



E-VLA: Event-augmented Vision-Language-Action Model for Dark and Blurred Scenes

Jiajun Zhai^{1,*}, Hao Shi^{2,1,*}, Shangwei Guo¹, Kailun Yang³, and Kaiwei Wang^{1†}

¹Zhejiang University, ²Ant Group, ³Hunan University

Abstract. Robotic Vision-Language-Action (VLA) models generalize well for open-ended manipulation, but their perception is fragile under sensing-stage degradations such as extreme low light, motion blur, and black clipping. We present E-VLA, an event-augmented VLA framework that improves manipulation robustness when conventional frame-based vision becomes unreliable. Instead of reconstructing images from events, E-VLA directly leverages motion and structural cues in event streams to preserve semantic perception and perception-action consistency under adverse conditions. We build an open-source teleoperation platform with a DAVIS346 event camera and collect a real-world synchronized RGB-event-action manipulation dataset across diverse tasks and illuminations. We also propose lightweight, pretrained-compatible event integration strategies and study event windowing for stable deployment. Experiments show that even a simple parameter-free fusion, *i.e.*, overlaying accumulated event maps onto RGB images, could substantially improve robustness in dark and heavy-blur scenes: on Pick-Place at 20 lux, success increases from 0% (image-only) to 60% with overlay fusion and to 90% with our event adapter; under severe motion blur (1000 ms-exposure proxy), Pick-Place improves from 0% to 20-25%, and Sorting from 5% to 32.5%. Overall, E-VLA provides systematic evidence that event-driven perception can be effectively integrated into VLA models, pointing toward robust embodied intelligence beyond conventional frame-based imaging. Code and dataset will be available at E-VLA.

Keywords: Vision-language-action models · Event cameras · Robotics

1 Introduction

Vision-Language-Action (VLA) models have recently emerged as a promising paradigm for robotic manipulation, enabling systems to map visual observations and language instructions directly to action sequences. Representative systems such as RT-1 [5], RT-2 [67], the π family [4, 26, 27], OpenVLA [28], RDT-1B [37],

* Equal contribution.

† Correspondence.

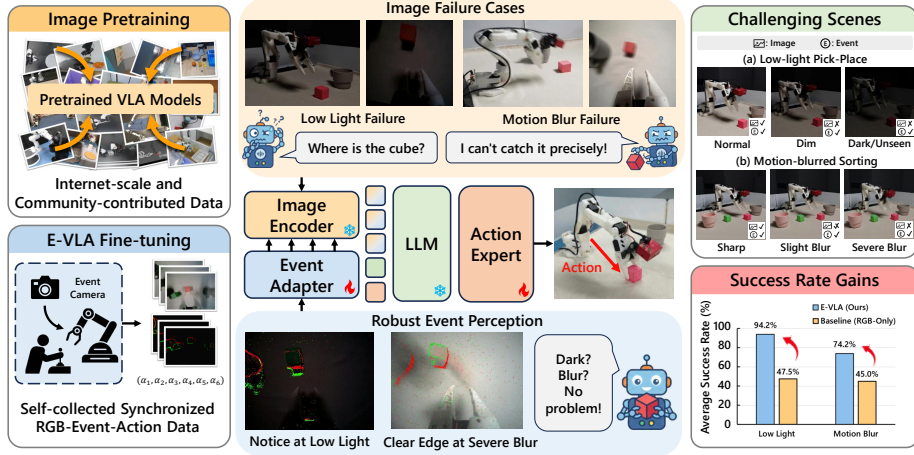


Fig. 1: Middle: Image-based VLA models degrade under low-light and motion-blurred conditions, leading to failures in object detection and imprecise manipulation. We propose E-VLA, which fuses stable event visual cues with image features within the VLA pipeline, preserving reliable performance under adverse conditions. Left: Existing VLA models (*e.g.*, SmolVLA [49]) are pretrained on large-scale text-image datasets and manipulation videos. We build a teleoperation system equipped with an event camera and collect synchronized RGB-event-action data for E-VLA training. Right: We evaluate our method across gradient low-light and motion-blurred scenes. E-VLA consistently achieves higher success rates than RGB baselines.

and Gemini Robotics [58] demonstrate a common trend: leveraging large-scale robot interaction data and pretrained vision-language models to learn versatile, language-conditioned manipulation policies. Most existing VLA approaches are built upon frame-based RGB perception, where pretrained visual encoders are coupled with policy learning modules to jointly model perception, reasoning, and control, and have shown strong performance on open-ended manipulation tasks in well-lit laboratory settings [15, 32, 40, 65].

However, real-world deployment exposes a critical weakness of current VLA systems: perceptual robustness under sensing-stage degradations. Recent surveys and benchmarks highlight illumination variation and visual domain shift as major factors limiting stable VLA performance [17, 32, 40]. In particular, low-light conditions substantially reduce signal quality, while increasing exposure time to recover brightness inevitably introduces motion blur and additional latency during fast manipulation. In extreme cases, severe under-exposure can lead to black clipping, where frame images become nearly unusable. Importantly, these failures originate from the limitations of frame-based imaging itself. Although image enhancement and data augmentation can improve visual appearance, they work on already degraded measurements and cannot recover information lost at capture time. Event cameras, which asynchronously capture brightness changes with high temporal resolution and wide dynamic range [14, 19], offer a compelling alternative for robust robotic perception under low light and fast motion. Yet integrating event sensing into VLA is nontrivial. First, event streams are sparse

and statistically different from dense RGB images, making them poorly matched to image-domain pretraining that modern VLA models rely on. Second, in manipulation settings, event triggering is tightly coupled with arm motion and control timing, resulting in highly non-stationary temporal distributions. These challenges make it unclear how to incorporate event signals into pretrained VLA architectures while preserving stable perception-action coupling.

To address this gap, we introduce **E-VLA**, an event-augmented VLA framework for robust manipulation under visually degraded conditions. We build an open-source teleoperation platform equipped with a DAVIS346 event camera and collect a real-world synchronized RGB-event-action dataset across multiple manipulation tasks and illumination conditions, enabling systematic study of event-augmented VLA learning. On top of a SmolVLA-style backbone, we design lightweight and pretrained-compatible event integration strategies, including a parameter-free event overlay mechanism and a hierarchical event adapter, and investigate event windowing/accumulation schemes to manipulation dynamics.

Extensive experiments show that E-VLA substantially improves robustness in low-light and blur-prone scenarios while preserving performance in normal lighting. Notably, under complete black clipping (an extreme low-light failure mode), E-VLA maintains task success rates above 80%, compared with 0% for the image-only baseline. We further show that event-count-based accumulation yields more stable perception-action coupling than time-based windowing, and that event augmentation improves generalization to unseen degraded conditions without requiring additional low-light training data.

To summarize, our contributions are as follows:

- We present **E-VLA**, the first Vision-Language-Action framework that integrates event-based visual sensing for robust robotic manipulation under low-light and motion-blur degradations.
- We build an open-source event-augmented teleoperation platform and collect a real-world synchronized RGB-event-action manipulation dataset across multiple tasks and illumination conditions, enabling systematic validating and training of event-enhanced VLA models.
- We propose lightweight, pretrained-compatible event integration strategies, including a parameter-free overlay fusion and a hierarchical event adapter, together with a manipulation-aware event windowing design.
- We conduct extensive experiments and ablations, providing the first systematic empirical evidence and practical design insights for integrating event-driven perception into scalable VLA learning.

2 Related Work

2.1 Vision-Language-Action Systems

Vision-Language-Action (VLA) models learn large-scale policies that directly map visual observations and language instructions to robot actions [32, 40, 65]. Representative systems, including the RT series [5, 67], the π family [4, 26, 27],

OpenVLA [28], RDT-1B [37], and Gemini Robotics [58], demonstrate that scaling multimodal transformers with large robot datasets enables generalizable language-conditioned manipulation across diverse tasks. Recent efforts further explore enhanced spatial reasoning via 3D representations [16, 31, 50, 55], tactile integration for contact modeling [3, 13, 24], reinforcement learning refinement [12, 23, 38], and efficient deployment strategies [49, 61, 63]. Despite architectural diversity, existing VLA systems predominantly rely on frame-based RGB cameras as their primary perceptual interface. Frame-based imaging integrates photons over an exposure interval, which inherently limits signal-to-noise ratio in low illumination and induces motion blur during fast movements. Such degradations occur at the sensing stage and propagate through pretrained visual encoders, ultimately affecting policy stability. While geometric or tactile extensions improve spatial reasoning and interaction modeling, they do not fundamentally address the imaging-level limitations of frame-based perception. As a result, the perceptual robustness of VLA models under extreme illumination and high-speed motion remains insufficiently explored.

2.2 Event-Based Perception in Robotics

Event cameras asynchronously record per-pixel brightness changes with microsecond latency and high dynamic range [14, 19]. By avoiding exposure-based temporal integration, event sensing naturally mitigates motion blur and maintains signal fidelity under extreme lighting conditions. These properties have motivated extensive research in event-based perception, including image reconstruction [7, 45], low-light enhancement [11, 35, 52], motion deblurring [39, 44, 48, 53], segmentation [1, 33], detection [10, 59], optical flow [21, 64, 66], and tracking [36, 60]. In robotics, event sensing has been applied to visual servoing and grasping [25, 30, 42], navigation [6, 47], slip detection [22, 46], and visual-tactile perception [18, 57]. However, most existing event-based robotic systems are designed for specific tasks and rely on modular pipelines with task-dependent representations and predefined control strategies. In contrast, modern VLA models aim to learn general-purpose, language-conditioned manipulation policies through large-scale multimodal pretraining. Despite the complementary sensing advantages of event cameras, their integration into scalable VLA frameworks remains largely unexplored. Bridging event-driven perception with pretrained VLA architectures therefore presents a promising direction for enhancing embodied robustness without sacrificing policy generality.

3 Methods

3.1 Overview

We build upon a SmolVLA-style [49] Vision-Language-Action framework and investigate how event-based visual sensing can be integrated in a lightweight and pretrained-compatible manner. Event streams captured by a DAVIS camera are

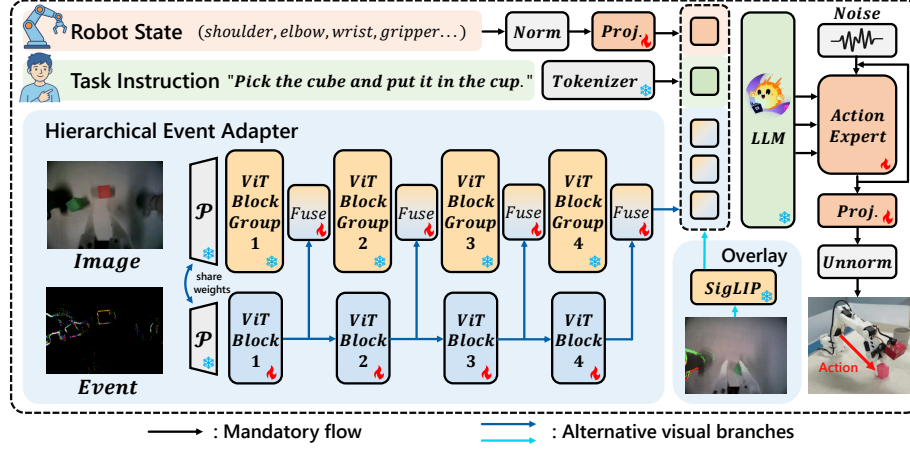


Fig. 2: An overview of our proposed E-VLA framework. Our architecture integrates event-based visual sensing with RGB frames and proprioceptive robot states to generate control sequences. We investigate two fusion strategies: (1) a Hierarchical Event Adapter that injects event features into intermediate layers of a frozen ViT encoder through trainable fusion modules, and (2) an Overlay strategy that directly combines events with RGB images prior to encoding via SigLIP. The resulting fusion visual tokens are concatenated with language tokens and state tokens and then processed by a frozen LLM backbone, which conditions an Action Expert to produce normalized robot actions. Snowflakes and flames denote frozen and trainable parameters, respectively.

spatially and temporally aligned with RGB frames and converted into frame-like representations through carefully designed windowing strategies, ensuring compatibility with standard visual encoders.

To incorporate event information without disrupting the pretrained token distribution, we investigate two alternative fusion designs that operate at different stages of the visual pipeline. The first design adopts a pre-encoding overlay strategy, directly combining accumulated event frames with RGB images before visual encoding, introducing no additional parameters. Instead of overlaying, the second design employs a lightweight hierarchical event adapter to inject event features into intermediate layers of the visual encoder for fine-grained cross-modal interaction. In each case, the resulting visual tokens are processed by the VLM backbone, and an action expert generates control sequences conditioned on the enriched perceptual representations. An overview of our E-VLA framework is depicted in Fig. 2 and a system schematic is shown in Fig. 3. The following sections detail the baseline architecture, event windowing and representation design, and the proposed fusion strategies.

3.2 SmolVLA Baseline

We adopt SmolVLA [49] as the image-based policy baseline to have a good balance between performance and efficiency. SmolVLA is an open-source lightweight vision-language-action model trained on community-contributed datasets with

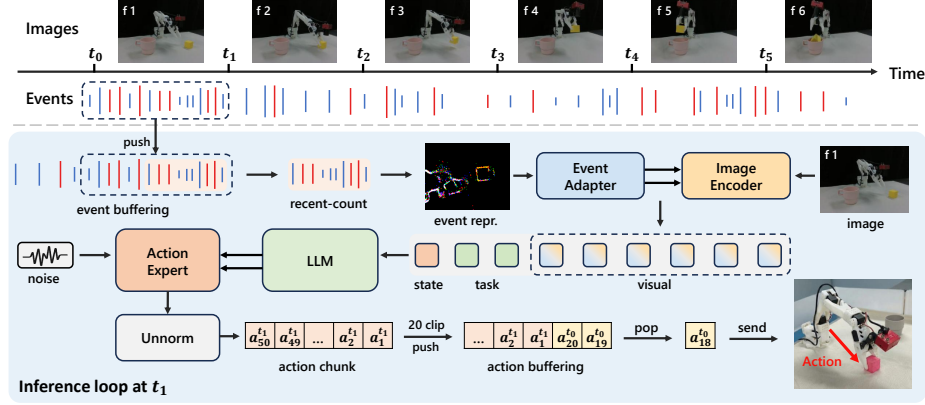


Fig. 3: A system schematic of our E-VLA pipeline. Incoming events are first buffered and aggregated via a recent-count representation, then converted into an event representation and aligned through an event adapter with visual features extracted by an image encoder. The fused multimodal visual tokens are then fed into LLM, which interacts with an action expert to generate chunked action outputs. The predicted actions are stored in an action buffer before being executed in a streaming manner.

about 0.5B parameters. It offers strong reproducibility compared to larger or closed-source VLA models, while maintaining competitive performance in real-world robot deployment. SmolVLA consists of a compact vision-language backbone (SmolVLM [41]) and a lightweight transformer-based action expert. Using SigLIP as its vision encoder, SmolVLM is pretrained on large-scale text-image and video data, encoding multi-view RGB observations, language instructions, and robot states into a unified token representation. The action expert then generates action chunks conditioned on SmolVLM features through the interaction design of interleaved self-attention and cross-attention. SmolVLA supports synchronous and asynchronous inference modes. In synchronous inference, a new policy inference process is only triggered when the current action queue is empty, which has a low perception delay while introducing execution pauses due to inference latency. In contrast, asynchronous mode performs inference in parallel with action execution, eliminating explicit waiting between action chunks while prolonging its perceptual delay. Overall, this established image baseline provides a solid foundation for our event-augmented VLA approach.

3.3 Event Windowing and Representation

Event cameras produce asynchronous and spatially sparse event streams, which record pixel-level brightness changes. The output is formulated as tuples:

$$\mathcal{E} = \{e_i\}_{i=1}^M, \quad e_i = (x_i, y_i, t_i, p_i). \quad (1)$$

where x_i and y_i denote the pixel coordinates of the event, t_i denotes its timestamp, and $p_i \in +1, -1$ is the polarity of the brightness change. The output

of this format cannot be directly consumed by standard vision-language models. Prior event-based vision work has explored windowing and representation strategies to convert raw event streams into structured forms, including fixed-duration windows and fixed-count windows for sampling, as well as event frames, voxel grids [66], time surfaces [29], and other specific methods [1, 56] for representation. Fixed-duration windowing is widely used in event-based methods [7, 21, 33, 36, 59] to maintain a frame rate or to synchronize with image exposure time.

However, in the VLA setting, the temporal distribution of event streams becomes highly unstable due to the tight coupling between perception and action. The wrist-mounted camera undergoes continuous ego-motion induced by robot actions, and the event rate varies substantially with the frequently changing motion speed across task phases. Under such conditions, event windowing becomes non-trivial: fixed-duration windows may produce inconsistent event responses, leading to perception failure at low speeds and motion blur at high speeds. Since the suitability of fixed-duration versus fixed-count windowing in VLA scenarios remains unclear, we empirically examine these strategies and adopt a *recent-count* scheme that retains the most recent N events, which can be formulated as Eq. 2, where t_e denotes the exposure end time for the current image $I \in \mathbb{R}^{H \times W \times 3}$, $\mathcal{W}_I$ denotes the event window corresponding to I .

$$\mathcal{E}_{t_e} = \{e_i \mid t_i \leq t_e\}, \quad \mathcal{W}_I = \{e_k\}_{k=|\mathcal{E}_{t_e}|-N+1}^{|\mathcal{E}_{t_e}|}. \quad (2)$$

This model demonstrates better event stability in practice. The DAVIS sensor provides events and RGB images with microsecond-resolution timestamps, enabling precise temporal alignment for implementing this windowing operation.

Compared to large-scale image-text and video datasets, event data remain limited in both scale and diversity, making large-scale pretraining or contrastive learning directly on event modalities impractical. To effectively leverage the knowledge embedded in pretrained VLMs, we adopt an image-like accumulated event representation that closely matches the input distribution of the SigLIP image encoder used in SmolVLA. Within each window, events are accumulated into a dense gray frame $\tilde{E} \in \mathbb{R}^{H \times W}$ via a polarity-agnostic way: the total amount of events is added up at each pixel regardless of polarity. This representation emphasizes event density and the absolute position while discarding the direction of intensity change, resulting in a structurally stable input format. We use a colored DAVIS346 camera equipped with a Bayer filter, enabling the capture of RGB event streams. So the accumulated event frames are first normalized to $[0, 1]$ and then demosaiced to obtain three-channel representations $E \in \mathbb{R}^{H \times W \times 3}$. The whole process can be formulated as Eq. 3, where $\delta(x, y)$ denotes the 2D Kronecker delta function, $\text{Norm}(\cdot)$ denotes the min-max normalization, and $\mathcal{D}(\cdot)$ denotes the demosaicing operation.

$$\tilde{E}(x, y) = \sum_{e_i \in \mathcal{W}_I} \delta(x - x_i, y - y_i), \quad E = \mathcal{D}(\text{Norm}(\tilde{E})). \quad (3)$$

3.4 Event Fusion Strategy

In the VLA architecture, visual observations are encoded into a sequence of tokens by a ViT-style encoder, then concatenated with language instruction tokens and robot state tokens before being fed into the LLM. Under this formulation, naively introducing additional event tokens would increase the token length and computational cost, and potentially disrupt the pretrained token distribution of the visual encoder. To preserve the original token length and leverage pretrained visual-language knowledge, we fuse event information into the existing visual tokens, rather than adding new event tokens.

We explore two event fusion strategies that operate at different stages of the visual processing pipeline: an overlay-based fusion that combines events and images before visual encoding, and a lightweight hierarchical event adapter that integrates event features within the visual encoder.

Overlay-based fusion. Given the spatial alignment between frame images and events provided by the DAVIS camera, a simple pre-encoding fusion can be implemented by directly overlaying the windowed events onto RGB images. Let $\mathcal{E}_{(x,y)}$ denotes the events at position (x, y) within the window:

$$\mathcal{E}_{(x,y)} = \{e_i \mid e_i \in \mathcal{W}_I, (x_i, y_i) = (x, y)\}. \quad (4)$$

The overlay process can be formulated as Eq. 5, where I° donates the overlay image, $\mathbf{c}(p_j)$ donates the polarity-to-color projection.

$$I^\circ(x, y) = \begin{cases} I(x, y), & |\mathcal{E}_{(x,y)}| = 0, \\ \mathbf{c}(p_j), & j = \operatorname{argmax}_i t_i, e_i \in \mathcal{E}_{(x,y)}. \end{cases} \quad (5)$$

This approach directly injects events into the visual input before ViT encoding, enriching the visual information without adding parameters, making it a plug-and-play and computationally efficient implementation.

Hierarchical event adapter. To enable progressive and fine-grained fusion of event and image information, we further propose a lightweight hierarchical event adapter that integrates event features into the image encoder during visual encoding. Specifically, accumulated event frames E are processed by a compact ViT-style event adapter which consists of a 16×16 weight-sharing patch embedding layer (sharing weights for closer feature distribution) and four stacked Transformer blocks. The outputs of these blocks are progressively fused with the corresponding intermediate features of the original visual encoder. For SigLIP in SmolVLA, we choose 3, 6, 9, and 12 as fusion layers. Let $\mathcal{G}_l(\cdot)$ denotes the l -th layer of the event adapter, $\mathcal{F}_l(\cdot)$ denote the l -th group of the SigLIP encoder layers, $E^{(l)}$ and $F^{(l)}$ denote their output features, respectively, and $\mathcal{P}(\cdot)$ denotes the weight-sharing patch embedding. Initially, we have $E^{(0)} = \mathcal{P}(E)$, $F^{(0)} = \mathcal{P}(I)$. Then the overall hierarchical fusion process can be formulated as Eq. 6, where $\text{Fuse}(\cdot)$ is implemented as $\text{MLP}(\text{Concat}(\cdot))$.

$$E^{(l+1)} = \mathcal{G}_{l+1}(E^{(l)}), \quad F^{(l+1)} = \mathcal{F}_{l+1}(\text{Fuse}(F^{(l)}, E^{(l)})). \quad (6)$$

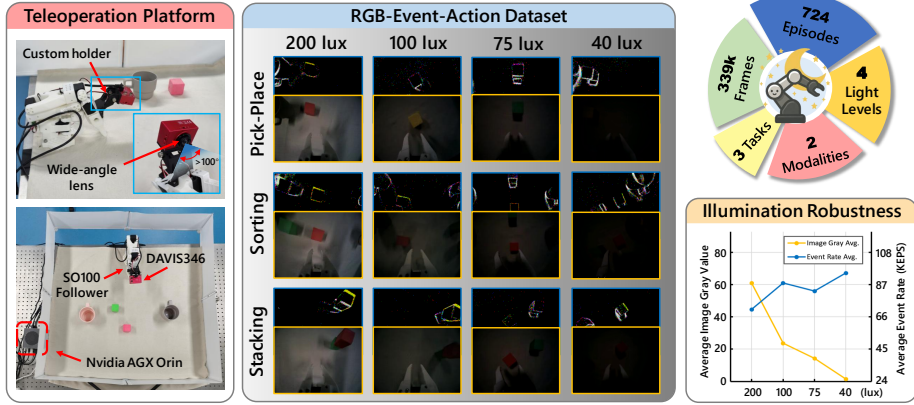


Fig. 4: Middle: The visualization of the proposed dataset. Events are represented as colored frames following Sec. 3.3. Left: Side and top views of our teleoperation platform based on LeRobot SO100 manipulator [8] and DAVIS346 event camera. Right: Above are the statistics of our dataset. The line chart below shows that even when the image signal rapidly decays with decreasing illumination, the event modality can still maintain a stable event rate ($\sim 87k$ events per second or KEPS).

The introduced event adapter and fusion module adopt smaller hidden dimensions and fewer layers, which only introduce 13M additional parameters, less than 3% of the total model, enabling a fine-grained and multi-level event-image fusion while preserving computational efficiency.

4 Dataset

4.1 Event-Augmented Teleoperation Platform

To support event-augmented VLA learning and real-world evaluation under visually degraded conditions, we build a unified teleoperation platform on the open-source SO100 6-DoF robotic arm [8], using structurally identical leader-follower arms for demonstration collection and closed-loop testing. The follower arm is equipped with an angle-driven gripper as the end-effector, and a DAVIS346 event camera is rigidly mounted on the wrist via a custom holder, providing pixel-aligned asynchronous events and RGB frames at 346×260 resolution with microsecond-level timestamps. Instead of adopting a multi-view setup, we deliberately use a monocular wrist-view camera to preserve system compactness and deployment practicality, while naturally coupling arm motion with event generation, providing an appropriate setting for integrating event cues to explore perception-action robustness under visual degradation. In practice, we find that a wide field of view is critical for wrist-view manipulation, as narrow-view lenses frequently miss task-relevant context; we therefore replace the original pinhole lens with a compact wide-angle lens (2.2 mm focal length, F/1.8), achieving a field of view greater than 100° with suitable depth of field for typical manipulation distances. The DAVIS camera operates with fixed hardware gain, all frames

are recorded at 30FPS, and unless otherwise specified, we use a 10 ms exposure time. To reflect realistic edge deployment constraints, both data collection and on-robot inference are conducted on an NVIDIA AGX Orin platform.

4.2 RGB-Event-Action Dataset

Existing event-based datasets [2, 20, 34, 43, 51, 54] mainly focus on perception tasks (*e.g.*, deblurring, optical flow, or reconstruction), while real-world event datasets for imitation learning and VLA-style manipulation remain extremely limited. To bridge this gap, we collect a new event-augmented manipulation dataset using the teleoperation system above; to the best of our knowledge, this is among the first real-world datasets designed specifically for studying synchronized *RGB-event-action* learning in language-conditioned manipulation under challenging illumination. Following SmolVLA [49], we include three representative tasks: **Pick-Place**, **Sorting**, and **Stacking**. Pick-Place requires grasping an object from a random location and placing it into a fixed cup; Sorting additionally requires color-based discrimination to place objects into designated cups; and Stacking requires precise placement of one cube on top of another under a constrained gripper-centric field of view. For each task, demonstrations are collected under one normal illumination condition (200 lux) and three lower illumination levels (100, 75, and 40 lux), with both fixed and randomized initial arm configurations to improve diversity and robustness. Due to the low hardware gain of the DAVIS camera, bright scenes under 200 lux still produce slightly dark images with about 60 of 255 grayscale value as shown in Fig. 4. All data are stored in the LeRobot format, where RGB frames and event streams are temporally aligned by camera timestamps and synchronized with robot states and actions. In summary, our dataset comprises 724 episodes (305 under normal illumination and 419 under three progressively decreasing low-light levels), totaling 339310 frames. Fig. 4 provides an overview of the dataset composition and capture setup.

5 Experiment

5.1 Implementation Details

We compare E-VLA against four representative alternatives: RetinexNet, Retinexformer, EvLight, and E2VID, using official pretrained weights and standard implementations. RetinexNet, Retinexformer, and EvLight are treated as low-light image enhancement baselines, where low-illumination images are enhanced offline for training and processed with the same enhancement pipeline at inference time; E2VID is used as an event-to-image reconstruction baseline, where event streams are first converted into color images and then introduced as additional visual inputs during both training and testing. For fair comparison, the VLM backbone is kept frozen in all settings, and imitation learning updates only the action expert and projection MLP layers. For our event-adaptor variant, we also

freeze the VLM backbone to preserve pretrained visual-language priors, and train the event branch with a two-stage strategy: we first optimize the newly introduced adapter from scratch, and then jointly fine-tune the adapter together with the action expert and projection layers. To avoid shortcut learning from RGB inputs and encourage effective event utilization, we apply a random dropout to the image branch during training. Unless otherwise specified, all methods are fine-tuned on the same multi-task, multi-illumination dataset for 20k iterations, and training is performed on a single NVIDIA A800 GPU.

5.2 Evaluation Protocol

We evaluate all methods on the same real-world teleoperation platform (SO100 + DAVIS346) over three manipulation tasks (Pick-Place, Sorting, Stacking) under multiple illumination conditions, with policy inference executed on an NVIDIA AGX Orin. For fair comparison, we sample task-specific test configurations once and keep them fixed: Pick-Place uses 10 test locations, and Sorting/Stacking uses 10 paired configurations with randomly swapped object placements. We adopt partial-credit scoring (Pick-Place: pick/place; Sorting: per-object pick/place; Stacking: pick/stack). We use synchronous inference with an action chunk size of 40 at 30 Hz (predicting ~ 1.33 s actions per forward pass). Success is measured within 1 minute for Pick-Place/Stacking and 1.5 minutes for Sorting. Additional implementation details are provided in the supplementary material.

Table 1: Task success rates (%). Higher is better. Results are reported under different illumination levels (lux) on the Pick-Place task.

Method	75 lux	40 lux	35 lux	30 lux	25 lux	20 lux	Average
Image (10ms exp.)	100	80	70	35	0	0	47.5
Image+RetinexNet [62]	100	100	85	80	25	10	66.7
Image+Retinexformer [9]	100	80	80	75	20	10	60.8
Image+EvLight [35]	100	95	95	75	45	10	70.0
Image+E2VID [45]	80	60	55	10	5	5	35.8
Ours overlay	100	100	85	75	65	60	<u>80.8</u>
Ours event adapter	100	100	95	90	90	90	94.2

Success rates (%): ■ ≥ 90 (Best) ■ $80 \sim 90$ ■ $70 \sim 80$ ■ $60 \sim 70$ ■ $50 \sim 60$ ■ ≤ 50 (Bad)

5.3 Low Illumination Performance

We evaluate task success rates under progressively reduced ambient illumination for image-only policies, image-based enhancement baselines, and the proposed E-VLA models (Pick-Place results in Tab. 1). Under well-lit settings, event integration does not hurt performance: at 75 lux all methods expect E2VID reach 100%, and at 40 lux both overlay fusion and the event adapter achieve 100% (vs. 80% image-only).

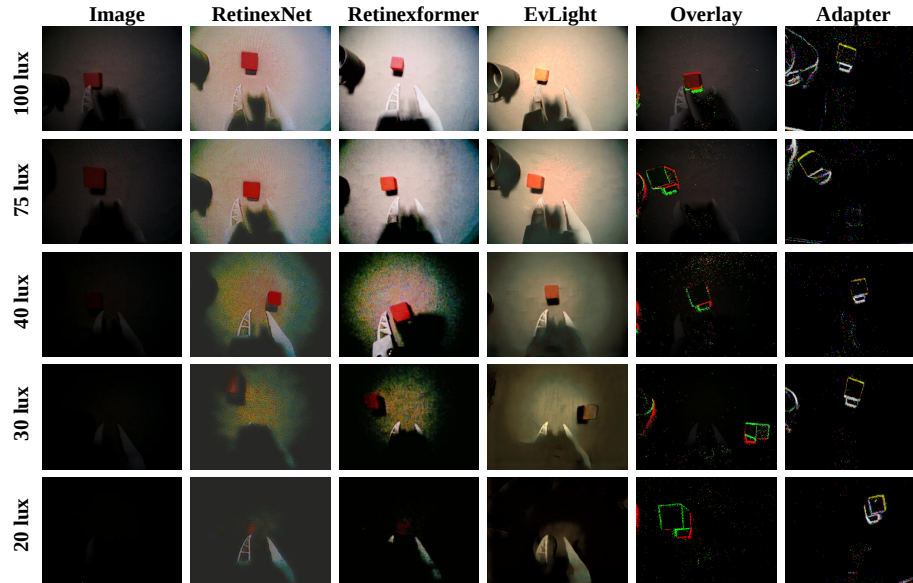


Fig. 5: Qualitative comparison of visual inputs under different illumination.

As illumination drops, the image-only baseline degrades from 70% (35 lux) to 35% (30 lux) and collapses to 0% (25/20 lux). Image enhancement helps only moderately (e.g., RetinexNet reaches 80% at 30 lux) but deteriorates at lower light (20–45% at 25 lux and 5–10% at 20 lux for several baselines), consistent with Fig. 5. In contrast, E-VLA is markedly more robust: overlay achieves 75% at 30 lux and 65%/60% at 25/20 lux, while the event adapter reaches 90% at 30 lux and maintains 90% at 25/20 lux. When frames are nearly black-clipped (25/20 lux), image-only fails (0%), whereas event-based variants retain substantial success (65–60% overlay; 90% adapter), showing that event streams provide actionable cues even when frames are uninformative. At even lower illumination, performance only drops near ~ 2 lux (Tab. S1, supplementary material), where event noise begins to dominate.

5.4 Generalization without Low-light Training

We further test OOD illumination generalization by training all models only on well-lit (200 lux) demonstrations and evaluating at lower lux (Tab. 2). The image-only policy drops from 65% (75 lux) to 30% (40 lux) and collapses to 0% (30/20 lux). In contrast, the event adapter remains robust: 85% (75 lux), 75% (40 lux), and 70%/45% (30/20 lux). This shows event-based perception improves robustness to unseen illumination shifts and reduces the need for extensive low-light data collection.

Table 2: Task success rates (%) for Out-Of-Distribution (OOD) low-light conditions on the Pick-Place task. In-Distribution (ID) refers to evaluating on low-light scenes that are included in training, for which the success rates are reported in Tab. 1. OOD refers to evaluation on unseen low-light scenes.

Method	100 lux		75 lux		40 lux		30 lux		20 lux		Average
	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD	OOD
Image (10ms exp.)	100	100	100	65	80	30	35	0	0	0	39.0
Image+Retinexformer [9]	100	100	100	70	80	50	75	10	10	0	46.0
Image+EvLight [35]	100	80	100	50	95	50	75	20	10	5	40.5
Ours overlay	100	100	100	95	100	55	75	10	60	10	54.0
Ours adapter	100	100	100	85	100	75	90	70	90	45	75.0

Success rates (%): ■ ≥ 90 (Best) ■ $80 \sim 90$ ■ $70 \sim 80$ ■ $60 \sim 70$ ■ $50 \sim 60$ ■ ≤ 50 (Bad)

5.5 Motion Blur Performance

We evaluate robustness under increasing motion blur by jointly adjusting illumination and exposure time (Tab. 3). Under mild blur, methods are comparable. As blur increases, image-only performance drops sharply: at 300 ms, overlay improves Pick-Place from 40% to 90% and Sorting from 52.5% to 72.5%, while the event adapter reaches 85%/85%. Under severe blur (1000 ms), image-only falls to 0% (Pick-Place) and 5% (Sorting), whereas overlay achieves 20%/32.5% and the event adapter further improves Pick-Place to 25% (Sorting: 32.5%). All models are trained without motion-blurred data, indicating gains stem from event sensing (high temporal resolution + recent-count windowing) rather than blur-specific adaptation, and mitigating the low-light/exposure trade-off of frame imaging.

Table 3: Task success rates (%). Higher is better. Results are reported under different motion blur levels (represented by exposure time) on Pick-Place and Sorting task.

Method	100 ms		200 ms		300 ms		400 ms		500 ms		1000 ms		Average	
	P&P	Sorting	P&P	Sorting	P&P	Sorting	P&P	Sorting	P&P	Sorting	P&P	Sorting	P&P	Sorting
Image	100	77.5	85	75.0	40	52.5	30	32.5	15	20.0	0	5.0	45.0	43.8
Ours overlay	100	95.0	100	95.0	90	72.5	60	70.0	60	55.0	20	32.5	71.7	70.0
Ours adapter	100	95.0	100	85.0	85	85.0	85	72.5	50	45.0	25	32.5	74.2	69.2

Success rates (%): ■ ≥ 90 (Best) ■ $80 \sim 90$ ■ $70 \sim 80$ ■ $60 \sim 70$ ■ $50 \sim 60$ ■ ≤ 50 (Bad)

5.6 Ablation Studies

We analyze key design choices for event-based VLA, including event windowing and training recipes.

Event windowing strategy and size. Tab. 4 shows time-based windows (5/20/40 ms) are unreliable under synchronous manipulation due to pauses and

Table 4: Task success rates (%) for different event windows. Higher is better. Results are reported under different illumination levels (lux) on the Pick-Place task.

Event Window	75 lux	40 lux	35 lux	30 lux	25 lux	20 lux	Average
5 ms window	80	70	65	55	50	45	60.8
20 ms window	100	85	85	80	75	70	82.5
40 ms window	100	90	90	75	50	45	75.0
500 events	100	100	60	50	50	45	67.5
2000 events	100	100	95	90	90	90	94.2
4000 events	100	100	85	85	75	70	<u>85.8</u>

Success rates (%): ■ ≥ 90 (Best) ■ $80 \sim 90$ ■ $70 \sim 80$ ■ $60 \sim 70$ ■ $50 \sim 60$ ■ ≤ 50 (Bad)

speed changes, while event-count windows are more stable. A moderate window performs best: 2000 events yields 95% at 35 lux and 90% at 30/25/20 lux, outperforming 500 events (60% at 35 lux; 45% at 20 lux) and 4000 events (85% at 35 lux; 70% at 20 lux).

Training recipe and weight-sharing. Tab. 5 confirms progressive training is critical: Action→Joint achieves 75%, while Event→Joint drops to 50%. Using both Action and Event stages improves to 80% (Action→Event→Joint, no sharing), and the full recipe with patch-embedding weight-sharing reaches 90%, supporting stable perception–action alignment and cross-modal integration.

Table 5: Task success rates (%) on different training recipes at 30 lux. Higher is better. ‘sharing’ represents weight-sharing for the patch embedding layer.

Strategy	Action	Event	Joint	Sharing	Success
Action→Joint	✓	✗	✓	✓	75.0
Event→Joint	✗	✓	✓	✓	50.0
Action → Event→Joint	✓	✓	✓	✗	80.0
Ours	✓	✓	✓	✓	90.0

5.7 Failure Cases: Color Semantics and Occlusion

We further investigate Sorting and Stacking, which introduce task-specific perceptual challenges beyond those in Pick-Place.

1) Color-based sorting. Unlike the Pick-Place task, the Sorting task additionally requires correct color-based placement. Although grasping remains reliable, failures mainly occur during placement due to the lower color classification accuracy of event-based methods when RGB frames are degraded. This limitation is inherent to event sensing: events encode intensity changes rather than absolute appearance, static object color vanishes after grasp stabilization, and sparse RGB events do not satisfy dense demosaicing assumptions. It suggests that,

while event cues are effective for maintaining object-level perception under low illumination, tasks that require reliable color discrimination remain supported by intensity-based visual signals. **2) Stacking with occlusion.** In the Stacking task, failures primarily arise during the stack phase due to the viewpoint occlusion caused by the grasped object blocking the wrist-mounted camera [40]. Therefore, the availability of event information does not substantially alter the perceptual bottleneck. Overall, these results indicate that E-VLA is effective for low-light manipulation tasks. Limitations observed in color-driven behaviors point to representation-level challenges that may be addressed with more advanced event-based color reconstruction, whereas failures under strong viewpoint occlusion reflect perceptual constraints that are not directly mitigated by event sensing. More detailed success rate analyses can be found in the supplementary material.

6 Conclusions

We present E-VLA, an event-augmented vision-language-action (VLA) framework that improves VLA robustness under visual degradations, including extreme low light, black clipping, and motion blur. Instead of reconstructing images from events, E-VLA integrates asynchronous event streams into a pre-trained VLA architecture in a lightweight, token-compatible manner, enabling robust manipulation without large-scale event pretraining. To facilitate systematic study, we build an open-source teleoperation platform with a DAVIS346 event camera and collect a synchronized RGB–event–action manipulation dataset spanning tasks, illumination levels, and motion regimes. Experiments show that even a parameter-free fusion (*i.e.*, overlaying accumulated event maps onto RGB images) improves success when frame inputs become unreliable, while preserving performance in bright scenes. On Pick-Place at 20 lux, success increases from 0% (image-only) to 60% with overlay fusion and to 90% with our event adapter; under severe motion blur (1000 ms exposure), Pick-Place improves from 0% to 20–25%, and Sorting from 5% to 32.5%. Moreover, E-VLA improves out-of-distribution generalization across illumination shifts and blur-heavy dynamics. Our ablations provide design insights for coupling event perception with vision-pretrained policy models, highlighting event windowing/accumulation, representation choice, and training balance for stable perception–action alignment. We hope this work encourages adoption of event-driven sensing in VLA learning and inspires robust embodied intelligence beyond frame-based imaging.

Acknowledgments

This research was funded by the Natural Science Foundation of Zhejiang Province (Grant No. LZ24F050003), the National Natural Science Foundation of China (Grant No. 62473139), the Hunan Provincial Research and Development Project (Grant No. 2025QK3019), and the opening project of the State Key Laboratory of Autonomous Intelligent Unmanned Systems (Grant No. ZZKF2025-2-10).

References

1. Alonso, I., Murillo, A.C.: EV-SegNet: Semantic segmentation for event-based cameras. In: CVPRW (2019)
2. Bao, Y., Sun, L., Ma, Y., Wang, K.: Temporal-mapping photography for event cameras. In: ECCV (2024)
3. Bi, J., Ma, K.Y., Hao, C., Shou, M.Z., Soh, H.: VLA-Touch: Enhancing vision-language-action models with dual-level tactile feedback. arXiv preprint arXiv:2507.17294 (2025)
4. Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., Jakubczak, S., Jones, T., Ke, L., Levine, S., Li-Bell, A., Mothukuri, M., Nair, S., Pertsch, K., Shi, L.X., Tanner, J., Vuong, Q., Walling, A., Wang, H., Zhilinsky, U.: π_0 : A vision-language-action flow model for general robot control. arXiv preprint arXiv:2410.24164 (2024)
5. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., Ibarz, J., Ichter, B., Irpan, A., Jackson, T., Jesmonth, S., Joshi, N.J., Julian, R., Kalashnikov, D., Kuang, Y., Leal, I., Lee, K., Levine, S., Lu, Y., Malla, U., Manjunath, D., Mordatch, I., Nachum, O., Parada, C., Peralta, J., Perez, E., Pertsch, K., Quiambao, J., Rao, K., Ryoo, M.S., Salazar, G., Sanketi, P.R., Sayed, K., Singh, J., Sontakke, S., Stone, A., Tan, C., Tran, H.T., Vanhoucke, V., Vega, S., Vuong, Q., Xia, F., Xiao, T., Xu, P., Xu, S., Yu, T., Zitkovich, B.: RT-1: Robotics transformer for real-world control at scale. In: RSS (2022)
6. Bugueño-Córdova, I.G., Ruiz-del-Solar, J., Verschae, R.: Human-robot navigation using event-based cameras and reinforcement learning. In: CVPRW (2025)
7. Cadena, P.R.G., Qian, Y., Wang, C., Yang, M.: SPADE-E2VID: Spatially-adaptive denormalization for event-based video reconstruction. IEEE Transactions on Image Processing (2021)
8. Cadene, R., Alibert, S., Soare, A., Gallouedec, Q., Zouitine, A., Palma, S., Kooijmans, P., Aractingi, M., Shukor, M., Aubakirova, D., Russi, M., Capuano, F., Pascal, C., Choghari, J., Moss, J., Wolf, T.: LeRobot: State-of-the-art machine learning for real-world robotics in pytorch. <https://github.com/huggingface/lerobot> (2024)
9. Cai, Y., Bian, H., Lin, J., Wang, H., Timofte, R., Zhang, Y.: Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In: ICCV (2023)
10. Cao, J., Zheng, X., Lyu, Y., Wang, J., Xu, R., Wang, L.: Chasing day and night: Towards robust and efficient all-day object detection guided by an event camera. In: ICRA (2024)
11. Chen, K., Liang, G., Lu, Y., Li, H., Wang, L.: EvLight++: Low-light video enhancement with an event camera: A large-scale real-world dataset, novel method, and more. IEEE Transactions on Pattern Analysis and Machine Intelligence (2026)
12. Chen, Z., Niu, R., Kong, H., Wang, Q., Xing, Q., Fan, Z.: TGRPO: Fine-tuning vision-language-action model via trajectory-wise group relative policy optimization. arXiv preprint arXiv:2506.08440 (2025)
13. Cheng, Z., Zhang, Y., Zhang, W., Li, H., Wang, K., Song, L., Zhang, H.: OmniVTLA: Vision-tactile-language-action model with semantic-aligned tactile sensing. arXiv preprint arXiv:2508.08706 (2025)
14. Delbrück, T.: Frame-free dynamic digital vision. In: International Symposium on Secure-Life Electronics, Advanced Electronics for Quality Life and Society (2008)

15. Deng, H., Wu, Z., Liu, H., Guo, W., Xue, Y., Shan, Z., Zhang, C., Jia, B., Ling, Y., Lu, G.: A survey on reinforcement learning of vision-language-action models for robotic manipulation. *Authorea Preprints* (2025)
16. Deng, S., Yan, M., Zheng, Y., Su, J., Zhang, W., Zhao, X., Cui, H., Zhang, Z., Wang, H.: StereoVLA: Enhancing vision-language-action models with stereo vision. *arXiv preprint arXiv:2512.21970* (2025)
17. Fei, S., Wang, S., Shi, J., Dai, Z., Cai, J., Qian, P., Ji, L., He, X., Zhang, S., Fei, Z., Fu, J., Gong, J., Qiu, X.: LIBERO-Plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626* (2025)
18. Funk, N., Helmut, E., Chalvatzaki, G., Calandra, R., Peters, J.: Evetac: An event-based optical tactile sensor for robotic manipulation. *IEEE Transactions on Robotics* (2024)
19. Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A.J., Conradt, J., Daniilidis, K., Scaramuzza, D.: Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022)
20. Gehrig, M., Aarents, W., Gehrig, D., Scaramuzza, D.: DSEC: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters* (2021)
21. Gehrig, M., Millh usler, M., Gehrig, D., Scaramuzza, D.: E-RAFT: Dense optical flow from event cameras. In: *3DV* (2021)
22. Guo, Q., Yu, Z., Fu, J., Lu, Y., Zweiri, Y., Gan, D.: Force-EvT: A closer look at robotic gripper force measurement with event-based vision transformer. In: *ReMAR* (2024)
23. Guo, Y., Zhang, J., Chen, X., Ji, X., Wang, Y.J., Hu, Y., Chen, J.: Improving vision-language-action model with online reinforcement learning. *arXiv preprint arXiv:2501.16664* (2025)
24. Huang, J., Wang, S., Lin, F., Hu, Y., Wen, C., Gao, Y.: Tactile-VLA: Unlocking vision-language-action model’s physical knowledge for tactile generalization. *arXiv preprint arXiv:2507.09160* (2025)
25. Huang, X., Halwani, M., Muthusamy, R., Ayyad, A., Swart, D., Seneviratne, L., Gan, D., Zweiri, Y.: Real-time grasping strategies using event camera. *Journal of Intelligent Manufacturing* (2022)
26. Intelligence, P., Amin, A., Aniceto, R., Balakrishna, A., Black, K., Conley, K., Connors, G., Darpinian, J., Dhabalia, K., DiCarlo, J., Driess, D., Equi, M., Esmail, A., Fang, Y., Finn, C., Glossop, C., Godden, T., Goryachev, I., Groom, L., Hancock, H., Hausman, K., Hussein, G., Ichter, B., Jakubczak, S., Jen, R., Jones, T., Katz, B., Ke, L., Kuchi, C., Lamb, M., LeBlanc, D., Levine, S., Li-Bell, A., Lu, Y., Mano, V., Mothukuri, M., Nair, S., Pertsch, K., Ren, A.Z., Sharma, C., Shi, L.X., Smith, L., Springenberg, J.T., Stachowicz, K., Stoeckle, W., Swerdlow, A., Tanner, J., Torne, M., Vuong, Q., Walling, A., Wang, H., Williams, B., Yoo, S., Yu, L., Zhilinsky, U., Zhou, Z.: $\pi_{0.6}^*$: a VLA that learns from experience. *arXiv preprint arXiv:2511.14759* (2025)
27. Intelligence, P., Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Galliker, M.Y., Ghosh, D., Groom, L., Hausman, K., Ichter, B., Jakubczak, S., Jones, T., Ke, L., LeBlanc, D., Levine, S., Li-Bell, A., Mothukuri, M., Nair, S., Pertsch, K., Ren, A.Z., Shi, L.X., Smith, L., Springenberg, J.T., Stachowicz, K., Tanner, J., Vuong, Q., Walke, H., Walling, A., Wang, H., Yu, L., Zhilinsky, U.: $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054* (2025)

28. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E.P., Sanketi, P.R., Vuong, Q., Kollar, T., Burchfiel, B., Tedrake, R., Sadigh, D., Levine, S., Liang, P., Finn, C.: OpenVLA: An open-source vision-language-action model. In: CoRL (2024)
29. Lagorce, X., Orchard, G., Galluppi, F., Shi, B.E., Benosman, R.: HOTS: A hierarchy of event-based time-surfaces for pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017)
30. Li, B., Cao, H., Qu, Z., Hu, Y., Wang, Z., Liang, Z.: Event-based robotic grasping detection with neuromorphic vision sensor and event-grasping dataset. *Frontiers in Neurorobotics* (2020)
31. Li, C., Wen, J., Peng, Y., Peng, Y., Feng, F., Zhu, Y.: PointVLA: Injecting the 3D world into vision-language-action models. *arXiv preprint arXiv:2503.07511* (2025)
32. Li, H., Chen, Y., Cui, W., Liu, W., Liu, K., Zhou, M., Zhang, Z., Zhao, D.: Survey of vision-language-action models for embodied manipulation. *arXiv preprint arXiv:2508.15201* (2025)
33. Li, H., Wang, J., Yuan, J., Li, Y., Weng, W., Peng, Y., Zhang, Y., Xiong, Z., Sun, X.: Event-assisted low-light video object segmentation. In: CVPR (2024)
34. Li, Y., Shen, Y., Huang, Z., Chen, S., Bian, W., Shi, X., Wang, F.Y., Sun, K., Bao, H., Cui, Z., Zhang, G., Li, H.: BlinkVision: A benchmark for optical flow, scene flow and point tracking estimation using rgb frames and events. In: ECCV (2024)
35. Liang, G., Chen, K., Li, H., Lu, Y., Wang, L.: Towards robust event-guided low-light image enhancement: A large-scale real-world event-image dataset and novel approach. In: CVPR (2024)
36. Liu, J., Wang, B., Tan, Z., Zhang, J., Shen, H., Hu, D.: Tracking any point with frame-event fusion network at high frame rate. In: IROS (2025)
37. Liu, S., Wu, L., Li, B., Tan, H., Chen, H., Wang, Z., Xu, K., Su, H., Zhu, J.: RDT-1B: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864* (2024)
38. Lu, G., Guo, W., Zhang, C., Zhou, Y., Jiang, H., Gao, Z., Tang, Y., Wang, Z.: VLA-RL: Towards masterful and general robotic manipulation with scalable reinforcement learning. *arXiv preprint arXiv:2505.18719* (2025)
39. Lu, Y., Liang, G., Wang, Y., Wang, L., Xiong, H.: UniINR: Event-guided unified rolling shutter correction, deblurring, and interpolation. In: ECCV (2025)
40. Ma, Y., Song, Z., Zhuang, Y., Hao, J., King, I.: A survey on vision-language-action models for embodied AI. *arXiv preprint arXiv:2405.14093* (2024)
41. Marafioti, A., Zohar, O., Farré, M., Noyan, M., Bakouch, E., Cuenca, P., Zakka, C., Allal, L.B., Lozhkov, A., Tazi, N., Srivastav, V., Lochner, J., Larcher, H., Morlon, M., Tunstall, L., von Werra, L., Wolf, T.: SmolVLM: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299* (2025)
42. Muthusamy, R., Ayyad, A., Halwani, M., Swart, D., Gan, D., Seneviratne, L., Zweiri, Y.: Neuromorphic eye-in-hand visual servoing. *IEEE Access* (2021)
43. Nah, S., Hyun Kim, T., Mu Lee, K.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: CVPR (2017)
44. Pan, L., Scheerlinck, C., Yu, X., Hartley, R., Liu, M., Dai, Y.: Bringing a blurry frame alive at high frame-rate with an event camera. In: CVPR (2019)
45. Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D.: Events-to-video: Bringing modern computer vision to event cameras. In: CVPR (2019)
46. Reinold, T., Ghosh, S., Gallego, G.: Combined physics and event camera simulator for slip detection. In: WACVW (2025)

47. Sanyal, S., Joshi, A., Kosta, A., Roy, K.: Real-time neuromorphic navigation: Guiding physical robots with event-based sensing and task-specific reconfigurable autonomy stack (2025)
48. Shang, W., Ren, D., Zou, D., Ren, J.S., Luo, P., Zuo, W.: Bringing events into video deblurring with non-consecutively blurry frames. In: ICCV (2021)
49. Shukor, M., Aubakirova, D., Capuano, F., Kooijmans, P., Palma, S., Zouitine, A., Aractingi, M., Pascal, C., Russi, M., Marafioti, A., Alibert, S., Cord, M., Wolf, T., Cadène, R.: SmolVLA: A vision-language-action model for affordable and efficient robotics. arXiv preprint arXiv:2506.01844 (2025)
50. Singh, I., Goyal, A., Birchfield, S., Fox, D., Garg, A., Blukis, V.: OG-VLA: 3D-aware vision language action model via orthographic image generation. arXiv preprint arXiv:2506.01196 (2025)
51. Stoffregen, T., Scheerlinck, C., Scaramuzza, D., Drummond, T., Barnes, N., Klee-man, L., Mahony, R.E.: Reducing the sim-to-real gap for event cameras. In: ECCV (2020)
52. Sun, L., Bao, Y., Zhai, J., Liang, J., Zhang, Y., Wang, K., Paudel, D.P., Van Gool, L.: Low-light image enhancement using event-based illumination estimation. In: ICCV (2025)
53. Sun, L., Sakaridis, C., Liang, J., Jiang, Q., Yang, K., Sun, P., Ye, Y., Wang, K., Gool, L.V.: Event-based fusion for motion deblurring with cross-modal attention. In: ECCV (2022)
54. Sun, L., Sakaridis, C., Liang, J., Sun, P., Zhang, K., Cao, J., Jiang, Q., Wang, K., Van Gool, L.: Event-based frame interpolation with ad-hoc deblurring. In: CVPR (2023)
55. Sun, L., Xie, B., Liu, Y., Shi, H., Wang, T., Cao, J.: GeoVLA: Empowering 3D representations in vision-language-action models. arXiv preprint arXiv:2508.09071 (2025)
56. Sun, Z., Fu, X., Huang, L., Liu, A., Zha, Z.J.: Motion aware event representation-driven image deblurring. In: ECCV (2024)
57. Taunyazov, T., Sng, W., Lim, B., See, H., Kuan, J., Ansari, A.F., Tee, B.C.K., Soh, H.: Event-driven visual-tactile sensing and learning for robots. In: RSS (2020)
58. Team, G.R., Abeyruwan, S., Ainslie, J., Alayrac, J.B., Arenas, M.G., Armstrong, T., Balakrishna, A., Baruch, R., Bauza, M., Blokzijl, M., et al.: Gemini robotics: Bringing AI into the physical world. arXiv preprint arXiv:2503.20020 (2025)
59. Tomy, A., Paigwar, A., Mann, K.S., Renzaglia, A., Laugier, C.: Fusing event-based and RGB camera for robust object detection in adverse conditions. In: ICRA (2022)
60. Wang, X., Yu, H., Yu, L., Yang, W., Xia, G.S.: Towards robust keypoint detection and tracking: A fusion approach with event-aligned image features. IEEE Robotics and Automation Letters (2024)
61. Wang, Y., Ding, P., Li, L., Cui, C., Ge, Z., Tong, X., Song, W., Zhao, H., Zhao, W., Hou, P., Huang, S., Tang, Y., Wang, W., Zhang, R., Liu, J., Wang, D.: VLA-Adapter: An effective paradigm for tiny-scale vision-language-action model. arXiv preprint arXiv:2509.09372 (2025)
62. Wei, C., Wang, W., Yang, W., Liu, J.: Deep retinex decomposition for low-light enhancement. arXiv preprint arXiv:1808.04560 (2018)
63. Wen, J., Zhu, Y., Li, J., Zhu, M., Wu, K., Xu, Z., Liu, N., Cheng, R., Shen, C., Peng, Y., Feng, F., Tang, J.: TinyVLA: Towards fast, data-efficient vision-language-action models for robotic manipulation. IEEE Robotics and Automation Letters (2025)
64. Ye, Y., Shi, H., Yang, K., Wang, Z., Yin, X., Sun, L., Wang, Y., Wang, K.: Towards anytime optical flow estimation with event cameras. Sensors (2025)

65. Zhong, Y., Bai, F., Cai, S., Huang, X., Chen, Z., Zhang, X., Wang, Y., Guo, S., Guan, T., Lui, K.N., Qi, Z., Liang, Y., Chen, Y., Yang, Y.: A survey on vision-language-action models: An action tokenization perspective. arXiv preprint arXiv:2507.01925 (2025)
66. Zihao Zhu, A., Yuan, L., Chaney, K., Daniilidis, K.: Unsupervised event-based optical flow using motion compensation. In: ECCVW (2018)
67. Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., Vuong, Q., Vanhoucke, V., Tran, H.T., Soricut, R., Singh, A., Singh, J., Sermanet, P., Sanketi, P.R., Salazar, G., Ryoo, M.S., Reymann, K., Rao, K., Pertsch, K., Mordatch, I., Michalewski, H., Lu, Y., Levine, S., Lee, L., Lee, T.E., Leal, I., Kuang, Y., Kalashnikov, D., Julian, R., Joshi, N.J., Irpan, A., Ichter, B., Hsu, J., Herzog, A., Hausman, K., Gopalakrishnan, K., Fu, C., Florence, P., Finn, C., Dubey, K.A., Driess, D., Ding, T., Choromanski, K.M., Chen, X., Chebotar, Y., Carbajal, J., Brown, N., Brohan, A., Arenas, M.G., Han, K.: RT-2: Vision-language-action models transfer web knowledge to robotic control. In: CoRL (2023)