EVENTBENCH: TOWARDS COMPREHENSIVE BENCHMARKING OF EVENT-BASED MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) have made significant advancements in event-based vision. However, the comprehensive evaluation of these models' capabilities within a unified benchmark remains a crucial yet largely unexplored aspect. In this paper, we introduce **EventBench**, which presents an evaluation benchmark covering 8 diverse task metrics and a large-scale event stream dataset. Our work distinguishes existing event-based benchmarks in four key aspects: **i) Openness in accessibility**, releasing all raw event streams and task instructions across 8 evaluation metrics; **ii) Diversity in task coverage**, covering diverse understanding, recognition, and spatial reasoning tasks, enabling comprehensive evaluation of model capability; **iii) Integration in spatial dimensions**, pioneering the design of 3D spatial reasoning task for event-based MLLMs; **iv) Scale in data volume**, with an accompanying training set of over one million event–text pairs supporting large-scale training and evaluation. With EventBench, we evaluate state-of-the-art closed-source models such as GPT-5 and Gemini-2.5 Pro, leading open-source models including Qwen2.5-VL and InternVL3, as well as event-based MLLMs like EventGPT that directly process raw event inputs. Extensive evaluation reveals that current event-based MLLMs perform well in event stream understanding, yet they still struggle with fine-grained recognition and spatial reasoning.

1 INTRODUCTION

Event cameras, offering high temporal resolution and high dynamic range, have been widely applied in challenging scenarios Gallego et al. (2020); Wang et al. (2025c); Zhang et al. (2024a); Wang et al. (2025d); Zhou et al. (2024); Yang et al. (2024); Liang et al. (2025); Klenk et al. (2024); Chen et al. (2025a); Verma et al. (2024); Chanda et al. (2025); Gehrig et al. (2020). With the rapid development of multimodal large language models (MLLMs) Bai et al. (2025); Zhang et al. (2024b); Team et al. (2025); Li et al. (2023) in recent years, event-based MLLMs Liu et al. (2025); Li et al. (2025b); Zhou & Lee (2025); Chen et al. (2025b); Qin et al. (2025); Zheng et al. (2023); Li et al. (2025a); Lu et al. (2025); Zhang et al. (2025) have increasingly demonstrated strong capabilities and generality in understanding asynchronous event streams. While video-based MLLMs have been extensively benchmarked across diverse tasks (e.g., understanding, recognition, and spatial reasoning) Fu et al. (2025); Li et al. (2024); Wang et al. (2025a); Zhou et al. (2025); Tan et al. (2025); Hong et al. (2025), comprehensive evaluation benchmarks for event-based MLLMs remain largely unexplored.

Most existing event-based MLLMs achieve strong performance on their respective evaluation tasks, lacking a publicly available and unified benchmark to comprehensively assess their general capabilities across diverse dimensions. Pioneering works such as EventGPT Liu et al. (2025) and EventVL Li et al. (2025b) present instructional datasets for scene understanding and visual question answering. Subsequent efforts, such as LLaFEA Zhou & Lee (2025) and LET-US Chen et al. (2025b), expand the scope to incorporate spatiotemporal localization and long-term event stream understanding. While these event-based MLLMs primarily focus on evaluating specific aspects of model capability, their assessment dimensions remain relatively limited compared with the mainstream benchmarks for video-based MLLMs (e.g., understanding, recognition, and spatial reasoning). In other words, this gap highlights the need for a unified benchmark to comprehensively evaluate event-based MLLMs across diverse dimensions.

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.

Table 1: Comparison of event-based MLLM datasets and benchmarks. We compare datasets in terms of scale, modalities, evaluation metrics, and four key attributes: Δd = Downstream task, Δs = Spatial task, Δt = Training data release, Δb = Benchmark release. Time ranges are categorized as *short* (50 ms), *medium* (50 ms–10 s), and *long* (>10 s). (RGB, Event Stream, and Text)

Dataset	Venue	Size	Modalities	Metric Types	Data Collection Source	Temporal Span	Attributes			
							Δd	Δs	Δt	Δb
N-ImageNet Kim et al. (2021)	ICCV'21	1M		–	Synthesis	Short	✗	✗	✓	✓
Talk2Event Kong et al. (2025)	NeurIPS'25	30.7K		–	Real World	Short	✓	✗	✗	✗
EventChatLiu et al. (2025)	CVPR'25	1.1M		3	Synthesis & Real World	Short	✗	✗	✓	✗
EventVL-Bench Li et al. (2025b)	arXiv'25	1.4M		4	Synthesis & Real World	Short & Medium	✓	✗	✗	✗
LLaFEA-Bench Zhou & Lee (2025)	ICCV'25	–		3	Synthesis & Real World	Short & Medium	✓	✗	✗	✗
EVQA-Bench Chen et al. (2025b)	arXiv'25	1.0M		4	Synthesis & Real World	Short & Medium & Long	✗	✗	✗	✗
EventBench	**Ours***	**1.4M**		**8**	**Synthesis & Real World**	**Short & Medium & Long**	✓	✓	✓	✓

Moreover, the rather limited public accessibility of existing event-based benchmarks Kong et al. (2025); Zhou et al. (2024); Yang et al. (2025); Zhu et al. (2018); De Tournemire et al. (2020); Perot et al. (2020); Orchard et al. (2015); Hu et al. (2020) further restricts fair comparison, making it challenging to accurately gauge progress in event-based MLLMs Lu (2026); Xu et al. (2025); Lou et al. (2025); Zhang et al. (2026).

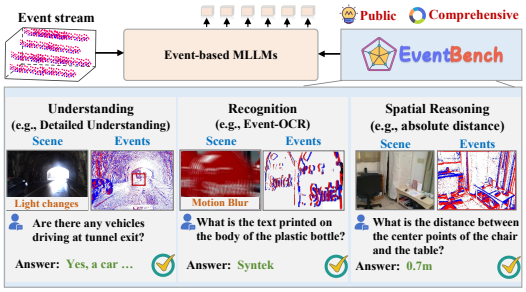


Figure 1: EventBench is an accessible and comprehensive evaluation benchmark for event-based MLLMs. It offers diverse task metrics across multiple dimensions (i.e., understanding, recognition, and spatial reasoning) and will accelerate research on event-based MLLMs in challenging scenarios.

In this work, we introduce EventBench, a comprehensive evaluation benchmark covering 8 diverse task metrics and a large-scale event stream dataset. To achieve our objective, two main challenges require to be addressed: (i) How can we design a systematic and comprehensive evaluation benchmark that measures capabilities across multiple dimensions, such as understanding, recognition, and spatial reasoning? How can we employ various MLLMs to validate the proposed benchmark's effectiveness across different tasks, ensuring it serves to foster progress in this topic?

To address these challenges, we introduce a comprehensive benchmark that systematically evaluates event-based MLLMs across diverse dimensions (see Fig. 1). It comprises 8 representative tasks covering understanding, recognition, and spatial reasoning. It integrates over 20 heterogeneous data sources drawn from both synthetic and real-world event streams to ensure rich scene diversity. Importantly, it pioneers the introduction of spatial reasoning evaluation for event-based MLLMs, filling a gap in assessing models' 3D understanding capabilities. Furthermore, we present a fully open and publicly accessible evaluation pipeline that enables different models to be evaluated under an unified benchmark, providing the opportunity to explore this evolving research topic. In addition, we construct a large-scale and carefully curated instruction dataset (i.e., EQA-1.4M) to enhance the baseline performance of open-source event-based MLLMs. We also establish objective metrics (e.g., multiple-choice accuracy, text matching, and numerical evaluation) to ensure fairness and reproducibility.

To evaluate the effectiveness of EventBench across diverse tasks, we assess both state-of-the-art closed-source models (i.e., GPT-5 OpenAI (2025), GPT-4o OpenAI (2024), Gemini 2.5-Pro Comanici et al. (2025), and Gemini 2.5-Flash Comanici et al. (2025)) as well as leading open-source MLLMs (i.e., Qwen2.5-VL Bai et al. (2025), InternVL 3.5 Wang et al. (2025b), MiMo-VL Team et al. (2025), and MiniCPM4-V Team (2025)). We further benchmark open-source event-based MLLMs (i.e., EventGPT Liu et al. (2025)). The results demonstrate that EventGPT+, trained with our large-scale instruction dataset EQA-1.4M, achieves the best overall performance. Moreover, fine-tuning open-source MLLMs on EQA-1.4M leads to substantial performance gains, highlighting the dataset's effectiveness in enhancing model performance.

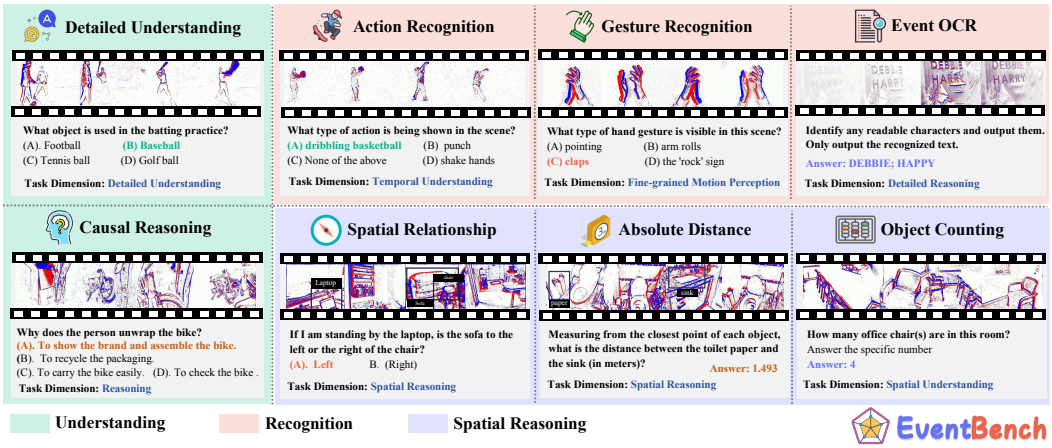


Figure 2: The comprehensive EventBench covers 8 diverse task metrics for systematically evaluating the capabilities of event-based MLLMs. These metrics can be broadly categorized into three types: understanding (i.e., detailed understanding), recognition (i.e., action recognition, gesture recognition, and event OCR), and spatial reasoning (i.e., spatial relationship, absolute distance, and object counting).

Our main contributions can be summarized as follows:

- We present *EventBench*, a publicly accessible evaluation benchmark with a large-scale event stream dataset, which will accelerate research on event-based MLLMs.
- We design *a comprehensive benchmark* covering 8 diverse task metrics for understanding, recognition, and spatial reasoning, which significantly expands the range of evaluation tasks compared to existing benchmarks.
- We build EQA-1.4M, *a large-scale dataset* of over one million synthetic and real-world event streams for event-based MLLMs, considering scene diversity across 20 data sources and various lengths of event streams.
- We conduct *extensive experiments* to compare state-of-the-art closed-source MLLMs, open-source MLLMs, and event-based MLLMs. The experimental results provide a guiding benchmark for event-based MLLMs.

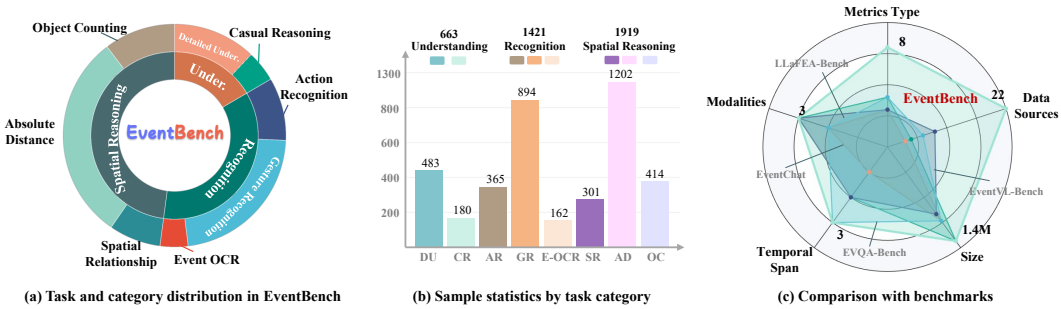
2 EventBench

2.1 Data Curation

EventBench are constructed through three stages: event stream collection, quality assurance, and instruction generation.

Event Stream Collection. To ensure scene diversity, we collect over 20 data sources covering both real-world and synthetic settings, including event datasets (e.g., DailyDVS-200 Wang et al. (2024a), HARDVS Wang et al. (2024b), DSEC Gehrig et al. (2021)) and large-scale video datasets (e.g., Kinetics-700 Carreira et al. (2019), ActivityNet Caba Heilbron et al. (2015), ScanNet Dai et al. (2017)). All videos are converted to event data using V2E Hu et al. (2021), resulting in diverse indoor and outdoor scenarios.

Quality Assurance. We adopt a two-stage filtering strategy. First, GPT-4o is used to retain videos with sufficient motion dynamics. Second, we manually annotate 10k samples and train a Qwen2.5-VL-7B Bai et al. (2025) model for large-scale quality filtering. This process produces 112k high-quality event streams. Most training and test samples are drawn from different data sources, and strict sample-level separation is enforced across all splits to prevent evaluation leakage. Given the scarcity of event-based spatial reasoning data, we further apply scene-level separation to the spatial subset, ensuring that no test scene appears during training.



(a) Task and category distribution in EventBench (b) Sample statistics by task category (c) Comparison with benchmarks

Figure 3: Data statistics of our EventBench. (a) Task and category distribution across three groups: understanding (i.e., DU and CR), recognition (i.e., AR, GR, and E-OCR), and spatial reasoning (i.e., SR, AD, and OC). (b) Sample statistics for each task category. (c) Comparison with existing event-based benchmarks across multiple dimensions (i.e., modality, metric type, data source, size, and temporal span). Note that EventBench provides a comprehensive benchmark for systematically evaluating.

Instruction Generation. We construct a unified QA-based pipeline covering understanding, recognition, and spatial reasoning. For understanding tasks, GPT-5 generates MCQA pairs involving temporal and causal reasoning, followed by human verification. For recognition tasks, structured annotations are reformulated into a unified QA format.

For spatial reasoning, we introduce three tasks—object counting, distance estimation, and spatial relation reasoning—based on instance-level annotations from ScanNet-V2 Dai et al. (2017) and ARKitScenes Baruch et al. (2021). These datasets provide metric 3D layouts for physically grounded evaluation. The construction follows three principles: (i) *geometry-consistent*, using metric 3D coordinates; (ii) *instance-level*, grounding each query to identifiable objects; and (iii) *human-centric*, expressing spatial relations in an egocentric reference frame.

Data statistics of our EventBench. (a) Task and category distribution across three groups: understanding (i.e., DU and CR), recognition (i.e., AR, GR, and E-OCR), and spatial reasoning (i.e., SR, AD, and OC). (b) Sample statistics for each task category. (c) Comparison with existing event-based benchmarks across multiple dimensions (i.e., modality, metric type, data source, size, and temporal span). Note that EventBench provides a comprehensive benchmark for systematically evaluating.

2.2 Dataset Statistics

EventBench comprises 8 task metrics for evaluating event-based MLLMs (see Fig. 2), including 112k event streams and 4,001 QA pairs. Each sample consists of an event stream, its paired video, and instructions. Compared to prior event-based benchmarks Chen et al. (2025b); Li et al. (2025b); Liu et al. (2025); Zhou & Lee (2025), which focus on limited evaluation scopes, EventBench provides a unified framework spanning understanding, recognition, and spatial reasoning. The overall composition is shown in Fig. 3. We summarize its key attributes from three perspectives.

Category Diversity. The benchmark is hierarchically organized into three major categories that reflect progressive levels of event stream understanding. The understanding category contains 663 QA pairs, which emphasize temporal comprehension and causal inference. The recognition category comprises 1,421 QA pairs focusing on the model's adaptability to real-world event-based vision tasks. The spatial reasoning category consists of 1,919 QA pairs designed to evaluate 3D geometric reasoning.

Temporal Duration. To capture diverse motion dynamics, EventBench includes both short and long-duration event streams, separated by a 10-second. The dataset consists of 22% long sequences and 78% short sequences, ensuring coverage across temporal scales.

Instruction Composition. The benchmark employs three types of objective question formats, including multiple choice, numerical, and text match. Each type contains a single correct answer to ensure reproducibility and precision in evaluation. Multiple-choice questions constitute 48.4% of

the dataset, numerical questions 40.0%, and text-match questions 11.6%. As depicted in Fig. 4, the multiple-choice options are evenly distributed across A, B, C, and D with proportions of 28.6%, 24.7%, 21.0%, and 25.7%, respectively. This balanced configuration minimizes bias and ensures that scores reflect genuine reasoning ability.

2.3 Evaluation Metrics

To comprehensively evaluate models across diverse capabilities, we adopt task-specific metrics tailored to each task:

Multiple-Choice Accuracy. For Detailed Understanding (DU), Causal Reasoning (CR), Action Recognition (AR), Gesture Recognition (GR), and Spatial Relationship (SR), we formulate evaluation as multiple-choice questions (MCQs) to ensure objective and standardized assessment. Each question is associated with a set of candidate answers containing plausible distractors, with a single correct option. For SR, the answer space is restricted to 2-3 predefined spatial relations. We report top-1 accuracy based on the selected option.

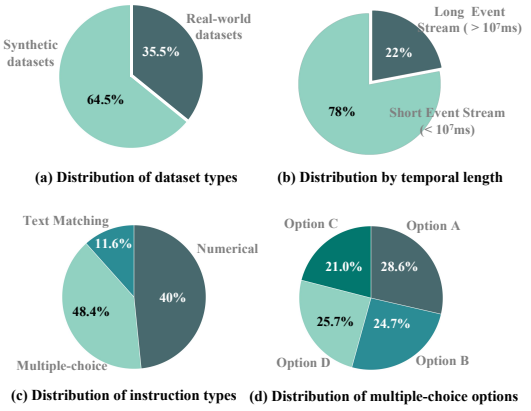


Figure 4: Data statistics in our EventBench across multiple dimensions. (a) Distribution of synthetic and real-world datasets. (b) Proportions of event streams by temporal length. (c)-(d) Distributions of instruction types and multiple-choice options.

Exact Matching. For tasks requiring precise and deterministic outputs, we adopt strict matching criteria. For Event OCR (E-OCR), we use exact string matching, where predictions must fully match the ground-truth text. For Object Counting (OC), we require exact numerical matching, where the predicted count must equal the ground-truth number of instances. This ensures objective evaluation of fine-grained recognition and numerical reasoning.

Threshold-based Accuracy. For Absolute Distance (AD), which involves continuous-valued predictions, we adopt a tolerance-based evaluation protocol. A prediction $\hat{y}$ is considered correct if the relative error with respect to the ground truth y satisfies $|\hat{y} - y|/y \leq 0.30$. We report the proportion of predictions that meet this criterion.

2.4 Comparison with Other Benchmarks

To highlight the advancements of the new benchmark, we compare it with related event-based datasets and benchmarks in Table 1. It indicates that our EventBench is the first open-source and comprehensive benchmark for event-based MLLMs. More precisely, EventBench incorporates 8 distinct metric types, effectively doubling the categorical coverage of the most competitive alternatives. Moreover, while existing benchmarks are often limited to short event sequences, EventBench offers a versatile suite of data with short, medium, and long-term event streams, facilitating a more thorough investigation of temporal modeling abilities.

Overall, the combination of a professionally designed instruction set and a large-scale event-based dataset makes EventBench a competitive benchmark with several key features: (i) Openness: all raw event streams and task instructions across eight evaluation metrics are released; (ii) Task diversity: it covers a wide range of understanding, recognition, and spatial reasoning tasks, enabling comprehensive evaluation; (iii) Spatial integration: it pioneers the inclusion of 3D spatial reasoning tasks for event-based MLLMs; and (iv) Data scale: it includes over one million event–text pairs, supporting large-scale training and evaluation.

Table 2: Evaluation results on EventBench, showing 3 categories of models: closed-source MLLMs, open-source MLLMs (SFT), and open-source event-based MLLMs. The 8 metrics are detailed understanding (DU), causal reasoning (CR), action recognition (AR), gesture recognition (GR), event-ocr (E-OCR), spatial relationship (SR), absolute distance (AD), and object counting (OC).

Model	Params	Backbone	Understanding		Recognition			Spatial Reasoning			Overall
			DU	CR	AR	GR	E-OCR	SR	AD	OC	
• Close-Source MLLMs											
GPT-5OpenAI (2025)	-	-	68.5	**70.1**	46.0	49.8	22.2	**47.8**	30.7	**25.6**	**45.1**
GPT-4oOpenAI (2024)	-	-	62.5	65.8	**55.3**	48.6	19.1	45.5	17.8	21.5	42.0
Gemini2.5-proComanici et al. (2025)	-	-	61.3	63.9	44.1	**49.9**	19.1	40.5	17.5	16.9	39.2
Gemini2.5-flashComanici et al. (2025)	-	-	54.7	61.1	42.7	43.4	21.0	25.6	15.6	14.5	34.8
Doubao-Seed-1.6-vision	-	-	**70.3**	65.2	46.6	45.4	**45.1**	11.6	**36.3**	23.0	42.9
• Open-Source MLLMs (SFT)											
Qwen2.5-VLBai et al. (2025)	3B	Qwen2.5	59.8	65.6	52.9	35.2	**45.6**	**44.9**	31.9	59.9	49.5
InternVL3.5Wang et al. (2025b)	4B	Qwen3	73.1	**75.0**	**69.3**	**45.2**	38.9	40.5	**46.8**	**61.1**	**56.2**
MiniCPM-V-4Team et al. (2025)	4.1B	MiniCPM4	**72.9**	63.9	58.9	37.1	44.4	44.2	43.1	55.6	52.5
MiMo-VLTeam (2025)	7B	MiMo	71.8	68.3	57.3	38.6	31.5	44.5	45.2	61.0	52.3
• Open-Source Event-Based MLLMs											
EventGPTLiu et al. (2025)	7B	Vicuna-v1.5	76.8	79.2	78.6	57.2	52.4	49.8	52.6	61.5	63.5
EventGPT+(B) (Ours*)	2B	Qwen2.5	78.3	78.2	79.8	55.5	53.6	50.2	51.9	62.8	63.8
EventGPT+(L) (Ours*)	7B	Qwen2.5	**79.1**	**79.6**	**81.6**	**58.7**	**55.4**	**53.2**	**52.2**	**64.6**	**65.5**

3 EXPERIMENTS

3.1 IMPLEMENTATION DETAILS

To evaluate our EventBench, we select four closed-source models (i.e., GPT-5, GPT-4o, Gemini-2.5-Pro, Gemini-2.5-Flash, and Doubao-Seed-1.6-Vision) and several state-of-the-art open-source models (i.e., Qwen2.5-VL, InternVL-3.5, MiniCPM4, and MiMo-VL), and the only publicly event-based MLLM (i.e., EventGPT) and our baseline model EventGPT+. For closed-source models, evaluation is conducted on EventBench by representing event streams as event videos sampled at 1 FPS. To ensure a fair comparison, all open-source models are fine-tuned on our EQA-1.4M dataset using the same event video inputs at 1 FPS, while event-based models are fine-tuned on raw event streams. All experiments are performed on 8 NVIDIA A800 GPUs, with API-based evaluation for closed-source models and single-GPU inference for all open-source settings.

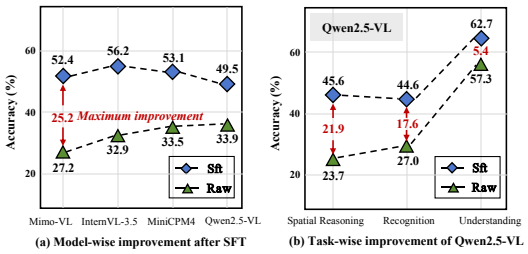


Figure 5: Overall performance improvement on EventBench after SFT. (a) Model-wise improvement with a 25.2% maximum gain. (b) Task-wise improvement of Qwen2.5-VL across spatial reasoning, downstream, and open-domain QA tasks.

3.2 EVALUATION ON EVENTBENCH

Quantitative Results. We conduct an extensive evaluation on EventBench to assess the generalization and reasoning capabilities of both video-based and event-based MLLMs. As summarized in Table 2, the evaluated models are grouped into three categories: advanced closed-source MLLMs, open-source video-based MLLMs fine-tuned on EQA-1.4M, and the only publicly available event-based MLLM (i.e., EventGPT). Meanwhile, our baseline EventGPT+ is designed to handle long event streams via dynamic event bin selection. In addition, Fig. 6 presents a radar-style comparison of model performance across EventBench, highlighting the distinctions among commercial models, open-source MLLMs, and event-based vision MLLMs. Specifically, GPT-5, Qwen2.5-VL, and EventGPT+ serve as representatives of these three categories, demonstrating the superior performance of EventGPT+(L) across multiple event understanding dimensions.

Among advanced closed-source models, GPT-5 achieves the highest overall accuracy, outperforming Gemini-2.5-Flash by 10.3%. While their strong general visual-language reasoning abilities, all closed-source models perform notably worse than the open-source video MLLMs after supervised fine-tuning on EQA-1.4M. This highlights the domain gap between standard video understanding and event stream perception. Within the open-source video-based category, InternVL-3.5 demonstrates the best performance, achieving an accuracy of 56.2%. Other models, such as Qwen2.5-VL and MiniCPM4, show consistent performance across tasks, particularly on general reasoning and action recognition. Fine-tuning on EQA-1.4M improves temporal reasoning and motion perception, indicating its effectiveness in bridging cross-modal gaps.

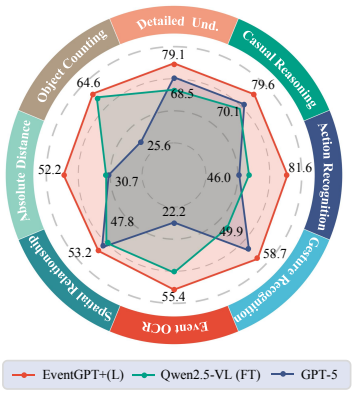


Figure 6: Comparison of EventGPT+(L), Qwen2.5-VL (FT), and GPT-5 on 8 evaluation tasks in EventBench.

Event-based MLLMs exhibit a significant advantage across nearly all categories, demonstrating clear superiority in spatiotemporal reasoning. Our EventGPT+ achieves the best overall performance among all models, surpassing GPT-5 by 20.4% and outperforming the best fine-tuned open-source video MLLM by 9.3%. This indicates that directly modeling raw event streams, allows the model to better capture high temporal resolution motion cues critical for event understanding. To further validate this advantage, we compare EventGPT+(L) with the only existing open-source event-based MLLMs (i.e., EventGPT). For a fair comparison, EventGPT is further fine-tuned on EQA-1.4M, with its fixed event bin constraint removed to support long duration event streams. Under identical conditions, EventGPT+(L) consistently outperforms EventGPT across all categories, achieving an overall improvement of 2.0%.

Qualitative Analysis.

We further conduct qualitative analyses on challenging scenarios from EventBench. As illustrated in Fig. 7, under severe motion blur, RGB-based models suffer from significant image degradation, which induces noticeable hallucinations across both advanced proprietary models such as GPT-5 and Gemini-2.5-Pro, and strong open-source models like Qwen2.5-VL. In contrast, event-based MLLMs exhibit higher robustness, effectively preserving fine-grained motion structures and temporal consistency in high-speed environments.

Table 3: Performance of open-source MLLMs before and after SFT on EQA-1.4M, evaluated across EventBench.

Task	Qwen-VL	SFT	Intern-VL	SFT
Understanding				
Detailed Understanding	53.8	59.8(+6.0)	47.2	**73.1**(+25.9)
Causal Reasoning	59.8	**65.6**(+5.8)	63.2	**75.0**(+11.8)
Recognition				
Action Recognition	29.3	**52.9**(+23.6)	24.1	**69.3**(+45.2)
Gesture Recognition	31.5	**35.2**(+3.7)	32.9	**45.2**(+12.3)
Event OCR	20.2	**45.6**(+25.4)	22.8	**38.9**(+16.1)
Spatial Reasoning				
Spatial Relationship	38.8	**44.9**(+6.1)	36.8	**40.5**(+3.7)
Absolute Distance	21.7	**31.9**(+10.2)	18.9	**46.8**(+27.9)
Object Counting	10.8	**59.9**(+49.1)	16.9	**61.1**(+44.2)

Moreover, we evaluate the Absolute Distance task in EventBench (see the Appendix). While most models reveal inherent limitations in spatial reasoning, EventGPT+ achieves the most accurate distance estimation with an error of 9.3%. Qwen2.5-VL also demonstrates a clear improvement after supervised fine-tuning, reducing its error from 33.3% to 12.9%.

Effect of Fine-tuning on EventBench. To ensure a fair comparison, some representative open-source video MLLMs (i.e., Qwen2.5-VL and InternVL-3.5) are fine-tuned on the EQA-1.4M dataset using event videos as input. Table 3 summarizes the performance of Qwen2.5-VL and InternVL-3.5 on EventBench before and after fine-tuning. Fine-tuning yields consistent gains across the three evaluation categories (i.e., understanding, recognition, and spatial reasoning) as visualized in Fig. 5. In particular, Qwen2.5-VL attains improvements of 49.1%, 10.2%, and 6.1% for three spatial reason-

ing tasks (i.e., object counting, absolute distance, and spatial relationship), respectively, meanwhile InternVL-3.5 achieves gains of 44.2%, 27.9%, and 3.7% on the same tasks. Although fine-tuning markedly enhances the temporal and spatial reasoning abilities of video-based MLLMs, models that directly operate on raw event streams still deliver superior overall performance on EventBench.

Effect of Sampling Interval. To assess the influence of the sampling interval on the final performance, we evaluate four settings with $\Delta t \in \{0.5, 1.0, 1.5, 2.0\}$. As shown in Table 4, varying Δt yields only minor fluctuations (within 2.5%) across all models, indicating strong robustness to temporal sampling and confirming the stability of EventBench. Given this consistent behavior and considering the balance between accuracy and computational efficiency, we recommend adopting $\Delta t = 1.0$ as the default configuration for evaluating event-based MLLMs using EventBench.

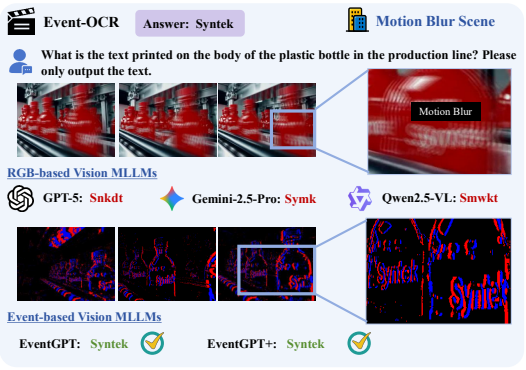


Figure 7: Comparison of frame-based and event-based MLLMs on Event-OCR under motion blur. Event-based models (i.e., EventGPT and EventGPT+) successfully recognize "Syntek", while RGB-based models fail due to motion blur.

Effect of Aggregation Window Size. To further examine the impact of temporal granularity, we analyze the effect of the aggregation window size on EventBench. As shown in Table 5, we evaluate four configurations with window sizes of $10\,ms$, $20\,ms$, $30\,ms$, and $40\,ms$. The results show that a $20\,ms$ window achieves the best overall performance. When the window size is too small, the number of events within each bin becomes insufficient, resulting in underpopulated or even empty windows that weaken temporal feature integration. In contrast, overly large windows produce coarser temporal segmentation, blurring fine-grained spatiotemporal cues and ultimately degrading the performance on tasks sensitive to temporal dynamics.

3.3 Experimental Analysis

We conduct further analysis to reveal the key factors influencing the performance of event-based MLLMs: *(i) model strengths and limitations*, *(ii) challenges in spatial reasoning*, and *(iii) the impact of SFT on the modality gap.*

Table 4: Performance comparison of open-source and event-based multimodal large language models (MLLMs) under different sampling intervals Δt to assess the impact of temporal granularity.

Model	Venue	Params	Sampling Interval (Δt)			
			$\Delta t = 0.5$	$\Delta t = 1.0$	$\Delta t = 1.5$	$\Delta t = 2.0$
Open-sources MLLMs						
Qwen2.5-VL	arXiv'25	3B	**49.7**	49.5	48.9	48.7
InternVL-3.5	arXiv'25	4B	**56.5**	56.2	55.9	55.6
Event-based vision MLLMs						
EventGPT	CVPR'25	7B	62.6	**63.5**	62.4	62.2
EventGPT+(L)	Ours	7B	64.9	**65.5**	64.3	63.9

How Do Event-based MLLMs Perform Across Tasks? In Tab. 2, we report the performance of advanced closed-source MLLMs, open source MLLMs with SFT, and event-based MLLMs across all tasks. From these results, we identify two principal bottlenecks. (1) *Spatial reasoning is the main limitation in event-based MLLMs*, where even the strongest closed-source MLLMs (e.g., GPT-5), open source MLLMs (e.g., InternVL-3.5), and advanced event-based MLLMs (e.g., EventGPT+(L)) still perform poorly. (2) *Spatial instruction tuning significantly improves spatial reasoning yet remains limited.* As shown in Tab. 3, spatial instruction tuning brings significant gains for open source MLLMs on spatial reasoning tasks (i.e., spatial relationship, absolute distance, and object counting), especially on object counting, suggesting that SFT helps reduce the cross-modal discrepancy between event streams and RGB inputs and enhances the spatial reasoning ability of MLLMs.

What Are the Principal Performance Bottlenecks? In Tab. 2, we report the performance of advanced closed-source MLLMs, open source MLLMs with supervised fine tuning (SFT), and event-based MLLMs across all tasks. From these results, we identify two principal bottlenecks. (1)

Spatial reasoning is the main limitation in event-based MLLMs, where even the strongest closed-source MLLMs (e.g., GPT-5), open source MLLMs (e.g., InternVL-3.5), and advanced event-based MLLMs (e.g., EventGPT+(L)) still perform poorly. (2) *Spatial instruction tuning significantly improves spatial reasoning yet remains limited.* As shown in Tab. 3, spatial instruction tuning brings significant gains for open source MLLMs on spatial reasoning tasks (i.e., spatial relationship, absolute distance, and object counting), especially on object counting, suggesting that supervised fine tuning (SFT) helps reduce the cross-modal discrepancy between event streams and RGB inputs and enhances the spatial reasoning ability of MLLMs.

Can Event SFT Mitigate the Modality Gap? To examine the impact of supervised fine-tuning (SFT) on the performance of RGB-based MLLMs in event-based vision tasks (i.e., understanding, recognition, and spatial reasoning), we evaluate two representative open-source models (i.e., Qwen2.5-VL and InternVL-3.5) both before and after SFT, and report the results in Tab. 3. Our analysis yields two key observations. (1) *SFT on event stream data contributes to reducing the modality gap.* As shown in Tab. 3, all task categories exhibit substantial performance gains after SFT, suggesting that adapting RGB-based MLLMs to event representations can effectively narrow the discrepancy between the two modalities. (2) *SFT alone is insufficient to close the gap relative to natively trained event-based MLLMs.* As shown in Tab. 2, although the fine-tuned open-source MLLMs achieve notable improvements, they continue to fall short of event-based MLLMs trained directly on event streams. This demonstrates that the modality disparity between event stream and RGB videos remains pronounced, and that fine-tuning with million-scale event samples can mitigate the gap, though considerable disparity remains.

Table 5: Performance comparison of open-source and event-based MLLMs under different temporal aggregation window sizes to assess the effect of event accumulation granularity.

Model	Venue	Params	Aggregation Window Size			
			10 ms	20 ms	30 ms	40 ms
Open-source MLLMs						
Qwen2.5-VL	arXiv'25	3B	49.2	**49.5**	48.6	48.2
InternVL-3.5	arXiv'25	4B	55.6	**56.2**	54.9	54.3
Event-based Vision MLLMs						
EventGPT	CVPR'25	7B	62.7	**63.5**	62.3	62.1
EventGPT+(L)	Ours	7B	64.8	**65.5**	64.5	63.9

3.4 Discussion

Our results yield two key insights. First, event-based MLLMs excel at general understanding—EventGPT+(L) averages 79.4% across the two understanding tasks—but remain weak on specialized recognition and spatial reasoning, indicating that LLM knowledge alone lacks the required temporal and geometric cues. Second, million-scale event SFT consistently improves RGB-based MLLMs (Tab. 3), narrowing but not closing the gap with raw-event models. Detailed analyses are provided in the supplementary material.

Beyond these empirical findings, EventBench lays the foundation for a fully open ecosystem for event-based MLLMs by releasing the training dataset, evaluation benchmark, baseline models, and inference and evaluation code. We will further build and maintain a leaderboard to support continuous and reproducible model evaluation, with the goal of providing foundational infrastructure for tracking and accelerating progress in event understanding.

4 Conclusion

In this paper, we introduced EventBench, a fully open and comprehensive benchmark for systematically evaluating the capabilities of event-based MLLMs. EventBench stands as a competitive benchmark compared with existing event-based benchmarks, offering several key advantages: openness, broad task diversity, integration of spatial reasoning tasks, and a large data scale. Extensive evaluations were conducted on state-of-the-art proprietary models, open-source RGB-based MLLMs, and event-based MLLMs. The results reveal that current event-based MLLMs perform well in event stream understanding, yet still struggle with fine-grained recognition and spatial reasoning. We believe that EventBench will help drive further progress in event-based MLLMs by providing a unified benchmark for fair comparison and consistent evaluation.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv*, 2025.

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv*, 2021.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 961–970, 2015.

Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv*, 2019.

Kaustav Chanda, Aayush Verma, Arpitsinh Vaghela, Yezhou Yang, and Bharatesh Chakravarthi. Event quality score (eqs): Assessing the realism of simulated event camera streams via distance in latent space. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5105–5113, 2025.

Kanghao Chen, Guoqiang Liang, Yunfan Lu, Hangyu Li, and Lin Wang. 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*, 2025a.

Rui Chen, Xingyu Chen, Shaoan Wang, Shihan Kong, and Junzhi Yu. Let-us: Long event-text understanding of scenes. *arXiv preprint arXiv:2508.07401*, 2025b.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv*, 2025.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5828–5839, 2017.

Pierre De Tournemire, Davide Nitti, Etienne Perot, Davide Migliore, and Amos Sironi. A large scale event-based detection dataset for automotive. *arXiv*, 2020.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24108–24118, 2025.

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

Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Recycling video datasets for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3586–3595, 2020.

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

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8450–8460, 2025.

Yuhuang Hu, Jonathan Binas, Daniel Neil, Shih-Chii Liu, and Tobi Delbruck. Ddd20 end-to-end event camera driving dataset: Fusing frames and events with deep learning for improved steering prediction. In *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*, pp. 1–6, 2020.

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.

Junho Kim, Jaehyeok Bae, Gangin Park, Dongsu Zhang, and Young Min Kim. N-imagenet: Towards robust, fine-grained object recognition with event cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2146–2156, 2021.

Simon Klenk, David Bonello, Lukas Koestler, Nikita Araslanov, and Daniel Cremers. Masked event modeling: Self-supervised pretraining for event cameras. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2378–2388, 2024.

Lingdong Kong, Dongyue Lu, Ao Liang, Rong Li, Yuhao Dong, Tianshuai Hu, Lai Xing Ng, Wei Tsang Ooi, and Benoit R Cottereau. Talk2event: Grounded understanding of dynamic scenes from event cameras. *arXiv*, 2025.

Dong Li, Jiandong Jin, Yuhao Zhang, Yanlin Zhong, Yaoyang Wu, Lan Chen, Xiao Wang, and Bin Luo. Semantic-aware frame-event fusion based pattern recognition via large vision–language models. *Pattern Recognition*, 158:111080, 2025a.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv*, 2023.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024.

Pengteng Li, Yunfan Lu, Pinghao Song, Wuyang Li, Huizai Yao, and Hui Xiong. Eventvl: Understand event streams via multimodal large language model. *arXiv*, 2025b.

Quanmin Liang, Qiang Li, Shuai Liu, Xinzi Cao, Jinyi Lu, Feidiao Yang, Wei Zhang, Kai Huang, and Yonghong Tian. Efficient event camera data pretraining with adaptive prompt fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8656–8667, 2025.

Shaoyu Liu, Jianing Li, Guanghui Zhao, Yunjian Zhang, Xin Meng, Fei Richard Yu, Xiangyang Ji, and Ming Li. Eventgpt: Event stream understanding with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29139–29149, 2025.

Hanyue Lou, Jiayi Zhou, Yang Zhang, Boyu Li, Yi Wang, Guangnan Ye, and Boxin Shi. Reconstruction as a bridge for event-based visual question answering. *arXiv preprint arXiv:2512.11510*, 2025.

Dongyue Lu. Towards robust and interpretable event–frame fusion for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 41062–41063, 2026.

Zihan Lu, Cuiying Sun, and Xiang Li. Multimodal large language model-enabled machine intelligent fault diagnosis method with non-contact dynamic vision data. *Sensors*, 25(18):5898, 2025.

OpenAI. Gpt-4o system card. `https://openai.com/index/gpt-4o-system-card/`, 2024.

OpenAI. Introducing gpt-5. `https://openai.com/index/introducing-gpt-5/`, 2025.

Garrick Orchard, Ajinkya Jayawant, Gregory K Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in Neuroscience*, 9:437, 2015.

Etienne Perot, Pierre De Tournemire, Davide Nitti, Jonathan Masci, and Amos Sironi. Learning to detect objects with a 1 megapixel event camera. *Proceedings of the Advances in Neural Information Processing Systems*, 33:16639–16652, 2020.

Haotong Qin, Cheng Hu, and Michele Magno. Event-priori-based vision-language model for efficient visual understanding. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 16–30, 2025.

Xichen Tan, Yuanjing Luo, Yunfan Ye, Fang Liu, and Zhiping Cai. Allvb: All-in-one long video understanding benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7211–7219, 2025.

MiniCPM Team, Chaojun Xiao, Yuxuan Li, Xu Han, Yuzhuo Bai, Jie Cai, Haotian Chen, Wentong Chen, Xin Cong, Ganqu Cui, et al. Minicpm4: Ultra-efficient llms on end devices. *arXiv*, 2025.

Xiaomi MiMo Team. Mimo-vl technical report. *arXiv*, 2025. URL https://arxiv.org/abs/2506.03569.

Aayush Atul Verma, Bharatesh Chakravarthi, Arpitsinh Vaghela, Hua Wei, and Yezhou Yang. Etram: Event-based traffic monitoring dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22637–22646, 2024.

Qi Wang, Zhou Xu, Yuming Lin, Jingtao Ye, Hongsheng Li, Guangming Zhu, Syed Afaq Ali Shah, Mohammed Bennamoun, and Liang Zhang. Dailydvs-200: A comprehensive benchmark dataset for event-based action recognition. In *Proceedings of the European Conference on Computer Vision*, pp. 55–72, 2024a.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22958–22967, 2025a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv*, 2025b.

Xiao Wang, Zongzhen Wu, Bo Jiang, Zhimin Bao, Lin Zhu, Guoqi Li, Yaowei Wang, and Yonghong Tian. Hardvs: Revisiting human activity recognition with dynamic vision sensors. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 5615–5623, 2024b.

Xiao Wang, Jingtao Jiang, Qiang Chen, Lan Chen, Lin Zhu, Yaowei Wang, Yonghong Tian, and Jin Tang. Estr-cot: Towards explainable and accurate event stream based scene text recognition with chain-of-thought reasoning. *arXiv*, 2025c.

Xiao Wang, Jingtao Jiang, Dong Li, Futian Wang, Lin Zhu, Yaowei Wang, Yongyong Tian, and Jin Tang. Eventstr: A benchmark dataset and baselines for event stream based scene text recognition. *arXiv*, 2025d.

Wenhao Xu, Xin Dong, Yue Li, Haoyuan Shi, and Zhiwei Xiong. Eventstu: Event-guided efficient spatio-temporal understanding for video large language models. *arXiv preprint arXiv:2511.18920*, 2025.

Yan Yang, Liyuan Pan, and Liu Liu. Event camera data dense pre-training. In *Proceedings of the European Conference on Computer Vision*, pp. 292–310, 2024.

Yan Yang, Liyuan Pan, Dongxu Li, and Liu Liu. Ezsr: Event-based zero-shot recognition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4628–4638, 2025.

Baoheng Zhang, Jiahui Liu, Gui Zhao, Weizhou Zhang, Yixuan Ma, Jun Jiang, Yingxian Chen, Wilton WT Fok, Xiaojuan Qi, and Hayden Kwok-Hay So. Learning to see through illumination extremes with event streaming in multimodal large language models. *arXiv preprint arXiv:2603.27558*, 2026.

Jinchang Zhang, Zijun Li, Jiakai Lin, and Guoyu Lu. Adaptive event stream slicing for open-vocabulary event-based object detection via vision-language knowledge distillation. *arXiv*, 2025.

Pengyu Zhang, Hao Yin, Zeren Wang, Wenyue Chen, Shengming Li, Dong Wang, Huchuan Lu, and Xu Jia. Evsign: Sign language recognition and translation with streaming events. In *Proceedings of the European Conference on Computer Vision*, pp. 335–351, 2024a.

Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model. `https://llava-vl.github.io/blog/2024-04-30-llava-next-video/`, 2024b.

Xu Zheng, Yexin Liu, Yunfan Lu, Tongyan Hua, Tianbo Pan, Weiming Zhang, Dacheng Tao, and Lin Wang. Deep learning for event-based vision: A comprehensive survey and benchmarks. *arXiv*, 2023.

Hanyu Zhou and Gim Hee Lee. Llafea: Frame-event complementary fusion for fine-grained spatiotemporal understanding in lmms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Jiazhou Zhou, Xu Zheng, Yuanhuiyi Lyu, and Lin Wang. Exact: Language-guided conceptual reasoning and uncertainty estimation for event-based action recognition and more. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18633–18643, 2024.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition*, pp. 13691–13701, 2025.

Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics and Automation Letters*, 3(3):2032–2039, 2018.