EXPO-SHAPE: A BENCHMARK AND EVENT-GUIDED NOVEL VIEW SYNTHESIS FOR EXPOSURE-DEGRADED OBJECT POSE ESTIMATION

Anonymous authors
Paper under double-blind review

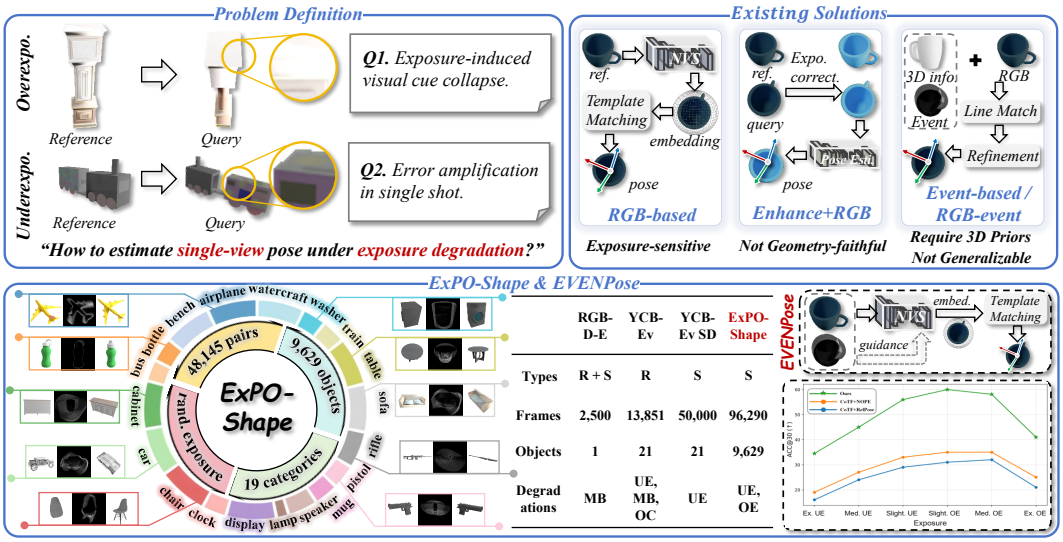


Figure 1: **Single-view pose estimation under exposure degradation is inherently difficult.** RGB-only methods are exposure-sensitive, restore-then-estimate pipelines are not geometry-faithful, and most Event/RGB-event approaches rely on 3D priors. We introduce **ExPO-Shape** (**48,145** paired samples, **9,629** objects, **19** categories, and controlled **under/over-exposure**) and **EVENPose**, a model-free event-guided novel-view synthesis baseline for exposure-robust 3D pose estimation. R and S denote the Real and Synthesized images.

ABSTRACT

Estimating single-view pose under exposure degradation requires both semantic appearance priors and illumination-robust geometric cues, neither of which can be provided by RGB or events alone. The current bottleneck of RGB–event pose estimators lies mainly in the lack of large-scale exposure-degraded benchmarks for unseen-object generalization and in the reliance on object 3D models. To fill this gap, we build **ExPO-Shape**, a large-scale **Ex**posure-degraded object **PO**se estimation benchmark comprising 48,145 paired RGB–event samples of 9,629 unique objects across 19 categories, rendered under randomly sampled under- and over-exposure conditions. To further facilitate research on model-free, generalizable RGB-event pose estimation, we propose **EVENPose**, an event-guided single-view framework for robust 3D pose estimation under exposure shifts. EVENPose stabilizes RGB embeddings with event cues instead of depth priors, guides novel-view synthesis with event-derived structure, and enforces event-level consistency to improve structural fidelity and generalization. Extensive experiments demonstrate that EVENPose serves as a strong baseline on ExPO-Shape and transfers competitively to existing RGB–event benchmarks. Overall, ExPO-Shape serves as a scalable training benchmark for model-free, generalizable RGB-event pose estimation and enables systematic evaluation of single-view robustness to exposure degradation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Estimating robust object pose from a single view is a practically significant yet rarely addressed task. Recent RGB-based methods Nguyen et al. (2024); Barroso-Laguna et al. (2024); Sun et al. (2022); He et al. (2022); Du et al. (2022); Liu et al. (2022); Wen et al. (2024); Li et al. (2023); Cheng et al. (2023); Zhou et al. (2025); Jain et al. (2023) have achieved impressive performance under standard imaging conditions and can generalize to unseen objects. In real-world scenarios, however, visual observations are often affected by exposure degradation, such as severe under-exposure or over-saturation. As shown in Figure 1 (top left), under these conditions, appearance-based visual cues become unreliable, leading to ambiguous or distorted features that can significantly degrade pose estimation accuracy Liu et al. (2024b); Dubeau et al. (2020); Rojtberg & Pöllabauer (2024). Directly cascading the enhancement method with the pose estimator improves visibility but may distort structures and accumulate errors, making pose estimates unreliable.

Event-based vision has been widely explored as a remedy for exposure degradation due to its high dynamic range and insensitivity to motion blur. Most existing event-based approaches treat events primarily as carriers of motion or geometric change, rather than as sources of rich spatial or semantic information. In practice, event streams are typically exploited either by enforcing geometric constraints derived from event-induced edges or 3D models (e.g., Liu et al. (2024a); Jawaid et al. (2022); Liu et al. (2024b)) or by estimating motion fields to track object motion incrementally and update pose over time (e.g., Chen et al. (2020)). More recent biologically inspired formulations operate directly on spatiotemporal event patterns to model motion consistency without reconstructing intensity images Zheng et al. (2022); Reverter Valeiras et al. (2016). While effective, these methods are limited by their reliance on 3D object priors, instance-level estimation, and motion-specific cues, making them unsuitable for single-view pose estimation under exposure degradation.

Pioneering studies have demonstrated the great potential of combining event and RGB data for object pose tracking Li et al. (2025); Dubeau et al. (2020); Li & Stueckler (2021). They typically exploit event streams to provide high-rate motion cues and update poses through event-based optical flow fused with low-frequency frame-based pose estimates Li et al. (2025), cascade event-to-frame refinement where an event network predicts relative pose increments that are further refined by an RGB-D tracker Dubeau et al. (2020), or generative event modeling with probabilistic inference on events followed by photometric alignment on intensity frames Li & Stueckler (2021).

While the aforementioned methods have shown promise in fusing RGB with event for object tracking, two fundamental problems arise when estimating object pose from a single view under exposure-degraded conditions: **(1) Dataset limitation.** Existing RGB–event benchmarks Rojtberg & Pöllabauer (2024); Rojtberg & Kühn (2025); Dubeau et al. (2020) are small-scale and instance-centric with limited object diversity (e.g., YCB-Ev Rojtberg & Pöllabauer (2024) with 21 objects, and RGB-D-E Dubeau et al. (2020) with only one object), which prevents systematic evaluation and learning of unseen-object generalization, especially in the challenging single-shot setting where exposure degradation severely disrupts RGB appearance cues. **(2) Strong reliance on 3D priors.** Most existing RGB–event pipelines require explicit 3D information, such as CAD models Li et al. (2025); Li & Stueckler (2021) or depth Dubeau et al. (2020), and therefore have limited practicality in real-world deployment, where such 3D cues are often unavailable and may further degrade or become unreliable under extreme under- and over-exposure.

We address the absence of a general large-scale dataset for single-view RGB–event paired object pose estimation by introducing **ExPO-Shape**, a large and diverse benchmark built upon the rich 3D asset pool of ShapeNetCore-v2 Chang et al. (2015). Following the rendering protocol of NOPE Nguyen et al. (2024), we use Blender to synthesize multi-view reference–target image pairs by randomly perturbing exposure (under-/over-exposure) and render corresponding pose annotations. ExPO-Shape provides **48,145** paired RGB–event samples covering **9,629** unique objects across **19** categories, enabling single-view pose evaluation for unseen objects and systematic robustness assessment under exposure degradation. As shown in Figure 1 (bottom), ExPO-Shape is 6.9× larger than YCB-Ev Rojtberg & Pöllabauer (2024) and 38.5× larger than RGB-D-E Dubeau et al. (2020) in sample scale, and offers dramatically higher object diversity (459× to YCB-Ev and YCB-Ev SD Rojtberg & Kühn (2025) and 9,629× to RGB-D-E, respectively). We empirically demonstrate ExPO-Shape as a scalable training benchmark, where naive event augmentation yields about 11-12% relative gains in single-view pose accuracy and noticeably improves robustness to exposure shifts.

To facilitate research on model-free single-view RGB–event pose estimation, we further propose **EVENPose**, which leverages event-derived geometric cues to guide novel view synthesis under exposure-degraded conditions. To address the instability of RGB embeddings under exposure collapse, EVENPose introduces an Event-Stabilized Embedding (ESE) to inject event-derived structural cues into the reference-view embedding in the shared latent space, stabilizing its distribution under exposure degradation. To improve novel view synthesis under exposure degradation, EVENPose proposes Event-Guided novel view synthesis (EG-NVS) that injects event-derived structural cues into the U-Net to provide a geometric "where-to-grow" shape bias in low-light and weak-texture regions. An Event Consistency Loss is adopted to supervise NVS in the event domain: instead of relying on exposure-sensitive pixel-wise losses, we enforce the events induced from the reference to the generated view to match the ground-truth events from the reference to the target, promoting geometrically faithful and exposure-robust synthesis.

In summary, (1) We introduce **ExPO-Shape**, the first large-scale RGB-event benchmark for model-free single-view object pose estimation, comprising 48,145 paired samples from 9,629 objects across 19 categories, with controlled under- and over-exposure; (2) we propose **EVENPose**, an event-guided model-free framework that leverages event-derived geometric cues to stabilize template representations and guide exposure-robust novel view synthesis for single-view 3D pose estimation; and (3) extensive experiments show that EVENPose serves as a strong baseline on ExPO-Shape and transfers competitively to existing RGB-event benchmarks. Training with ExPO-Shape yields up to 12% higher single-view pose accuracy and substantially improved robustness to exposure shifts.

2 Related work

RGB-event Pose Estimation Datasets. Public RGB-event pose datasets remain limited in scale and object diversity. RGB-D-E Dubeau et al. (2020) contains only one object and 10 sequences (2,500 frames), while YCB-Ev Rojtberg & Pöllabauer (2024) extends coverage to 21 YCB objects but provides only 21 real sequences (13,851 frames). YCB-Ev SD Rojtberg & Kühn (2025) further scales to 50,000 synthetic sequences, yet retains the same 21 object instances. Existing benchmarks therefore remain largely object-specific and provide limited support for evaluating severe exposure degradation or unseen-object generalization, motivating a larger and more diverse RGB-event benchmark. Such diversity is essential for reliably evaluating cross-object generalization and robustness to realistic exposure variations.

Event-based Pose Estimation. *Event-only object pose estimation.* Event cameras provide high dynamic range and reliable motion cues under challenging illumination, making them attractive for pose estimation. Existing methods mainly follow model-based pipelines, including event-to-contour alignment Reverter Valeiras et al. (2016), time-surface or geometry-guided refinement Liu et al. (2024a), line-based matching and optimization Liu et al. (2024b), and event-frame-based keypoint estimation followed by PnP Jawaid et al. (2022). However, these approaches typically rely on explicit geometric priors such as CAD models or point clouds, and often require iterative refinement or accurate initialization, limiting their applicability to model-free single-view inference.

RGB-event object pose estimation. RGB-event methods mainly target high-speed 6DoF tracking by combining event motion cues with frame-based appearance. Representative approaches fuse event-based optical flow with frame-based pose estimates Li et al. (2025), predict high-rate event pose updates followed by RGB-D refinement Dubeau et al. (2020), or optimize continuous object trajectories using event and photometric constraints Li & Stueckler (2021). While these methods demonstrate the complementarity of RGB and events, they are predominantly designed for temporal tracking and often require depth or known object geometry. In contrast, exposure-degraded, model-free pose estimation for unseen objects remains largely unexplored.

Single-view Novel View Synthesis. Single-view novel view synthesis aims to generate geometrically consistent appearances under unseen viewpoints. Existing methods improve view consistency through multi-view or geometric constraints Watson et al. (2022); Zheng & Vedaldi (2024); Liu et al. (2023); Tseng et al. (2023), or by incorporating stronger 3D priors from explicit representations Chan et al. (2023) and video diffusion models Kwak et al. (2024). However, these approaches are typically developed for well-exposed imagery and rely heavily on appearance cues, which can become unreliable under severe exposure degradation.

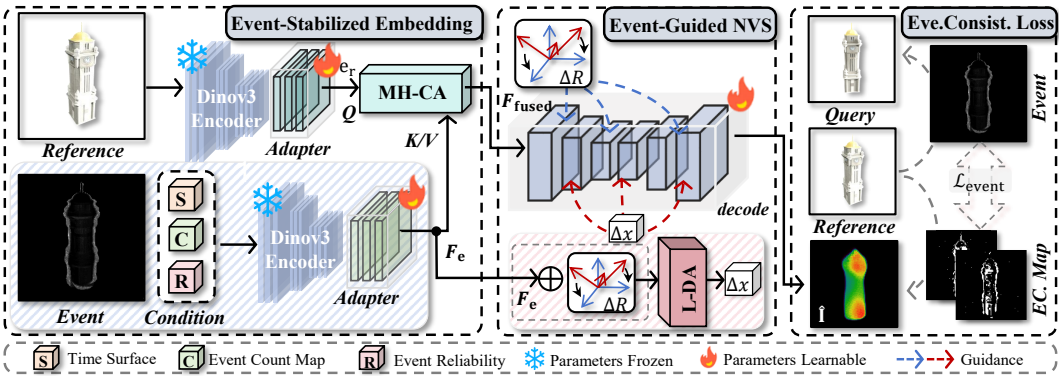


Figure 2: **Overview of EVENPose.** EVENPose performs model-free single-view RGB-event 3D pose estimation under exposure shifts via (i) **Event-Stabilized Embeddings** (left) that inject event cues F_e to stabilize the reference embedding F_r in a shared latent space, (ii) **Event-Guided Novel View Synthesis** (middle) that steers synthesis with event-derived 2D offset field Δx, and (iii) an **Event Consistency Loss** (right) that uses reference-to-target ground-truth events as supervision and encourages consistency in the event domain for geometry-faithful NVS.

3 METHOD

Overview. As shown in Figure 2, EVENPose performs model-free single-view pose estimation by learning to synthesize novel-view embeddings under event guidance. During training, the reference and query images and their event stream are encoded by frozen DINOv3 encoders Siméoni et al. (2025) with learnable adapters. The Event-Stabilized Embedding (ESE) fuses reference and event features, which are then fed into Event-Guided Novel View Synthesis (EG-NVS) to predict the query-view embedding conditioned on the relative viewpoint and event cues. Training is supervised by an L2 loss to the query embedding and an Event Consistency Loss that enforces consistency between the events induced by the synthesized view and the ground-truth reference-to-query events. At inference, ESE is used to generate novel-view embeddings over a set of template relative viewpoints, and the object pose is obtained by retrieving the synthesized embedding that best matches the query embedding.

3.1 EVENT-STABILIZED EMBEDDING

Under severe exposure degradation, RGB images tend to lose texture and edge information while being contaminated by noise. This further leads to unstable and less discriminative embeddings, which adversely affect the subsequent reference-conditioned pose reasoning. In contrast, event cameras capture only logarithmic brightness changes rather than absolute intensity, making event data robust to extreme illumination variations. Moreover, event data are sensitive to motion at object edges, thereby capturing geometric shapes and texture details effectively. Leveraging these properties of event signals, we propose the Event-Stabilized Embedding (ESE), which integrates events into an embedding space to enhance the distribution's robustness under severe exposure degradation.

As shown in Figure 2 (left), we first preprocess the event data to extract features along three dimensions. Given the event stream $E = \{(x, y, t, p)\}$ from $\mathrm{I_r}$ to $\mathrm{I_q}$, we compute the Time Surface S, the Event Count Map C, and the Event Reliability R as follows:

$$\begin{aligned} S(x,y) &= \exp\big(-(t_0 - t_{last}(x,y,p))/\tau\big), \\ C(x,y) &= N_+(x,y) + N_-(x,y), \\ R(x,y) &= \sqrt{C(x,y)}, \end{aligned} \tag{1}$$

where $t_{last}(x, y, p)$ denotes the timestamp of the most recent event at pixel (x, y) with polarity p, τ is the decay constant, and $N_\pm$ is the number of positive/negative events triggered within the time window from $\mathrm{I_r}$ to $\mathrm{I_q}$.

Subsequently, we embed the event representations, the reference image $\mathrm{I_r}$, and the query image $\mathrm{I_q}$ separately, followed by multi-head cross-attention (MH-CA) for feature fusion with the reference embedding as Query and the event embedding as Key-Value pairs.

$$e_{\mathrm{r}} = \mathrm{ADP_I}\left(\mathcal{D}(\mathrm{I_r})\right), e_{\mathrm{q}} = \mathrm{ADP_I}\left(\mathcal{D}(\mathrm{I_q})\right) \tag{2}$$

$$F_{\mathrm{e}} = \mathrm{ADP_e}\left(\mathcal{D}(\mathrm{Concat(S, C, R)})\right), \tag{3}$$

$$F_{\text{fused}} = \text{MH-CA}(Q = e_{\mathrm{r}}, K = F_{\mathrm{e}}, V = F_{\mathrm{e}}), \tag{4}$$

where e_{r}, e_{q}, and F_{e} are the embeddings of the reference image, query image, and event representations, respectively, and $\mathcal{D}$ represents the pretrained DINOv3 encoder. All DINOv3 encoders share weights and are frozen. Features from $\mathcal{D}$'s 6^{th}, 9^{th}, and 11^{th} layers for both the reference image and event representations are fed into the image adapter $\mathrm{ADP_I}$ and the event adapter $\mathrm{ADP_e}$, respectively. These adapters are composed of several upsampling convolutional layers to learn features from both modalities separately. Finally, the fused feature F_{fused} is passed into a U-Net for NVS.

3.2 Event-Guided Novel View Synthesis

Existing NVS approaches using U-Net typically embed and inject the relative pose as conditions Nguyen et al. (2024). However, when both input features and supervisory images are degraded under extreme lighting conditions, relying solely on relative pose is insufficient to recover or compensate for visual content missing or corrupted in the RGB input due to severe exposure problems, making it difficult to reconstruct clear, geometrically correct structures. To address this issue, we propose Event-Guided Novel View Synthesis (EG-NVS), which injects event information into the embedding space. The event stream provides illumination-invariant motion edges and texture features, enabling the generation of novel-view embeddings that are both geometrically accurate and visually clear, even under extreme illumination.

As shown in Figure 2 (middle), we first encode the relative pose ΔR by an MLP, then a Learnable Deformable Attention head produces a 2D offset field Δx using event features F_{e} and pose embedding f_{pose}:

$$f_{\text{pose}} = \mathrm{MLP}(\Delta R), \qquad \Delta x = \mathrm{LDA}\left([F_{\mathrm{e}}, f_{\text{pose}}]\right). \tag{5}$$

Finally, the offset field Δx is applied to the key and value matrices in the U-Net's cross-attention mechanism, along with a condition generated solely on ΔR. Through Event-Guided Novel View Synthesis, we obtain the final average embedding $\hat{e} = e(e_r, \Delta R)$ with exposure-robust event-aware guidance. During training, this embedding is used to compute the loss for supervision with the reference image and the query embedding.

3.3 Event Consistency Loss

Existing event-based pose methods often treat events as image-like signals, such as edge maps, without explicitly exploiting their physical triggering mechanism. An event $e(x, y, t, p)$ is generated when the log-intensity change at pixel (x, y) crosses a polarity-dependent threshold $t_{\text{pos/neg}}$. Motivated by this mechanism, we supervise NVS in the event domain rather than raw pixel space. This avoids exposure-sensitive pixel losses that may favor brightness agreement over pose-relevant geometry. We therefore enforce consistency between the events induced by the reference and synthesized views and the ground-truth reference-to-target events. As the intermediate motion is unknown and non-unique, supervision is anchored at the two endpoints through their induced event statistics.

Specifically, given the estimated novel-view embedding $\hat{e}$ from the U-Net, we first obtain its image $\hat{\mathrm{I}}$ using a pretrained VAE decoder. Then, with the reference image $\mathrm{I_r}$ and the events from $\mathrm{I_r}$ to $\mathrm{I_q}$, the Event Consistency Loss is computed as follows:

$$M_{\text{count}}(x, y) = N_+(x, y) - N_-(x, y), \tag{6}$$

$$M_{\text{pos}} = \mathbb{I}(M_{\text{count}} > 0), \qquad M_{\text{neg}} = \mathbb{I}(M_{\text{count}} < 0), \tag{7}$$

$$T_{\text{pos}} = \frac{\mathrm{L}(\hat{I}) - \mathrm{L}(I_r)}{t_{\text{pos}}}, \qquad T_{\text{neg}} = \frac{\mathrm{L}(I_r) - \mathrm{L}(\hat{I})}{t_{\text{neg}}}, \tag{8}$$

$$\mathcal{L}_{\text{pos}} = \mathbb{E}\Big[\big((T_{\text{pos}} - N_{+}) \odot M_{\text{pos}}\big)^2\Big], \qquad \mathcal{L}_{\text{neg}} = \mathbb{E}\Big[\big((T_{\text{neg}} - N_{-}) \odot M_{\text{neg}}\big)^2\Big], \tag{9}$$

$$\mathcal{L}_{\text{event}} = \mathcal{L}_{\text{pos}} + \mathcal{L}_{\text{neg}}, \tag{10}$$

where M_{count}, M_{pos}, and M_{neg} represent the polarity event count map, the positive mask, and the negative mask of the events from $\mathrm{I_r}$ to $\mathrm{I_q}$. t_{pos} and t_{neg} denote the thresholds for positive and negative events, respectively. T_{pos} and T_{neg} are the positive and negative event count maps of the estimated events between the decoded image of novel-view embedding $\hat{\mathrm{I}}$ and $\mathrm{I_r}$. The function $\mathrm{L(I)} = \log(\mathrm{I}+\epsilon)$ represents the log transformation on the grayscale images of $\hat{\mathrm{I}}$ and $\mathrm{I_r}$. $\mathbb{I}(\cdot)$ is the indicator function, $\mathbb{E}[\cdot]$ indicates averaging over all spatial locations and batch dimensions, and $\odot$ stands for element-wise multiplication. Finally, combining the $\mathcal{L}_2 = \frac{1}{N}||e_q - \hat{e}||_2^2$ loss between the query embedding e_q and $\hat{e}$, where $N = B \times H \times W$ is the total number of pixels, the overall training loss is $\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{event}} + \mathcal{L}_2$, where α is the weight of $\mathcal{L}_{\text{event}}$.

3.4 ExPO-Shape Dataset

Current RGB-event datasets for object pose estimation often face issues such as a limited number of objects and categories, as well as registration errors between event cameras and RGB cameras. These problems not only hinder the pose estimation accuracy of models on known objects but also significantly limit their capability to estimate poses for unseen objects. To address this gap, we propose ExPO-Shape (Figure 1), the first large-scale RGB-event benchmark for model-free single-view object pose estimation, comprising 48,145 paired samples from 9,629 objects across 19 categories under controlled under-/over-exposure conditions.

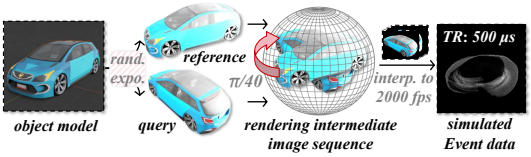


Figure 3: ExPO-Shape dataset construction pipeline.

Why not directly simulate events from two rendered views? A simple alternative is to generate events directly between rendered reference and query images, as in two-view ShapeNet-based pose datasets such as NOPE. However, this becomes unreliable for larger viewpoint gaps: beyond moderate rotations (e.g., beyond $\sim$20°), the implied two-frame motion is highly non-physical, producing distorted and unstable event patterns unsuitable for ground-truth supervision.

Rendering and view sampling. We largely follow the rendering protocol of NOPE to render reference and query images on object models from ShapeNetCore-v2 Chang et al. (2015). As shown in Figure 3, for each 3D model, we randomly sample camera poses to render five reference images and five query images using BlenderProc Denninger et al. (2019) as the rendering engine. For evaluation, we construct a test split where reference views are sampled from front views with $|\text{azimuth}| \leq 45°$ for five held-out categories. This design avoids making the test setting dominated by large viewpoint gaps, since our focus is on exposure degradation rather than extreme-viewpoint pose estimation.

Exposure scheduling. Each reference-query pair is assigned an exposure condition randomly sampled from six levels: *Extreme/Medium/Slight Under-Exposure* and *Extreme/Medium/Slight Over-Exposure*. We apply the same sampled exposure setting to both the reference and the query to avoid a harder but orthogonal cross-exposure matching setting (e.g., over-exposed reference vs. under-exposed query), so that the evaluation isolates pose estimation robustness to exposure-induced appearance cue collapse, as summarized in Figure 1 (right-bottom).

Event synthesis. As shown in Figure 3, given the sampled reference and query poses, we interpolate the motion trajectory via spherical linear interpolation (slerp) between the corresponding quaternions. The number of interpolation steps is set as $n = \Delta\theta/(\pi/40)$, where $\Delta\theta$ is the angular distance. We then render the intermediate image sequence under the same exposure setting. To generate realistic events, we adopt v2e Hu et al. (2021) following Qi et al. (2023); Yu et al. (2025): we first interpolate frames to 2000 fps and simulate events with a temporal resolution of 500 μs. This high temporal resolution prevents inter-frame appearance jumps that would otherwise produce non-physical and unstable events. Finer temporal resolution primarily affects the stability of event-derived structural cues.

4 EXPERIMENT

Table 1: Quantitative comparison on ExPO-Shape. The best is in **bold**, while the second is underlined.

Metrics	ACC 30 ↑						Median ↓					
Methods	Bus	Clock	Mug	Train	Washer	Mean	Bus	Clock	Mug	Train	Washer	Mean
PIZZA	21.6	23.6	20.5	23.9	16.7	21.3	71.8	59.6	76.3	65.7	77.8	70.2
RelPose	26.4	25.3	24.6	23.5	26.1	25.2	51.9	71.7	64.3	69.7	51.3	61.8
NOPE	28.7	25.2	23.0	29.7	21.6	25.6	56.1	56.4	54.2	57.7	65.8	58.0
PIZZA+E	25.3	24.0	28.2	23.4	27.3	25.6	57.1	72.6	61.8	82.2	58.3	66.4
RelPose+E	27.8	24.1	26.6	25.6	26.4	26.1	50.2	74.3	60.4	55.8	52.3	58.6
CoTF+PIZZA	23.1	25.3	22.4	26.7	21.6	23.8	58.9	65.6	70.1	64.2	79.3	67.6
CoTF+RelPose	26.5	24.3	23.4	25.8	27.4	25.5	57.4	63.1	69.9	54.5	59.2	60.8
CoTF+NOPE	30.2	26.7	22.1	35.8	40.2	31.0	54.4	50.6	64.7	41.1	39.3	50.0
Ours	**59.7**	**46.0**	**60.0**	**48.0**	**50.5**	**52.8**	**19.1**	**32.1**	**16.7**	**34.6**	**27.8**	**26.1**

4.1 EXPERIMENTAL SETTINGS

Implementation details. Our model was trained on an NVIDIA A100 GPU using the AdamW optimizer, with an initial learning rate of 5×10^{-5} and the hyperparameter α set to 1×10^{-3}. The image size and epoch settings follow those in Nguyen et al. (2024).

Datasets and metrics. To evaluate unseen-category generalization, we split ExPO-Shape by category into 14 training categories ($\sim$9,130 objects) and five test categories (bus, clock, mug, train, washer with $\sim$100 objects each), rendering five reference-query pairs per object. We also evaluate on YCB-Ev SD Rojtberg & Kühn (2025) with 500 reference-query pairs covering all 21 objects in multi-object scenes and the real YCB-Ev Rojtberg & Pöllabauer (2024) with 230 pairs, which contains real event data with underexposure, motion blur, and occlusion. Following Nguyen et al. (2024), we report ACC 30 and Median Pose Error to measure accuracy and overall robustness, respectively.

Table 2: Quantitative comparison on YCB-Ev SD and real YCB-Ev. Mean_{21} denotes the mean over all 21 objects. Best in **bold**, second best underlined.

(a) YCB-Ev SD

Metrics	ACC 30 ↑				Median ↓			
Methods	Bott.	Drill	Mug	Mean_{21}	Bott.	Drill	Mug	Mean_{21}
PIZZA	14.8	16.5	12.0	15.2	78.0	70.0	88.0	79.2
RelPose	24.8	28.0	21.0	25.3	82.0	75.0	90.0	81.5
NOPE	15.3	16.2	13.1	15.4	64.5	63.0	74.5	69.1
PIZZA+E	19.0	22.5	15.5	20.4	87.0	80.0	95.0	88.3
RelPose+E	24.5	27.2	21.8	25.2	76.0	69.0	83.0	75.8
CoTF+PIZZA	17.2	20.0	14.0	19.4	90.0	82.0	96.0	90.9
CoTF+RelPose	25.6	27.5	21.5	25.8	72.5	66.0	70.5	72.0
CoTF+NOPE	18.0	21.0	14.8	19.2	66.0	61.5	76.0	65.5
Ours	**31.5**	**34.0**	**27.5**	**30.1**	**52.0**	**40.0**	**58.0**	**47.4**

(b) YCB-Ev

Metrics	ACC 30 ↑						Median ↓					
Methods	Buck.	Box	Scis.	Bott.	Can	Mean_{21}	Buck.	Box	Scis.	Bott.	Can	Mean_{21}
PIZZA	11.2	10.4	7.8	8.6	9.9	9.6	94.0	98.7	109.5	103.8	96.4	100.2
RelPose	18.3	17.1	13.8	15.4	16.9	16.5	79.2	83.1	91.4	87.6	81.5	84.7
NOPE	13.7	12.9	9.6	10.8	12.7	12.1	77.8	81.6	89.3	85.2	79.8	82.9
Ours	**27.9**	**25.8**	**21.6**	**23.4**	**26.2**	**24.7**	**55.4**	**59.1**	**68.7**	**64.5**	**57.2**	**61.8**

Comparative methods. We compare three recent single-reference relative pose estimators on unseen objects: RelPose Zhang et al. (2022), PIZZA Du et al. (2022), and NOPE Nguyen et al. (2024). We first retrain all methods on the ExPO-Shape training set to assess their robustness under various exposures. To validate the effectiveness of ExPO-Shape, we equip RelPose and PIZZA with the ESE to fuse RGB and event and retrain them on ExPO-Shape. NOPE is treated as the primary baseline and is analyzed separately in our ablations. We also test a decoupled "recover-then-estimate" pipeline by training the three pose estimators on clean ShapeNetCore-v2 renders Chang et al. (2015) (with the same split as ExPO-Shape), and at test time applying a SOTA exposure correction method CoTF Li et al. (2024) to ExPO-Shape images before pose estimation.

4.2 QUANTITATIVE COMPARISON

Tables 1 and 2 show that EVENPose consistently outperforms existing methods on ExPO-Shape, YCB-Ev SD, and real YCB-Ev, demonstrating strong exposure robustness and cross-benchmark generalization. We summarize the quantitative findings from four aspects:

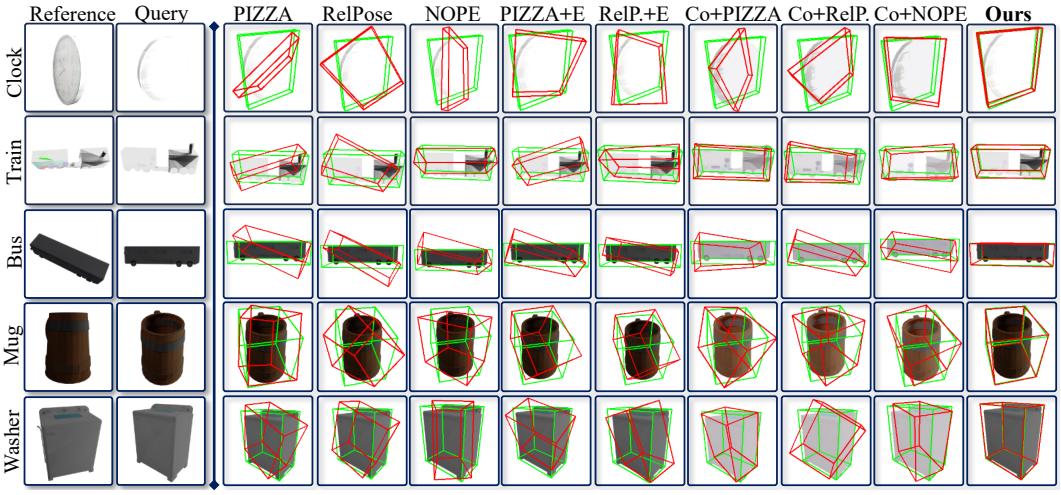


Figure 4: Qualitative results on ExPO-Shape. The first two rows show Clock and Train under extreme and medium over-exposure, respectively. The remaining three rows show Bus, Mug, and Washer under extreme, medium, and slight under-exposure, respectively.

First, ExPO-Shape supports scalable training and exposes the benefits of events. On ExPO-Shape, augmenting RGB-only baselines with events yields consistent improvements: the mean ACC 30 of PIZZA increases from 21.3 to 25.6 and that of RelPose from 25.2 to 26.1, corresponding to an average relative gain of $\sim$11-12%. Meanwhile, the Median error changes only marginally (PIZZA: 70.2$\rightarrow$66.4; RelPose: 61.8$\rightarrow$58.6), suggesting that naive event fusion improves overall accuracy but still leaves occasional large failures unresolved under challenging exposure conditions.

Second, recover-then-estimate is less consistent than direct event augmentation. Although exposure recovery improves visual appearance, it remains inferior to event augmentation for both PIZZA and RelPose in mean ACC 30 (CoTF+PIZZA: 23.8 $<$ PIZZA+E: 25.6; CoTF+RelPose: 25.5 $<$ RelPose+E: 26.1). This suggests that exposure enhancement does not necessarily preserve geometry-faithful structures, and residual artifacts or shape distortions may propagate to downstream pose estimation and impair reliable geometric reasoning.

Third, ExPO-Shape provides a targeted benchmark for exposure robustness. NOPE exhibits a substantial gap between clean ShapeNet renders (mean ACC 30: 59.8 / Median: 24.4) Nguyen et al. (2024) and exposure-degraded ExPO-Shape (mean ACC 30: 25.6 / Median: 58.0), highlighting the difficulty of single-view pose estimation under severe exposure shifts.

Finally, the results on real YCB-Ev further validate real-world transfer. RGB-only methods degrade substantially under real under-exposure, motion blur, and occlusion, achieving only 9.6 to 16.5 mean ACC 30 with Median errors of 82.9 to 100.2. In contrast, EVENPose reaches 24.7 mean ACC 30 and reduces the Median error to 61.8. These results indicate that the event-guided structural cues learned on ExPO-Shape transfer effectively to real event sensors and remain robust under compound real-world degradations and sensor imperfections.

4.3 Qualitative Results

Figure 4 qualitatively compares pose predictions under different exposure conditions and object geometries. RGB-only methods, including PIZZA, RelPose, and NOPE, often produce noticeably misaligned poses when severe exposure degradation suppresses appearance and edge cues, as observed for Clock, Train, and Mug. The recover-then-estimate pipelines partially restore visibility, but the recovered images may contain hallucinated textures or distorted structures (e.g., Clock and Washer), which subsequently propagate to pose estimation and still lead to inaccurate predictions. In contrast, EVENPose consistently produces poses that align more closely with the ground truth across both over- and under-exposure cases. This suggests that directly exploiting event-derived structural cues is more reliable than either degraded RGB appearance or appearance recovery alone, particularly when exposure corruption removes pose-relevant visual evidence.

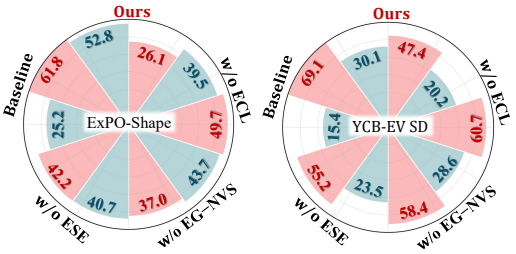


Figure 5: Quantitative ablation of the components on ACC 30 and Median.

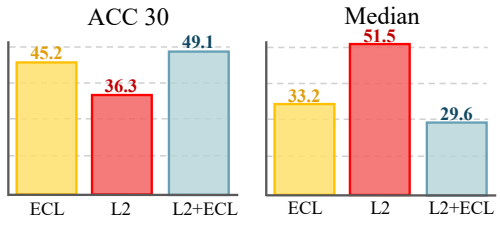


Figure 6: Quantitative ablation on the loss choice.

4.4 Ablation Study

Effectiveness of components. To assess the contribution of each component, we conduct ablation by removing the ECL, EG-NVS, and ESE. Figure 5 reports ACC 30 and Median pose error on ExPO-Shape and YCB-Ev SD. Among all components, ECL is the most critical: removing it causes a pronounced drop in accuracy on both datasets (13.3 and 9.9 points), indicating the importance of event-consistent supervision for exposure-robust pose reasoning.

Impact of loss. We further study how supervision affects performance by training three variants: ECL-only, $\mathcal{L}_2$-only, and ECL+$\mathcal{L}_2$. Figure 6 reports ACC 30 and Median pose error on ExPO-Shape and YCB-Ev SD. ECL-only model slightly underperforms the combined setting, indicating that event supervision alone may be under-constrained and can tolerate appearance drift. In contrast, removing ECL causes a substantial drop (ACC 30 $\downarrow$12.8 and Median error $\uparrow$21.9), confirming that ECL contributes a distinct, event-domain geometric supervision beyond $\mathcal{L}_2$.

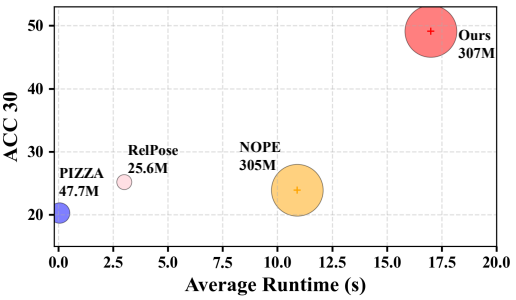


Figure 7: Efficiency analysis.

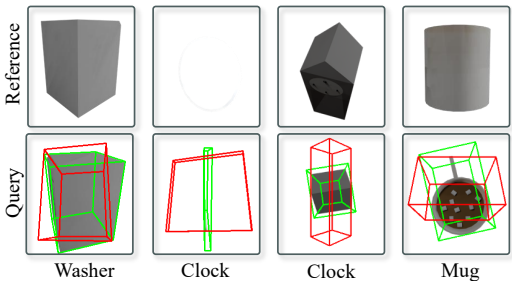


Figure 8: Failure case analysis.

4.5 Efficiency and Failure Analysis

Efficiency. As shown in Figure 7, event guidance increases inference time while adding only 2M parameters over NOPE (307M vs. 305M). This is accompanied by a substantial improvement in pose accuracy, while real-time inference remains an important direction for future work.

Failure cases. Figure 8 shows that failures mainly arise from object symmetry, extreme exposure, and ambiguous viewpoints. These conditions respectively introduce pose ambiguity, unreliable RGB/event cues, and insufficient geometric evidence. Addressing them may require symmetry-aware reasoning and stronger temporal or geometric constraints.

5 Conclusion

In this work, we study model-free single-view pose estimation under severe exposure degradation. We introduce ExPO-Shape, a large-scale RGB-event benchmark for unseen-object evaluation, and EVENPose, which leverages event cues for robust embedding stabilization and geometry-faithful novel-view synthesis. Experiments show strong performance on ExPO-Shape and competitive transfer to existing RGB-event benchmarks.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, figure refinement, and feedback on the presentation of the methodology and experiments. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final content of this paper.

REPRODUCIBILITY STATEMENT

We provide detailed model, training, dataset construction, and evaluation settings in the paper and appendix. Upon acceptance, we will publicly release the complete EVENPose codebase, pretrained models, and the ExPO-Shape dataset, together with the corresponding generation scripts and splits.

REFERENCES

Axel Barroso-Laguna, Sowmya Munukutla, Victor Adrian Prisacariu, and Eric Brachmann. Matching 2d images in 3d: Metric relative pose from metric correspondences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4852–4863, 2024.

Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4217–4229, 2023.

Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015.

Haosheng Chen, David Suter, Qiangqiang Wu, and Hanzi Wang. End-to-end learning of object motion estimation from retinal events for event-based object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 10534–10541, 2020.

Zezhou Cheng, Matheus Gadelha, and Subhransu Maji. Accidental turntables: Learning 3d pose by watching objects turn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2113–2122, 2023.

Maximilian Denninger, Martin Sundermeyer, Dominik Winkelbauer, Youssef Zidan, Dmitry Olefir, Mohamad Elbadrawy, Ahsan Lodhi, and Harinandan Katam. Blenderproc. *arXiv preprint arXiv:1911.01911*, 2019.

Yuming Du, Yang Xiao, Michael Ramamonjisoa, Vincent Lepetit, et al. Pizza: A powerful image-only zero-shot zero-cad approach to 6 dof tracking. In *2022 International Conference on 3D Vision (3DV)*, pp. 515–525. IEEE, 2022.

Etienne Dubeau, Mathieu Garon, Benoit Debaque, Raoul de Charette, and Jean-François Lalonde. Rgb-de: Event camera calibration for fast 6-dof object tracking. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 127–135. IEEE, 2020.

Xingyi He, Jiaming Sun, Yuang Wang, Di Huang, Hujun Bao, and Xiaowei Zhou. Onepose++: Keypoint-free one-shot object pose estimation without cad models. *Advances in Neural Information Processing Systems*, 35:35103–35115, 2022.

Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic dvs events. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1312–1321, 2021.

Nishant Jain, Suryansh Kumar, and Luc Van Gool. Enhanced stable view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13208–13217, 2023.

Mohsi Jawaid, Ethan Elms, Yasir Latif, and Tat-Jun Chin. Towards bridging the space domain gap for satellite pose estimation using event sensing. *arXiv preprint arXiv:2209.11945*, 2022.

Jeong-gi Kwak, Erqun Dong, Yuhe Jin, Hanseok Ko, Shweta Mahajan, and Kwang Moo Yi. Vivid-1-to-3: Novel view synthesis with video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6775–6785, 2024.

Fu Li, Shishir Reddy Vutukur, Hao Yu, Ivan Shugurov, Benjamin Busam, Shaowu Yang, and Slobodan Ilic. Nerf-pose: A first-reconstruct-then-regress approach for weakly-supervised 6d object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2123–2133, 2023.

Haolong Li and Joerg Stueckler. Tracking 6-dof object motion from events and frames. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14171–14177. IEEE, 2021.

Zhichao Li, Arren Glover, Chiara Bartolozzi, and Lorenzo Natale. 6-dof object tracking with event-based optical flow and frames. *arXiv preprint arXiv:2508.14776*, 2025.

Ziwen Li, Feng Zhang, Meng Cao, Jinpu Zhang, Yuanjie Shao, Yuehuan Wang, and Nong Sang. Real-time exposure correction via collaborative transformations and adaptive sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2984–2994, 2024.

Yuan Liu, Yilin Wen, Sida Peng, Cheng Lin, Xiaoxiao Long, Taku Komura, and Wenping Wang. Gen6d: Generalizable model-free 6-dof object pose estimation from rgb images. In *European Conference on Computer Vision*, pp. 298–315. Springer, 2022.

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023.

Zibin Liu, Banglei Guan, Yang Shang, Shunkun Liang, Zhenbao Yu, and Qifeng Yu. Optical flow-guided 6dof object pose tracking with an event camera. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 6501–6509, 2024a.

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 Transactions on Image Processing*, 2024b.

Van Nguyen Nguyen, Thibault Groueix, Georgy Ponimatkin, Yinlin Hu, Renaud Marlet, Mathieu Salzmann, and Vincent Lepetit. Nope: Novel object pose estimation from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17923–17932, 2024.

Yunshan Qi, Lin Zhu, Yu Zhang, and Jia Li. E2nerf: Event enhanced neural radiance fields from blurry images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13254–13264, 2023.

David Reverter Valeiras, Garrick Orchard, Sio-Hoi Ieng, and Ryad B Benosman. Neuromorphic event-based 3d pose estimation. *Frontiers in neuroscience*, 9:522, 2016.

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 1.1: Event-vision dataset for 6dof object pose estimation. In *European Conference on Computer Vision*, pp. 1–13. Springer, 2024.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He, Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou. Onepose: One-shot object pose estimation without cad models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6825–6834, 2022.

Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16773–16783, 2023.

Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022.

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17868–17879, 2024.

Wangbo Yu, Chaoran Feng, Jianing Li, Jiye Tang, Jiashu Yang, Zhenyu Tang, Meng Cao, Xu Jia, Yuchao Yang, Li Yuan, et al. Evagaussians: Event stream assisted gaussian splatting from blurry images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24780–24790, 2025.

Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose: Predicting probabilistic relative rotation for single objects in the wild. In *European Conference on Computer Vision*, pp. 592–611. Springer, 2022.

Chuanxia Zheng and Andrea Vedaldi. Free3d: Consistent novel view synthesis without 3d representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9720–9731, 2024.

Yajing Zheng, Zhaofei Yu, Song Wang, and Tiejun Huang. Spike-based motion estimation for object tracking through bio-inspired unsupervised learning. *IEEE Transactions on Image Processing*, 32: 335–349, 2022.

Han Zhou, Wei Dong, and Jun Chen. Lita-gs: Illumination-agnostic novel view synthesis via reference-free 3d gaussian splatting and physical priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21580–21589, 2025.

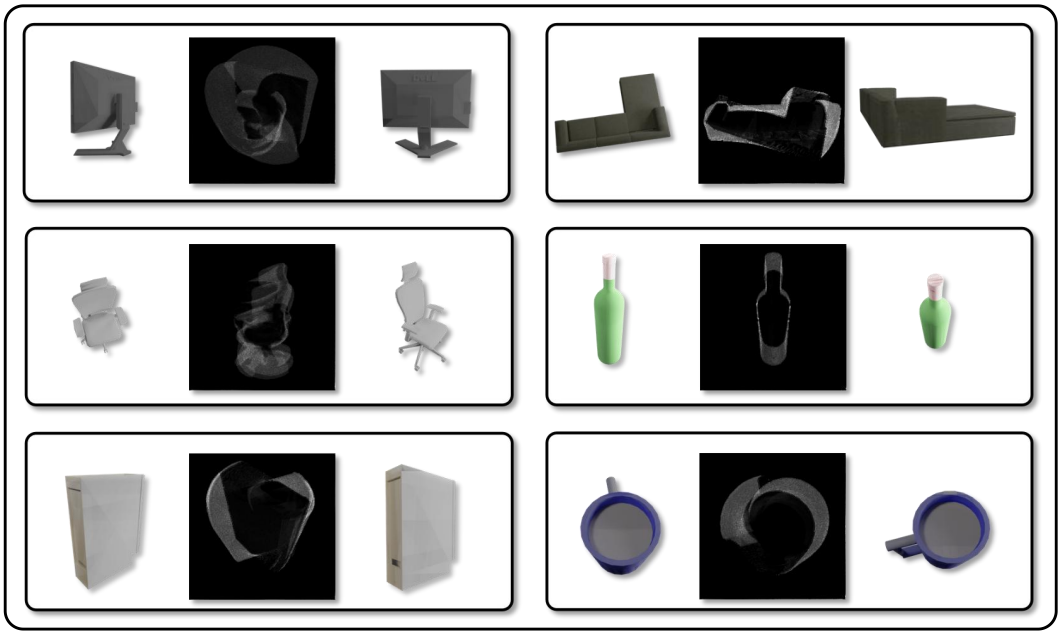


Figure 9: Samples of ExPO-Shape. Each sample shows the reference, event, and the query, respectively.

A SUMMARY

We provide additional qualitative evidence and implementation details here to complement the main paper. (1) We illustrate the background of EVENPose in B. (2) We present representative samples of ExPO-Shape in C to illustrate its category diversity, exposure scheduling, and RGB-event characteristics. (3) We visualize the failure of directly simulating two-view events from a single reference-query pair in D, showing that the event signal can be severely distorted when the viewpoint gap is large. This supports our design choice of sequence-based rendering and dense interpolation for stable event synthesis. (4) We visualize additional experiments in E.

B BACKGROUND

Given a reference image $\mathrm{I_r}$ of a target object, a query image $\mathrm{I_q}$ of the same object, and the intervening event stream $\mathcal{E}_{r\to q}$, our goal is to estimate the relative pose of the single query frame with respect to the reference. Here, "single-view" refers to single-query inference: only one query RGB frame is used for pose prediction, while $\mathcal{E}_{r\to q}$ serves as an auxiliary asynchronous observation rather than a sequence of query frames for temporal pose tracking. Our method identifies a specific discretized relative pose ΔR that maximizes the probability $p(\Delta R \mid \mathrm{I_r}, \mathrm{I_q}, \mathcal{E}_{r\to q})$. We model this probability in the image embedding space as

$$p(\Delta R \mid \mathrm{I_r}, \mathrm{I_q}, \mathcal{E}_{r\to q}) = \mathcal{N}\big(e_q \mid e(e_r, \mathcal{E}_{r\to q}, \Delta R), \Sigma(e_r, \mathcal{E}_{r\to q}, \Delta R)\big), \tag{11}$$

where e_r and e_q denote the embeddings of $\mathrm{I_r}$ and $\mathrm{I_q}$, respectively. The predicted mean $e(e_r, \mathcal{E}_{r\to q}, \Delta R)$ represents the novel-view embedding under candidate pose ΔR, with event cues providing auxiliary structural guidance. Pose estimation is then performed by matching the query embedding against the generated novel-view embeddings over the discretized pose candidates.

C SAMPLES OF EXPO-SHAPE

Figure 9 visualizes representative samples from ExPO-Shape, covering six object categories. Each panel shows a reference-query pair rendered under an exposure-shifted setting, together with the corresponding event observation synthesized along the reference-to-query motion. In contrast to

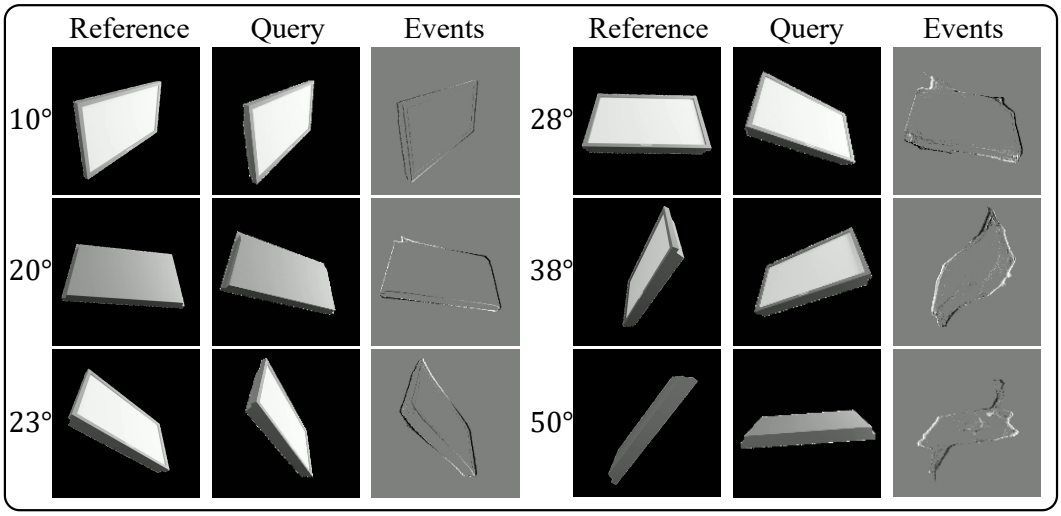


Figure 10: Example of simulating two-view events.

frame-based images whose appearance can suffer from severe saturation or under-exposure, the event signal highlights motion-induced brightness changes and preserves clear structural outlines around object boundaries. ExPO-Shape poses a challenging yet well-controlled benchmark for evaluating single-view RGB-event pose estimation under exposure shifts.

D Failure of Two-view Event

Figure 10 visualizes why we do not directly simulate events between the two rendered views of existing datasets (i.e., using only a reference and a query frame).

When the viewpoint change is small (e.g., $\sim$10°-20°), the two-view difference can still produce edge-like event responses that roughly follow the object boundaries. However, once the relative rotation exceeds $\sim$20°, the simulated events become increasingly distorted and fragmented: many responses no longer align with true geometric contours, while spurious structures and missing edges appear due to large inter-frame motion and severe temporal aliasing. Such artifacts make the resulting event maps unreliable as supervision or as a ground-truth event signal.

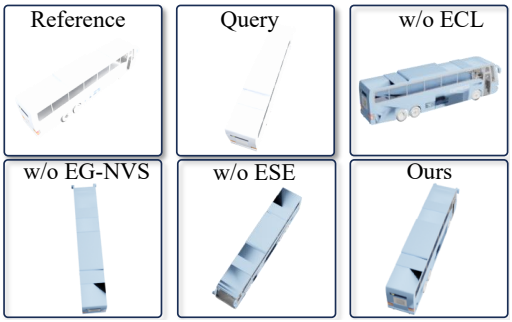


Figure 11: Qualitative ablation of the components.

This observation motivates our event synthesis strategy in ExPO-Shape: instead of relying on a single two-frame difference, we re-render a short intermediate sequence along the reference-to-query motion trajectory and generate events from densely interpolated frames. By providing sufficient temporal sampling, the simulated events remain stable across large viewpoint changes, yielding robust, physically plausible event measurements for exposure-degraded single-view pose estimation.

E Additional Visualization

Figure 11 shows the qualitative ablation of the components, and Figure 12 presents the additional visual comparison on ExPO-Shape.

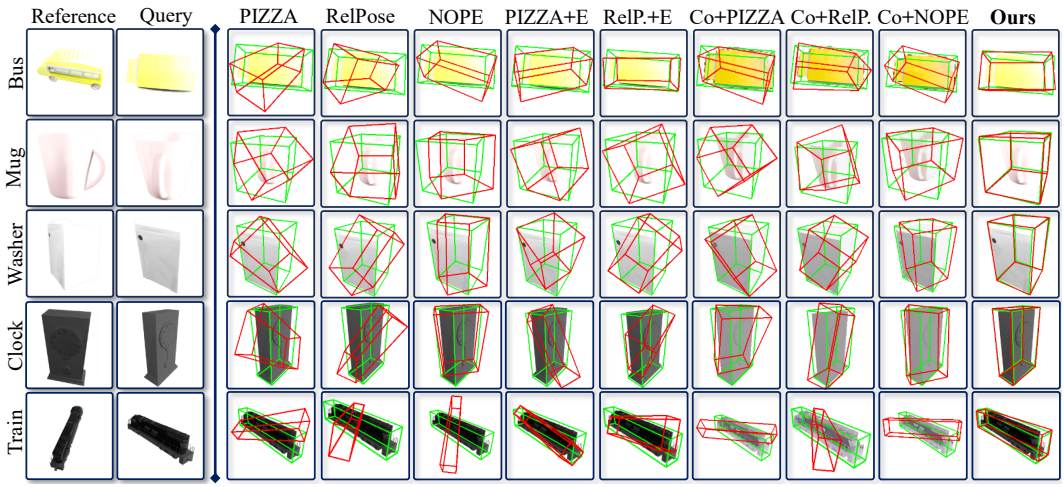


Figure 12: Additional qualitative results on ExPO-Shape.