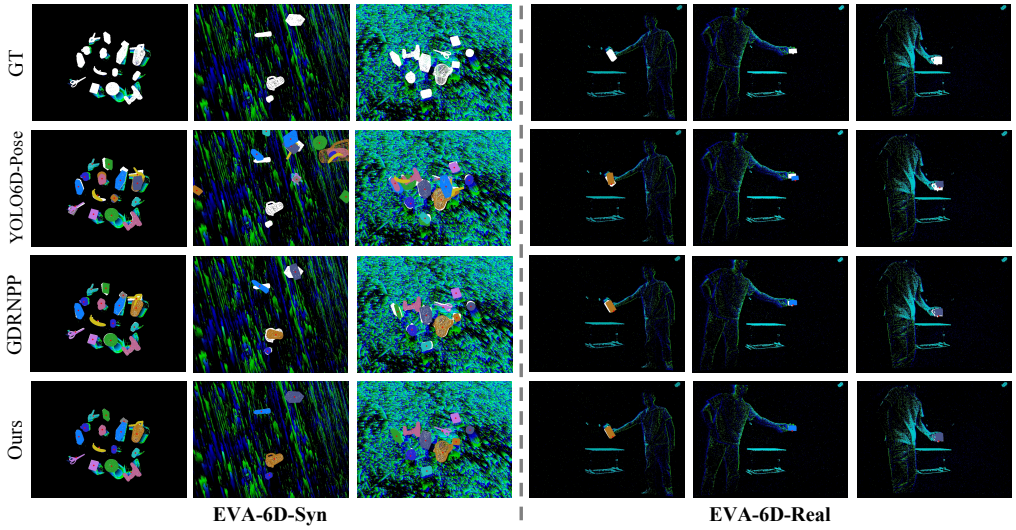


Figure 4: **Qualitative comparison on EVA-6D-Syn and EVA-6D-Real.**

Table 2: **Class-wise quantitative comparison on EVA-6D-Real (%).** TS denotes Time Surfaces. Best and second-best results are highlighted.

Object	**SAM6D**	**BoxDreamer**	**FoundationPose**	**PoseGAM**	**YOLO6D-Pose**	**GDRNPP**	**Ours-TS**	**Ours**
Venue	CVPR'24	ICCV'25	CVPR'24	CVPR'26	3DV'24	TPAMI'25	–	–
Paradigm	Feature	Feature	Template	Template	Direct	Direct	Direct	Direct
Input	TS + Depth	TS	TS + Depth	TS	TS	TS	TS	Event
juice_box	25.7	1.2	19.9	1.1	35.2	23.6	43.0	49.7
black_tea	22.6	2.6	14.2	1.2	34.9	37.5	35.4	41.0
green_tea	16.7	0.0	9.1	0.0	19.4	23.4	31.2	41.7
camera_box	14.9	0.0	7.8	0.0	17.3	12.4	24.1	29.6
elephant	15.1	0.2	6.3	0.0	17.6	20.8	22.3	34.7
cup	4.6	0.0	2.2	0.0	8.8	6.7	7.4	11.3
Mean	16.6	0.7	9.9	0.4	22.2	20.7	27.2	34.7

points and achieving the best result on all six objects. The 7.5-point gain over Ours-TS further demonstrates the benefit of PRAEE under real-world dynamics. As shown in Fig. 4, YOLO6D-Pose exhibits localization offsets and GDRNPP shows evident rotation errors despite using ground-truth masks, whereas our method achieves better pose alignment without such priors.

Table 3: **Mean results on YCB-Ev SD.** Best and second-best results are highlighted.

	SAM6D	**BoxDreamer**	**FoundationPose**	**PoseGAM**	**YOLO6D-Pose**	**GDRNPP**	**Ours-TS**	**Ours**
Mean	25.1	5.9	17.5	8.9	29.8	36.1	50.2	56.8

Results on YCB-Ev SD. Table 3 reports the mean results on YCB-Ev SD, which provides a complementary synthetic benchmark with a different event distribution. EventGeoPose achieves 56.8%, outperforming the best baseline, GDRNPP (36.1%), by 20.7 percentage points. Ours-TS reaches 50.2%, and PRAEE provides a further 6.6-point improvement. These results demonstrate consistent gains across different event distributions. Full class-wise results are provided in Appendix C.2.

4.4 Ablation Studies and Analyses

All ablation studies are conducted on EVA-6D-Syn using the training configuration in Sec. 4.2.

Contributions of PRAEE and EGCR. Table 4 evaluates the individual and joint contributions of PRAEE and EGCR. The baseline spiking pose network achieves 62.5% accuracy. Adding PRAEE or EGCR individually improves the result to 65.7% and 66.1%, respectively, while combining both reaches 71.2%, an 8.7-point gain over the baseline. The joint improvement exceeds the sum of the individual gains by 1.9 points, suggesting complementary effects between representation-level temporal modeling and observation-aware geometric supervision. All variants retain 24.4 M parameters and similar theoretical inference energy (135.5-137.1 mJ), indicating negligible additional inference cost.

Analysis of PRAEE. Table 5 compares different event representations. Time Surface achieves 66.1%, uniform temporal binning improves to 69.3%, and the full PRAEE reaches 71.2%. As illustrated in Fig. 5, Time Surface tends to accumulate motion trajectories, whereas fixed-duration binning produces uneven event densities under non-uniform motion. PRAEE instead adapts temporal resolution to event activity, reducing motion-state mixing and balancing evidence across timesteps. Removing the duration-normalized event-rate channel decreases accuracy to 67.6%, showing that equal-event partitioning alone is insufficient. The event-rate channel provides complementary duration-dependent activity cues for the variable-length bins, explaining the benefit of combining adaptive discretization with event-rate modeling.

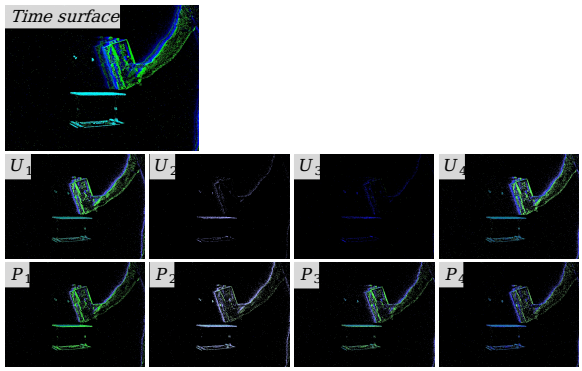


Figure 5: **Comparison of event representations under high-speed non-uniform motion.** From top to bottom: Time Surface, uniform temporal binning, and PRAEE.

Table 4: **Contributions of PRAEE and EGCR.**

Method	PRAEE	EGCR	Params (M)	Energy (mJ)	ADD(-S) (%)
Baseline	–	–	24.4	136.4	62.5
+ PRAEE	✓	–	24.4	137.1	65.7
+ EGCR	–	✓	24.4	136.0	66.1
Full (Ours)	✓	✓	24.4	**135.5**	**71.2**

Table 5: **Ablation analysis of PRAEE.**

Event Representation	Energy (mJ)	ADD(-S) (%)
Time Surface	136.0	66.1
Uniform Temporal	**129.3**	69.3
PRAEE w/o Event Rate	134.0	67.6
PRAEE (Ours)	135.5	**71.2**

Effect of Spiking Neuron Design. Table 6 compares different neuron formulations in terms of pose accuracy and theoretical inference energy. The ANN variant achieves the highest accuracy of 76.6%, but requires 866.7 mJ, approximately 6.4× that of I-LIF. Standard LIF reduces the energy to 153.1 mJ but achieves only 67.4%, reflecting the limited numerical fidelity of binary activations for continuous pose regression. In comparison, I-LIF reaches 71.2% with only 135.5 mJ, improving over LIF by 3.8 percentage points while further reducing energy by 17.6 mJ. These results support the use of spike-driven computation for energy-constrained event-based perception, particularly on continuously operating edge platforms such as mobile robots.

Table 6: **Ablation of different neurons.**

Neuron	Energy (mJ)	ADD(-S) (%)
ANN	866.7	**76.6**
LIF	153.1	67.4
I-LIF	**135.5**	<u>71.2</u>

5 Conclusion

This paper presents EventGeoPose, a directly trained SNN framework for event-camera-based 6D object pose estimation. At the computational level, we build the spiking backbone with I-LIF neurons to improve the numerical fidelity of continuous pose regression while retaining spike-driven inference. At the representation level, PRAEE combines activity-adaptive temporal discretization with duration-normalized local event-rate cues to mitigate motion-state mixing and better capture pose-relevant event dynamics. At the supervision level, EGCR introduces complementary observation-aware geometric supervision by enforcing consistency between predicted 3D object geometry and observed event structures. We further construct two event-based 6D pose datasets, EVA-6D-Syn and EVA-6D-Real, covering diverse motion patterns, illumination conditions, and real sensor observations. Extensive experiments on the proposed datasets and a public benchmark demonstrate consistent improvements over representative methods, while ablation studies validate the effectiveness of PRAEE and EGCR. The spiking framework also achieves a favorable trade-off between pose accuracy and theoretical inference energy, supporting energy-efficient event-driven perception. Overall, these results demonstrate the potential of combining event cameras with directly trained SNNs for robust and energy-efficient 6D object pose estimation in highly dynamic environments. Future work will investigate continuous multi-object pose tracking, cross-scene and cross-object generalization, and deployment on neuromorphic hardware with hardware-level energy evaluation.

AI USE STATEMENT

Generative AI tools were used to assist with language translation, manuscript editing, and the drafting and revision of portions of the paper, including improvements to clarity, grammar, and presentation. They were also used to assist with method implementation and software code editing. All AI-assisted text and code were critically reviewed and verified by the authors. In particular, technical formulations, experimental settings, numerical results, citations, and scientific claims were checked against the implemented methods, experimental records, and original sources, while AI-assisted code was reviewed and tested for correctness before being used in the reported experiments. All final research, implementation, and writing decisions were made by the authors. The authors take full responsibility for the final content of this work, including any text, code, claims, or other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the methodological definitions and experimental details necessary to reproduce Event-GeoPose in the main paper and supplementary material. Pose parameterization and training objectives are detailed in Appendix A, while training hyperparameters, implementation settings, evaluation protocols, and baseline configurations are reported in Secs. 4.2 and 4.3. The construction of EVA-6D-Syn and EVA-6D-Real, including event generation, data acquisition, calibration, synchronization, pose annotation, and dataset preprocessing, is described in Appendix B. The ablation protocols and additional analyses are documented in Sec. 4.4 and Appendix C, providing the experimental details required to reproduce and verify the reported results.

REFERENCES

Pablo Aguirrezabal, Iker Aguinaga, and Aitor Alvarez-Gila. Monocular rgb 6d object pose estimation for augmented reality: a survey. *Virtual Real.*, 30(2):57, 2026.

Kemal Alperen Çetiner and Hazım Kemal Ekenel. Yolo-key-6d: Single stage monocular 6d pose estimation with keypoint enhancements. *arXiv preprint arXiv:2603.03879*, 2026.

Bharatesh Chakravarthi, Aayush Atul Verma, Kostas Daniilidis, Cornelia Fermuller, and Yezhou Yang. Recent event camera innovations: A survey. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 342–376, 2024.

Jianqi Chen, Biao Zhang, Xiangjun Tang, and Peter Wonka. Posegam: Robust unseen object pose estimation via geometry-aware multi-view reasoning. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 7197–7208, 2026.

Jonathan Courtois, Benoît Miramond, and Alain Pegatoquet. Spiking monocular event based 6d pose estimation for space application. *arXiv preprint arXiv:2501.02916*, 2025.

Junming Fan, Pai Zheng, Shufei Li, and Lihui Wang. An integrated hand-object dense pose estimation approach with explicit occlusion awareness for human-robot collaborative disassembly. *IEEE Trans. Autom. Sci. Eng.*, 21(1):147–156, 2022.

Yimeng Fan, Wei Zhang, Changsong Liu, Mingyang Li, and Wenrui Lu. Sfod: Spiking fusion object detector. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 17191–17200, 2024.

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 Trans. Pattern Anal. Mach. Intell.*, 44(1):154–180, 2020.

Arren Glover, Luna Gava, Zhichao Li, and Chiara Bartolozzi. Edopt: Event-camera 6-dof dynamic object pose tracking. In *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pp. 18200–18206, 2024.

Oussama Abdul Hay, Xiaoqian Huang, Abdulla Ayyad, Eslam Sherif, Randa Almadhoun, Yusra Abdulrahman, Lakmal Seneviratne, Abdulqader Abusafieh, and Yahya Zweiri. E-pose: A large scale event camera dataset for object pose estimation. *Sci. Data*, 12(1):245, 2025a.

Oussama Abdul Hay, Xiaoqian Huang, Muhammad Ahmed Humais, Abdulla Ayyad, Randa Almadhoun, and Yahya Zweiri. Learning 6d object pose estimation with event cameras using synthetic data and domain randomization. *IEEE Robot. Autom. Lett.*, 11(2):1690–1697, 2025b.

Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic dvs events. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, pp. 1312–1321, 2021.

Jae-Young Kang, Hoonhee Cho, Taeyeop Lee, Minjun Kang, Bowen Wen, Youngho Kim, and Kuk-Jin Yoon. Event6d: Event-based novel object 6d pose tracking. *arXiv preprint arXiv:2603.28045*, 2026.

Aupendu Kar, Vishnu Raj, and Guan-Ming Su. Event camera guided visual media restoration and 3d reconstruction: a survey. *APSIPA Trans. Signal Inf. Process.*, 15(1):110–162, 2026.

Yann Labbé, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, and Josef Sivic. Megapose: 6d pose estimation of novel objects via render & compare. *arXiv preprint arXiv:2212.06870*, 2022.

Taeyeop Lee, Bowen Wen, Minjun Kang, Gyuree Kang, In So Kweon, and Kuk-Jin Yoon. Any6d: Model-free 6d pose estimation of novel objects. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 11633–11643, 2025.

Zhichao Li, Nicola A Piga, Franco Di Pietro, Massimiliano Iacono, Arren Glover, Lorenzo Natale, and Chiara Bartolozzi. Hybrid object tracking with events and frames. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, pp. 9057–9064, 2023.

Zhichao Li, Chiara Bartolozzi, Lorenzo Natale, and Arren Glover. Event-based motion & appearance fusion for 6d object pose tracking. *arXiv preprint arXiv:2603.08264*, 2026.

Jiehong Lin, Lihua Liu, Dekun Lu, and Kui Jia. Sam-6d: Segment anything model meets zero-shot 6d object pose estimation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 27906–27916, 2024.

Daikun Liu, Lei Cheng, Teng Wang, and Changyin Sun. Edcflow: Exploring temporally dense difference maps for event-based optical flow estimation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1984–1993, 2025a.

Jian Liu, Wei Sun, Chongpei Liu, Xing Zhang, and Qiang Fu. Robotic continuous grasping system by shape transformer-guided multiobject category-level 6-d pose estimation. *IEEE Trans. Ind. Informat.*, 19(11):11171–11181, 2023a.

Xingyu Liu, Gu Wang, Ruida Zhang, Chenyangguang Zhang, Federico Tombari, and Xiangyang Ji. Unopose: Unseen object pose estimation with an unposed rgb-d reference image. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 22023–22034, 2025b.

Xingyu Liu, Ruida Zhang, Chenyangguang Zhang, Gu Wang, Jiwen Tang, Zhigang Li, and Xiangyang Ji. Gdrnpp: A geometry-guided and fully learning-based object pose estimator. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(7):5742–5759, 2025c.

Yun Liu, Xiaomeng Xu, Weihang Chen, Haocheng Yuan, He Wang, Jing Xu, Rui Chen, and Li Yi. Enhancing generalizable 6d pose tracking of an in-hand object with tactile sensing. *IEEE Robot. Autom. Lett.*, 9(2):1106–1113, 2023b.

Zibin Liu, Banglei Guan, Yang Shang, Qifeng Yu, and Laurent Kneip. Line-based 6-dof object pose estimation and tracking with an event camera. *IEEE Trans. Image Process.*, 33:4765–4780, 2024.

Xinhao Luo, Man Yao, Yuhong Chou, Bo Xu, and Guoqi Li. Integer-valued training and spike-driven inference spiking neural network for high-performance and energy-efficient object detection. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 253–272, 2024.

Wolfgang Maass. Networks of spiking neurons: the third generation of neural network models. *Neural Netw.*, 10(9):1659–1671, 1997.

Debapriya Maji, Soyeb Nagori, Manu Mathew, and Deepak Poddar. Yolo-6d-pose: Enhancing yolo for single-stage monocular multi-object 6d pose estimation. In *Proc. Int. Conf. 3D Vis. (3DV)*, pp. 1616–1625, 2024.

Mayura Manawadu, Sieun Park, and Soon-Yong Park. Advancing 6d pose estimation in augmented reality–overcoming projection ambiguity with uncontrolled imagery. *arXiv preprint arXiv:2403.13434*, 2024.

Sungphill Moon, Hyeontae Son, Dongcheol Hur, and Sangwook Kim. Co-op: Correspondence-based novel object pose estimation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 11622–11632, 2025.

Tam Ngoc-Bang Nguyen, Mohsi Jawaid, and Tat-Jun Chin. Robust pnp on a neuromorphic processor for object pose estimation. *arXiv preprint arXiv:2607.16834*, 2026.

Van Nguyen Nguyen, Thibault Groueix, Mathieu Salzmann, and Vincent Lepetit. Gigapose: Fast and robust novel object pose estimation via one correspondence. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 9903–9913, 2024.

Nassim Ali Ousalah, Peyman Rostami, Vincent Gaudillière, Emmanuel Koumandakis, Anis Kacem, Enjie Ghorbel, and Djamila Aouada. Cov2pose: Leveraging spatial covariance for direct manifold-aware 6-dof object pose estimation. *arXiv preprint arXiv:2603.19961*, 2026.

Pavel Rojtberg and Julius Kühn. Ycb-ev sd: Synthetic event-vision dataset for 6dof object pose estimation. *arXiv preprint arXiv:2511.11344*, 2025.

Pavel Rojtberg and Thomas Pöllabauer. Ycb-ev: Event-vision dataset for 6 dof object pose estimation. In *Proc. Int. Conf. Pattern Recognit. Artif. Intell.*, pp. 449–456, 2024.

Ranjan Sapkota, Rahul Harsha Cheppally, Ajay Sharda, and Manoj Karkee. Yolo26: Key architectural enhancements and performance benchmarking for real-time object detection. *arXiv preprint arXiv:2509.25164*, 2025.

Qiaoyi Su, Yuhong Chou, Yifan Hu, Jianing Li, Shijie Mei, Ziyang Zhang, and Guoqi Li. Deep directly-trained spiking neural networks for object detection. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 6532–6542, 2023.

Zhe Wang, Qijin Song, Zihao Li, Jingyu Xiao, and Weibang Bai. Keypoint-based dynamic object 6-dof pose tracking via event camera. *arXiv preprint arXiv:2604.23387*, 2026.

Bowen Wen, Wenzhao Lian, Kostas Bekris, and Stefan Schaal. Catgrasp: Learning category-level task-relevant grasping in clutter from simulation. In *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pp. 6401–6408, 2022.

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 17868–17879, 2024.

Wanqing Xia, Hao Zheng, Weiliang Xu, and Xun Xu. Large vision-language models enabled novel objects 6d pose estimation for human-robot collaboration. *Robot. Comput.-Integr. Manuf.*, 95: 103030, 2025.

Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017.

Yan Yang, Liyuan Pan, Dongxu Li, and Liu Liu. Ezsr: Event-based zero-shot recognition. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 4628–4638, 2025.

Man Yao, JiaKui Hu, Tianxiang Hu, Yifan Xu, Zhaokun Zhou, Yonghong Tian, Bo Xu, and Guoqi Li. Spike-driven transformer v2: Meta spiking neural network architecture inspiring the design of next-generation neuromorphic chips. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, volume 2024, pp. 52885–52907, 2024.

Yuanhong Yu, Xingyi He, Chen Zhao, Junhao Yu, Jiaqi Yang, Ruizhen Hu, Yujun Shen, Xing Zhu, Xiaowei Zhou, and Sida Peng. Boxdreamer: Dreaming box corners for generalizable object pose estimation. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 9374–9384, 2025.

Pengjie Zhang, Lin Zhu, Xiao Wang, Lizhi Wang, and Hua Huang. Ematch: A unified framework for event-based optical flow and stereo matching. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 5845–5855, 2025.

Yan Zhao, Shaobo Zhang, Wanqing Zhao, Ying Wei, and Jinye Peng. Augmented reality system based on real-time object 6d pose estimation. In *Proc. Int. Conf. Image Process. Media Comput. (ICIPMC)*, pp. 27–34, 2023.

A Pose Parameterization and Loss Function

A.1 Pose Parameterization

To enable stable regression of the 6D pose, we decouple pose parameterization into a rotation and a translation branch Maji et al. (2024), as illustrated in Fig. 2. For rotation, the network predicts a continuous 6D representation $[\hat{\mathbf{a}}_1, \hat{\mathbf{a}}_2] \in \mathbb{R}^6$, which is converted into a valid rotation matrix through Gram-Schmidt orthogonalization:

$$\hat{\mathbf{r}}_1 = \frac{\hat{\mathbf{a}}_1}{\|\hat{\mathbf{a}}_1\|_2}, \tag{16}$$

$$\tilde{\mathbf{r}}_2 = \hat{\mathbf{a}}_2 - (\hat{\mathbf{r}}_1^\top \hat{\mathbf{a}}_2)\hat{\mathbf{r}}_1, \qquad \hat{\mathbf{r}}_2 = \frac{\tilde{\mathbf{r}}_2}{\|\tilde{\mathbf{r}}_2\|_2}, \tag{17}$$

$$\hat{\mathbf{r}}_3 = \hat{\mathbf{r}}_1 \times \hat{\mathbf{r}}_2, \qquad \hat{\mathbf{R}} = [\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2, \hat{\mathbf{r}}_3]. \tag{18}$$

For translation, the network predicts a projected-center offset $\hat{\boldsymbol{\delta}}_{xy}$ and a depth parameter $\hat{\delta}_z$. Given the feature-grid location $\mathbf{g} = (g_x, g_y)^\top$ and feature stride s, the projected object center $\hat{\mathbf{c}} = (\hat{u}, \hat{v})^\top$ and normalized depth are decoded as

$$\hat{\mathbf{c}} = \begin{bmatrix} \hat{u} \\ \hat{v} \end{bmatrix} = s\left(\hat{\boldsymbol{\delta}}_{xy} + \mathbf{g} - \frac{1}{2}\right), \tag{19}$$

$$\hat{z}_\ell = \exp\left[\mathrm{clip}(\hat{\delta}_z, -10, 6.9)\right]. \tag{20}$$

Given the camera intrinsics (f_x, f_y, c_x, c_y) and depth scale factor γ, the 3D translation $\hat{\mathbf{t}} = (\hat{T}_x, \hat{T}_y, \hat{T}_z)^\top$ is recovered as

$$\hat{T}_z = \gamma \hat{z}_\ell, \quad \hat{T}_x = \frac{(\hat{u} - c_x)\hat{T}_z}{f_x}, \quad \hat{T}_y = \frac{(\hat{v} - c_y)\hat{T}_z}{f_y}. \tag{21}$$

A.2 Loss Function

The network is jointly optimized for object detection and pose estimation. The detection objective Luo et al. (2024); Sapkota et al. (2025) consists of bounding-box regression, classification, and distribution focal losses, denoted by $\mathcal{L}_{\mathrm{box}}$, $\mathcal{L}_{\mathrm{cls}}$, and $\mathcal{L}_{\mathrm{dfl}}$, respectively. Let $\mathcal{P}$ denote the set of positive predictions, $g(n)$ the ground-truth instance assigned to prediction n, and $\mathcal{P}_v \subseteq \mathcal{P}$ the subset associated with valid 6D pose annotations. All pose-specific losses below are computed over $\mathcal{P}_v$. In the datasets used in this work, valid 6D pose annotations are available for all instances used for pose supervision. If $\mathcal{P}_v = \varnothing$, the pose-specific losses are set to zero.

For rotation, we supervise the first two columns of the predicted rotation matrix:

$$\mathcal{L}_R = \frac{1}{6|\mathcal{P}_v|} \sum_{n \in \mathcal{P}_v} \left(\|\hat{\mathbf{r}}_{1,n} - \mathbf{r}^*_{1,g(n)}\|_1 + \|\hat{\mathbf{r}}_{2,n} - \mathbf{r}^*_{2,g(n)}\|_1\right), \tag{22}$$

where $(\cdot)^*$ denotes the corresponding ground-truth quantity.

Translation is supervised in the projected-center and depth spaces. The center loss is normalized by the ground-truth bounding-box area $A_{g(n)}$ to reduce the influence of object-scale variation:

$$\mathcal{L}_{xy} = \frac{1}{|\mathcal{P}_v|} \sum_{n \in \mathcal{P}_v} \left[1 - \exp\left(-\frac{\|\hat{\mathbf{c}}_n - \mathbf{c}^*_{g(n)}\|_2^2}{4\sigma^2 A_{g(n)} + \epsilon}\right)\right], \tag{23}$$

where $\sigma = 0.026$ controls the scale sensitivity and $\epsilon = 10^{-9}$ ensures numerical stability.

The depth loss is normalized by the CAD model diameter $d_{c_{g(n)}}$:

$$\mathcal{L}_z = \frac{1}{|\mathcal{P}_v|} \sum_{n \in \mathcal{P}_v} \frac{\gamma \left|\hat{z}_{\ell,n} - z^*_{\ell,g(n)}\right|}{d_{c_{g(n)}}}, \tag{24}$$

where $\gamma = 100$ is the depth scaling factor and $z^*_{\ell,g(n)} = T^*_{z,g(n)}/\gamma$ denotes the normalized ground-truth depth.

To jointly constrain rotation and translation in the 3D model space, we further employ a model-based geometric loss. For the matched category $c_{g(n)}$, the model $\mathcal{M}_{c_{g(n)}} = \{\mathbf{q}_i\}_{i=1}^{N_{c_{g(n)}}}$ is transformed by the predicted and ground-truth poses:

$$\hat{\mathbf{p}}_{n,i} = \hat{\mathbf{R}}_n \mathbf{q}_i + \hat{\mathbf{t}}_n, \qquad \mathbf{p}^*_{n,i} = \mathbf{R}^*_{g(n)} \mathbf{q}_i + \mathbf{t}^*_{g(n)}. \tag{25}$$

For asymmetric objects, the instance-level geometric loss is

$$\ell^{\text{asym}}_{\text{geom},n} = \frac{1}{d_{c_{g(n)}}} \sqrt{\frac{1}{N_{c_{g(n)}}} \sum_{i=1}^{N_{c_{g(n)}}} \left\| \hat{\mathbf{p}}_{n,i} - \mathbf{p}^*_{n,i} \right\|_2^2}. \tag{26}$$

For symmetric objects, we use nearest-neighbor matching:

$$\ell^{\text{sym}}_{\text{geom},n} = \frac{1}{d_{c_{g(n)}}} \sqrt{\frac{1}{N_{c_{g(n)}}} \sum_{i=1}^{N_{c_{g(n)}}} \min_j \left\| \hat{\mathbf{p}}_{n,i} - \mathbf{p}^*_{n,j} \right\|_2^2}. \tag{27}$$

The corresponding geometric loss is

$$\mathcal{L}_{\text{geom}} = \frac{1}{|\mathcal{P}_v|} \sum_{n \in \mathcal{P}_v} \ell_{\text{geom},n}, \tag{28}$$

where $\ell_{\text{geom},n}$ is selected according to the symmetry of the corresponding object.

The complete training objective is

$$\begin{aligned} \mathcal{L} = \lambda_{\text{box}}\mathcal{L}_{\text{box}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{dfl}}\mathcal{L}_{\text{dfl}} + \lambda_R\mathcal{L}_R \\ + \lambda_{xy}\mathcal{L}_{xy} + \lambda_z\mathcal{L}_z + \lambda_g\mathcal{L}_{\text{geom}} + \lambda_{\text{EGCR}}(e)\mathcal{L}_{\text{EGCR}}. \end{aligned} \tag{29}$$

where $\lambda_{\text{EGCR}}(e)$ follows the warm-up schedule defined in Sec. 3.4.

B Additional Datasets

Table 7: **Comparison of the event-based 6D pose estimation datasets used in our experiments.**

Dataset	Type	Objects	Scale	Motion Setting	Illumination
YCB-Ev SD Rojtberg & Kühn (2025)	Synthetic	21	50,000	Camera motion	Standard
EVA-6D-Syn	Synthetic	21	10,000	Camera / Object / Coupled	Normal / Low-light / Dynamic
EVA-6D-Real	Real	6	13,128	Object motion	Normal / Low-light

We evaluate EventGeoPose on the public YCB-Ev SD benchmark Rojtberg & Kühn (2025) and two event-based 6D pose datasets introduced in this work, EVA-6D-Syn and EVA-6D-Real. Their main characteristics are summarized in Table 7. YCB-Ev SD is derived from the 21 YCB-Video object categories Xiang et al. (2017) and provides synthetic event observations under simulated camera motion. In the following, we provide detailed descriptions of the construction and annotation procedures of EVA-6D-Syn and EVA-6D-Real.

B.1 EVA-6D-Syn

Generation Pipeline. EVA-6D-Syn is constructed using the 21 object categories from YCB-Video Xiang et al. (2017). High-frame-rate intensity sequences are rendered using BlenderProc and Blender Cycles and subsequently converted into asynchronous event streams using v2e Hu et al. (2021). All samples are generated at a resolution of 640×480. During rendering, object categories, object instances, object poses, camera viewpoints, backgrounds, and illumination conditions are randomized to increase scene diversity. Natural occlusion and partial visibility are also introduced through multi-object scene composition.

Motion and Illumination Settings. To provide diverse dynamic conditions, EVA-6D-Syn includes camera motion, independent object motion, and coupled camera–object motion. The dataset further incorporates high-speed motion, low-light scenes, and dynamically varying illumination to evaluate pose estimation under conditions where conventional frame observations may become unreliable. These factors are varied across sequences rather than being restricted to a single motion or illumination pattern, resulting in substantial variation in object appearance, motion magnitude, occlusion, and event activity.

Annotations and Data Modalities. EVA-6D-Syn contains 10,000 sequences covering all 21 YCB-Video object categories. Since the data are generated in simulation, accurate camera intrinsics and object poses are directly available from the rendering environment. Each sequence provides synchronized RGB images, depth maps, object masks, camera intrinsics, 6D object poses, and raw event streams, enabling both event-based pose estimation and controlled multimodal analysis.

B.2 EVA-6D-REAL

Acquisition System. EVA-6D-Real is collected using an Intel RealSense D435i RGB-D camera, an iniVation DVXplorer Micro event camera, and a $240\,\mathrm{Hz}$ optical motion-capture system. The RGB-D and event cameras remain fixed during acquisition, while objects undergo continuous 6D motion within the common field of view. The resulting sequences contain substantial variations in translation, rotation, apparent object scale, occlusion, and visibility. Each object is recorded under both normal and low-light conditions. In total, EVA-6D-Real contains 6 objects, 12 motion sequences, and 13,128 synchronized samples.

Calibration and Synchronization. Accurate 6D pose annotation requires establishing the spatial relationships among the object model, motion-capture rigid body, motion-capture coordinate system, and camera coordinate system. We therefore calibrate the corresponding rigid transformations before propagating the motion-capture trajectories to the camera frame. Temporal synchronization is additionally performed between the motion-capture system and the visual sensors so that object poses can be associated with the corresponding RGB-D and event observations. For paired low-light recordings, the spatial calibration is retained while the temporal offset is recalibrated.

Pose Annotation. Dense ground-truth poses are obtained by combining manual RGB-D keyframe annotation with calibrated motion-capture trajectory propagation. Multiple keyframes covering diverse object viewpoints are manually annotated to establish the transformation between the object model coordinate system and the tracked rigid-body coordinate system. The calibrated $240\,\mathrm{Hz}$ motion-capture trajectories are then transformed into the camera coordinate frame and interpolated to the timestamps of the recorded samples, producing dense 6D pose annotations throughout each sequence. This procedure avoids manually annotating every frame while retaining the geometric consistency provided by the motion-capture measurements.

B.3 DATASET SPLITS

For YCB-Ev SD, we follow the official training and testing protocol provided with the dataset Rojtberg & Kühn (2025). For EVA-6D-Syn, the 10,000 independently generated sequences are randomly partitioned into 6,000 training sequences and 4,000 test sequences. Since each sequence is independently synthesized with randomized scene, motion, and illumination parameters, the random sequence-level split draws the training and test sets from the same generation distribution. For EVA-6D-Real, we adopt a chronological split within each recorded sequence. Specifically, the first 60% of samples are used for training and the remaining 40% for testing, avoiding random interleaving of temporally adjacent observations between the two subsets.

C ADDITIONAL EXPERIMENTAL ANALYSES

The quantitative ablations below are conducted on EVA-6D-Syn using the same training and evaluation settings as in the main paper.

Table 8: **Sensitivity to the number of temporal steps on EVA-6D-Syn.**

Temporal Steps T	Params (M)	Energy (mJ)	ADD(-S) (%)
1	24.4	**39.5**	63.3
2	24.4	74.1	68.3
4	24.4	135.5	**71.2**
8	24.4	310.2	67.8

C.1 Sensitivity to Temporal Steps

Table 8 analyzes the effect of the number of encoding timesteps T on pose accuracy and theoretical inference energy. Unless otherwise specified, we use $T = 4$ as the default setting in all other experiments. Under a fixed event window of approximately 33-34 ms, increasing T from 1 to 4 improves the average accuracy from 63.3% to 71.2%, suggesting that finer temporal decomposition can mitigate motion-state mixing and preserve more pose-relevant temporal information. However, further increasing T to 8 raises the theoretical inference energy from $135.5\,\mathrm{mJ}$ to $310.2\,\mathrm{mJ}$ while reducing the accuracy to 67.8%. A possible explanation is that excessive temporal subdivision leaves fewer events within each timestep and may fragment local event structures, weakening pose estimation despite the increased computational cost. These results indicate that the number of timesteps should balance temporal resolution, structural integrity of event observations, and computational efficiency.

C.2 Full Class-wise Results on YCB-Ev SD

Table 9 reports the complete per-object results on YCB-Ev SD. These results complement the mean performance reported in the main paper and illustrate the variation in pose estimation accuracy across all 21 object categories.

Table 9: **Full class-wise quantitative comparison on YCB-Ev SD (%).** TS denotes Time Surfaces. Best and second-best results are highlighted.

Object	**SAM6D**	**BoxDreamer**	**FoundationPose**	**PoseGAM**	**YOLO6D-Pose**	**GDRNPP**	**Ours-TS**	**Ours**
Venue	CVPR'24	ICCV'25	CVPR'24	CVPR'26	3DV'24	TPAMI'25	–	–
Paradigm	Feature	Feature	Template	Template	Direct	Direct	Direct	Direct
Input	TS + Depth	TS	TS + Depth	TS	TS	TS	TS	Event
master_chef_can	25.1	8.6	11.7	28.1	49.9	51.1	72.3	73.5
cracker_box	36.4	4.6	16.6	2.3	57.5	67.0	62.6	66.0
sugar_box	16.6	3.6	15.1	4.1	23.0	38.2	50.0	55.5
tomato_soup_can	17.3	1.5	17.3	1.5	25.8	32.3	38.4	44.9
mustard_bottle	10.8	5.1	20.5	2.6	15.8	35.4	45.7	55.1
tuna_fish_can	10.4	0.0	8.0	0.8	20.7	17.0	14.6	11.5
pudding_box	21.0	0.0	14.4	2.4	20.9	29.5	37.7	44.8
gelatin_box	17.7	0.5	13.4	2.1	13.6	17.4	26.1	33.2
potted_meat_can	20.9	1.2	13.3	2.0	23.7	33.4	36.8	42.7
banana	19.5	7.0	22.9	7.8	12.6	24.6	36.7	44.4
pitcher_base	28.3	1.6	23.6	3.9	42.4	60.2	50.4	56.1
bleach_cleanser	33.5	5.6	20.5	1.4	20.5	42.1	53.1	61.3
bowl	20.3	0.1	11.2	19.0	52.9	43.9	72.3	75.3
mug	28.7	0.0	20.0	2.0	44.5	45.7	36.2	74.1
power_drill	40.1	0.7	21.9	4.2	42.3	48.8	55.2	62.8
wood_block	26.0	33.4	11.2	24.4	38.7	40.9	75.7	75.9
scissors	37.3	4.0	24.3	13.4	12.2	23.6	37.7	47.4
large_marker	25.2	3.4	20.4	7.6	21.5	9.8	46.3	49.5
large_clamp	35.6	13.8	22.6	23.8	23.5	35.6	69.3	73.0
extra_large_clamp	35.1	17.6	22.3	22.3	32.5	47.4	73.5	76.8
foam_brick	21.7	11.0	15.5	12.1	30.9	14.3	63.5	69.5
Mean	25.1	5.9	17.5	8.9	29.8	36.1	50.2	56.8

C.3 Qualitative Examples under Challenging Conditions

Fig. 6 presents representative RGB- and event-based pose predictions under low-light, extreme-illumination, and high-speed-motion conditions. In the shown low-light and extreme-illumination examples, severe degradation of the RGB observations is accompanied by missed or false detections, while the high-speed-motion example exhibits a larger rotation error. In these cases, the corresponding event-based predictions remain visually closer to the ground truth, illustrating the potential of event sensing to provide complementary structural and motion cues when frame-based observations are strongly degraded.

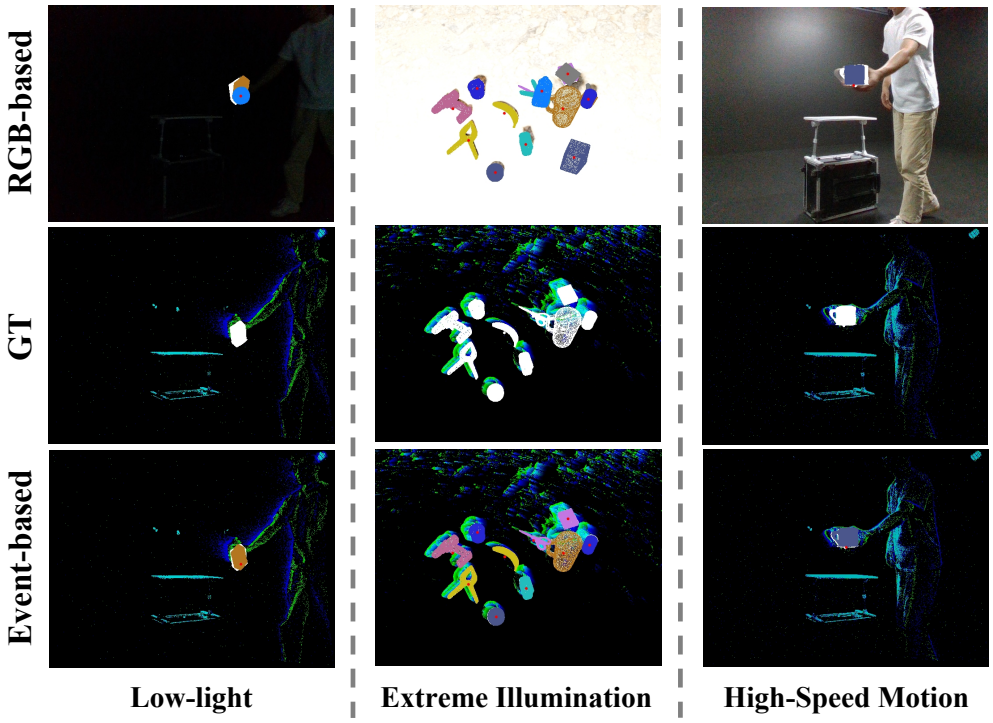


Figure 6: **Qualitative comparison of RGB- and event-based 6D pose estimation under low-light, extreme illumination, and high-speed motion conditions.** The top, middle, and bottom rows show predictions from the RGB-trained model, ground-truth poses, and predictions from the event-based model, respectively.