# EventBench: Towards Comprehensive Benchmarking of Event-based MLLMs

Shaoyu Liu[1,2], Jianing Li[2], Guanghui Zhao[1], Yunjian Zhang[2], Xiangyang Ji[2]

[1]Xidian University [2]Tsinghua University

**Project Page & Code**: eventbench.github.io **Dataset:** eventbench.huggingface

## 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 [9, 11, 20, 21, 23, 29, 35, 44, 47, 53–55, 60, 61, 63, 64, 67, 69, 74]. With the rapid development of multimodal large language models (MLLMs) [3, 32, 43, 51, 70] in recent years, event-based MLLMs [12, 27, 31, 34, 37, 38, 46, 68, 71, 72] 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) [19, 24, 33, 50, 58, 75], comprehensive evaluation benchmarks for event-based MLLMs remain largely unexplored.

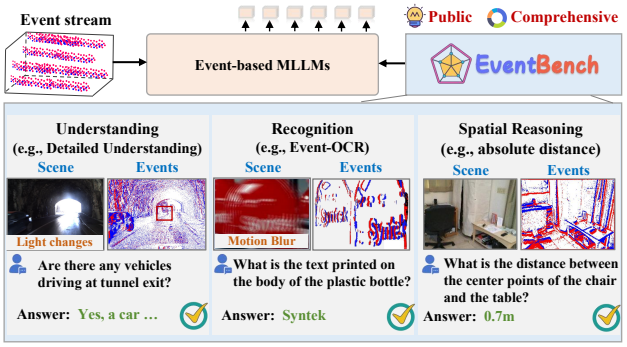



Figure 1. Our EventBench is a publicly accessible and comprehensive evaluation benchmark for event-based MLLMs. It offers diverse task metrics across multiple dimensions (i.e., understanding, recognition, and spatial reasoning) and will accelerate research on event-based MLLMs in challenging scenarios.

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 [37] and EventVL [34] present instructional datasets for scene understanding and visual question answering. Subsequent efforts, such as LLaFEA [72] and LET-US [12], 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,

Table 1. Comparison of event-based MLLM datasets and benchmarks. We compare datasets in terms of scale, modalities, evaluation metrics, and four key attributes: $\Delta$d = Downstream task support, $\Delta$s = Spatial task support, $\Delta$t = Training data release, $\Delta$b = 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 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | $\Delta$d | $\Delta$s | $\Delta$t | $\Delta$b |
| **N-ImageNet**[28] | ICCV'21 | 1M | | – | Synthesis | Short | ✗ | ✗ | ✓ | ✓ |
| **Talk2Event**[30] | NeurIPS'25 | 30.7K | | – | Real World | Short | ✓ | ✗ | ✗ | ✗ |
| **EventChat**[37] | CVPR'25 | 1.1M | | 3 | Synthesis & Real World | Short | ✗ | ✗ | ✓ | ✗ |
| **EventVL-Bench**[34] | arXiv'25 | 1.4M | | 4 | Synthesis & Real World | Short & Medium | ✓ | ✗ | ✗ | ✗ |
| **LLaFEA-Bench**[72] | ICCV'25 | – | | 3 | Synthesis & Real World | Short & Medium | ✓ | ✗ | ✗ | ✗ |
| **EVQA-Bench**[12] | arXiv'25 | 1.0M | | 4 | Synthesis & Real World | Short & Medium & Long | ✗ | ✗ | ✗ | ✗ |
| **EventBench** | **Ours*** | **1.4M** | | **8** | **Synthesis & Real World** | **Short & Medium & Long** | ✓ | ✓ | ✓ | ✓ |

and spatial reasoning). In other words, this gap highlights the need for a unified benchmark to comprehensively evaluate event-based MLLMs across diverse dimensions. Moreover, the limited public accessibility of existing event-based benchmarks [16, 25, 30, 42, 45, 65, 74, 76] further restricts fair comparison, making it challenging to accurately gauge progress in event-based MLLMs.

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 different tasks, we assess both state-of-the-art closed-source models (i.e., GPT-5 [41], GPT-4o [40], Gemini 2.5-Pro [14], and Gemini 2.5-Flash [14]) as well as leading open-source MLLMs (i.e., Qwen2.5-VL [3], InternVL 3.5 [57], MiMo-VL [51], and MiniCPM4-V [52]). We further benchmark open-source event-based MLLMs (i.e., EventGPT [37]). 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 event-based model performance.

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. Related Works

**Event-based MLLMs.** Early works (e.g., EventCLIP [62] and EventBind [73]) in event-language multimodal learning tightly coupled event streams with language [30, 65], achieving efficient event-language alignment and enabling downstream tasks such as event retrieval and object recognition. With the rapid advancement of LLMs [32, 36, 39, 66], event-based vision MLLMs have seen significant progress. Pioneering models, such as EventGPT [37] and EventVL [34], design event-based LLMs to leverage inher-

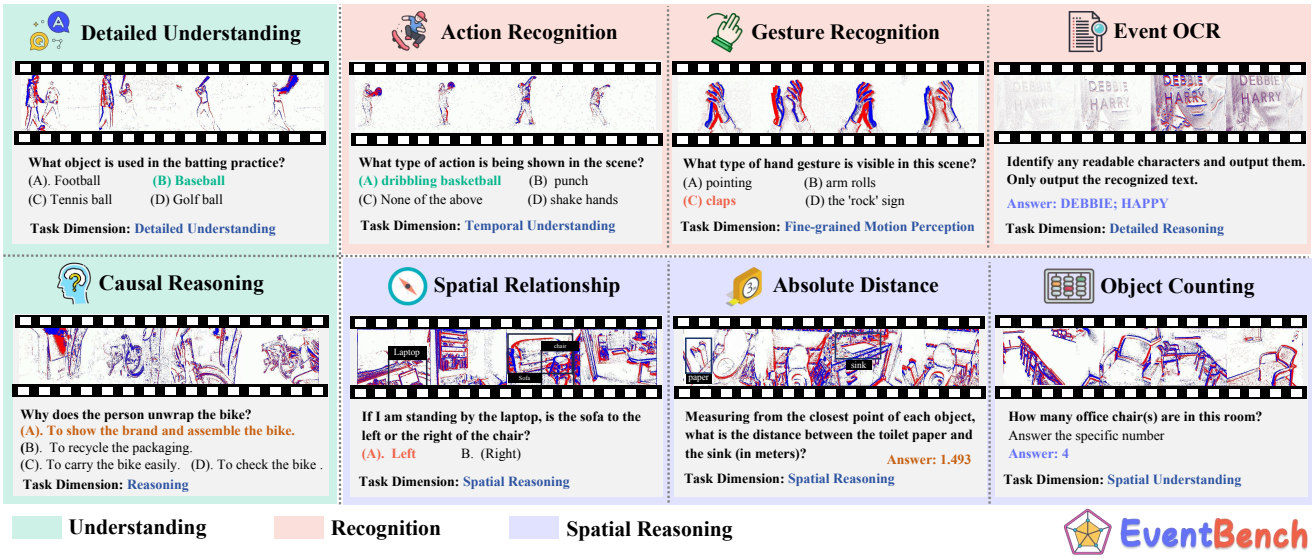



Figure 2. The comprehensive EventBench covers 8 diverse task metrics for systematically evaluating the capabilities of event-based MLLMs. These metrics can be broadly categorized into three types: understanding (i.e., detailed understanding), recognition (i.e., action recognition, gesture recognition,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

ent world knowledge for event stream understanding. In addition, LLaFEA [72] extends region-level perception using both events and frames, while LET-US [12] explores the capabilities of event-based vision MLLMs in modeling long-duration event streams. Nevertheless, these event-based MLLMs achieve strong performance on their respective evaluation tasks, yet they lack a unified benchmark to assess their general capabilities across diverse dimensions.

**Benchmarks for Event-based MLLMs.** Effective benchmarks are vital for advancing event-based MLLMs, as they provide guidance for evaluating model capabilities and fostering new research. For instance, EventChat [37] is the first self-constructed event-based dataset for event-based MLLMs, designed to assess models' performance in event-driven captioning and visual question answering. LLaFEA-Bench [72] focuses on region-level understanding by incorporating object localization, while EVQA-Bench [12] targets long-term event stream comprehension. As shown in Table 1, these benchmarks provide quantitative assessments of specific abilities but remain narrow in scope, typically focusing on a limited range of tasks. Furthermore, they are often not publicly accessible, limiting reproducible comparisons across models and research on this evolving topic.

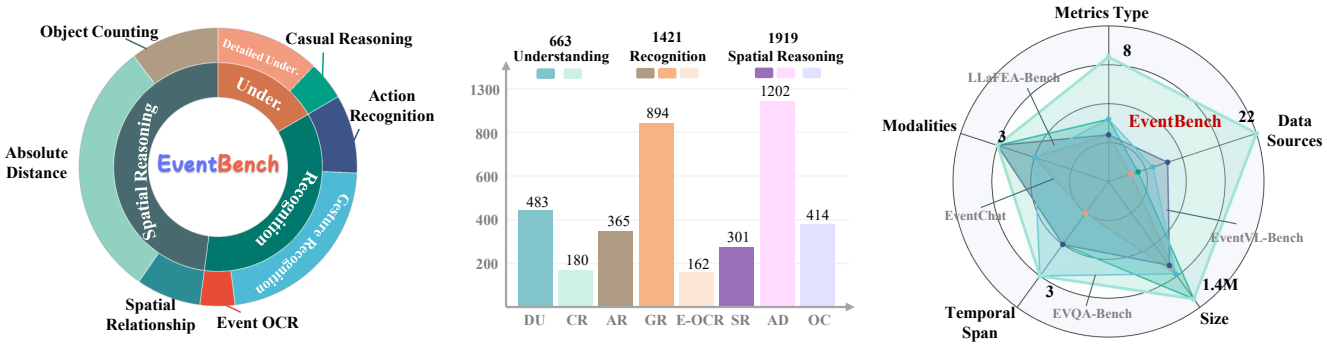
## 3. EventBench

### 3.1. Data Curation

The data preparation process of EventBench and EQA-1.4M consists of three main stages: event stream collection and synthesis, quality assurance, and instruction generation. Through this process, we aim to establish a comprehensive and high-quality evaluation platform specifically designed to assess the capabilities of event-based MLLMs. The detailed process is described as follows.

**Event Stream Collection and Synthesis.** To ensure scene diversity, we select over 20 data sources covering both real-world and synthetic event streams. These include widely used real-world event-based datasets such as DailyDVS-200 [56], HARDVS [59], Bullying-10k [17], EB-HandGesture [1], EHWGesture [2], EventSTR [61], DSEC [22], and N-ImageNet [28]. To further enhance scene diversity, we also incorporate multiple scene-centric video datasets featuring dynamic scenarios, including Kinetics-700 [8], ActivityNet [7], Charades [48], MotionBench [24], MotionSight [18], Wevid-10M [4], SportsSloMo [10], ScanNet [15], PE-Data [6], and PLM-Data [13]. This integration ensures a balanced coverage across indoor and outdoor scenarios, providing a rich foundation for evaluation. All video datasets are subsequently converted into event streams using the V2E [26] to build a large-scale synthesized dataset.

**Quality Assurance.** To ensure reliable and high-quality raw data sources, we implement a two-stage quality assurance strategy. In the first stage, we prioritize source videos exhibiting rich motion dynamics. For videos accompanied by captions, GPT-4o is employed to verify whether the descriptions correspond to dynamic scenes, and only those meeting this criterion are retained. In the second stage, we develop an automated filtering pipeline. A subset of 10k converted samples is manually annotated for quality assessment, and these annotations are used to fine-tune a Qwen2.5-VL-7B [3] evaluation model. The trained model




(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 the capabilities of event-based MLLMs.

| Category | Evaluation Dimension | Data Sources |
|---|---|---|
| **Understanding** | Detailed Understanding | Kinetics [8], SportSloMo [10], MotionBench [24] |
| | Causal Reasoning | PLM-Data [13], PE-Data [6], DSEC [22], N-ImageNet [28] ActivityNet [7], Charades [48], MotionSight [18], WebVid-10M [4] |
| **Recognition** | Action Recognition | DailyDVS-200 [56], Bullying-10k [17], UCF-101 [49], HARDVS [59] |
| | Gesture Recognition | EB-HandGesture [1], EHWGesture [2] |
| | Event-OCR | EventSTR [60] |
| **Spatial Reasoning** | Spatial Relationship | ScanNet-v2 [15], ARKitScenes [5] |
| | Absolute Distance | ScanNet-v2, ARKitScenes |
| | Object Counting | ScanNet-v2, ARKitScenes |

Table 2. Evaluation dimensions and data sources in EventBench. EventBench includes 8 representative tasks covering understanding, recognition, and spatial reasoning. It brings together over 20 data sources from both synthetic and real-world event streams.

is subsequently applied to the entire dataset to automatically remove low quality samples. This process yields a high quality corpus of 112k event streams that span both task-specific and general dynamic scenarios.

**Instruction Generation.** We construct a unified instruction generation pipeline that formulates diverse tasks under a consistent QA paradigm. The 8 tasks in EventBench are organized into three categories: understanding, recognition, and spatial reasoning, each reflecting different aspects of event-based MLLM capabilities.

For understanding tasks (i.e., detailed understanding and causal reasoning), GPT-5 is prompted to synthesize multiple choice question–answer (MCQA) pairs based on the RGB videos paired with event streams. A reasoning-oriented prompting strategy guides the model to generate explicit temporal and causal inference chains, while human verification ensures factual accuracy and modality alignment. For recognition tasks (i.e., action recognition, gesture recognition, and event OCR), structured annotations are reformulated into natural language QA form. Each instruction concisely defines the recognition objective and contextual scope, and the candidate labels are converted into multiple choice options, thereby unifying recognition and understanding within a shared evaluation format.

Beyond understanding and recognition, we further introduce spatial reasoning tasks to evaluate models' 3D geometric reasoning. We design three tasks (i.e., object counting, absolute distance estimation, and spatial relational reasoning), each grounded in the intrinsic geometry of the scene and constructed from instance-level annotations in ScanNet-V2 [15] and ARKitScene [5]. The dataset provides metrically accurate spatial layouts that ensure geometric consistency, enabling physically interpretable and quantitatively verifiable evaluation of spatial reasoning ability. To ensure both reliability and consistency of the constructed data, we adhere to three guiding principles: (i) *Geometry consistent*, where all spatial relationships are derived directly from metric 3D coordinates rather than image projections, ensuring geometric fidelity; (ii) *Instance level*, where each query is anchored to uniquely identified object instances, avoiding semantic ambiguity across categories; (iii) *Human centric*, where spatial relations are represented in egocentric form, aligning geometric definitions with intuitive human perception.

## 3.2. Dataset Statistics

EventBench encompasses 8 diverse task metrics for event-based MLLMs, comprising 112k event streams and 4,003 QA pairs. Each sample includes an event stream, its paired video, and a carefully designed textual instruction. In contrast to previous event-based benchmarks [12, 34, 37, 72] that define only limited evaluation scopes, EventBench establishes a broader and more systematic benchmark unifying understanding, recognition, and spatial reasoning. The overall task composition is illustrated in Fig. 3, which summarizes the hierarchical distribution across categories and individual task scales. We further analyze the attributes of the benchmark from the following 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 both short-term and long-term motion dynamics, EventBench integrates event streams with diverse temporal lengths divided by a threshold of 10 second. As shown in Fig. 4, long event streams account for 22% of all samples, while short ones make up the remaining 78%. This temporal distribution provides balanced and comprehensive coverage across motion scales, thereby enabling consistent evaluation of model robustness under both short-term and long-term event streams.

*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.

### 3.3. Task Definition

To establish a unified benchmark for systematically evaluating event-based vision MLLMs, we define a structured set of eight tasks covering understanding, recognition, and spatial reasoning, as detailed below.

**Detailed Understanding (DU).** This task evaluates a model's ability to interpret scene content in event streams, requiring precise identification of objects, activities, and contextual cues. Typical questions include identifying what is happening in the scene or recognizing key elements (*"What is the person doing?"*). This capability reflects the model's fundamental competence in understanding fine-grained semantics within dynamic event streams.

**Causal Reasoning (CR).** Inferring causal relationships from event streams requires interpreting temporal dependencies and intention cues (*"Why did the object fall?"*). This task evaluates a model's ability to uncover the underlying mechanisms behind actions and outcomes.

**Action Recognition (AR).** Identifying human actions from event stream requires detecting motion dynamics and inferring activity categories (*"What action is being performed?"*). This task evaluates a model's ability to capture coherent temporal dynamics under rapid motion.

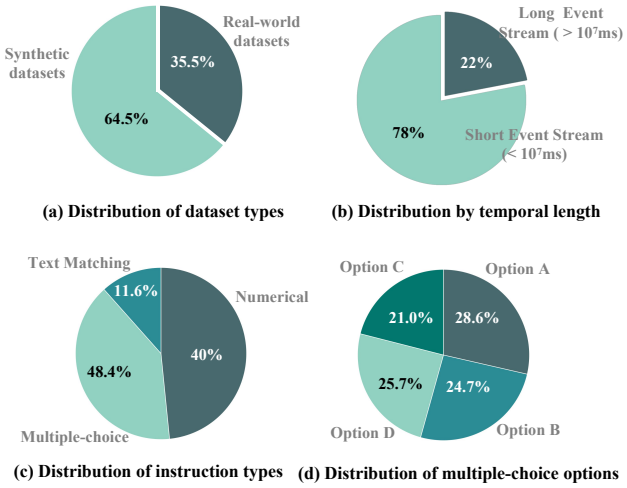



Figure 4. Data statistics in our EventBench across multiple dimensions. (a) Distribution of synthetic and real-world datasets. (b) Proportions of event streams by temporal length. (c)-(d) Distributions of instruction types and multiple-choice options.

**Gesture Recognition (GR).** Identifying human actions from the event stream requires detecting motion dynamics and inferring activity categories (*"What gesture is being performed?"*). This task evaluates a model's ability to capture temporal continuity under rapid motion, which is essential for reliable event-driven activity understanding.

**Event OCR (E-OCR).** Recognizing textual content in event streams requires identifying characters and reading dynamic text under highly challenging visual conditions (*"What text appears in the scene?"*). Since event cameras possess high temporal resolution and high dynamic range, they offer a distinct advantage for overcoming specific challenges like fast motion blur and sudden illumination changes. This enables robust text recognition in scenarios where conventional cameras fail.

**Spatial Relationship (SR).** Understanding relative object positions within 3D space requires determining where one object lies with respect to another (*"Where is X relative to Y?"*). This capability depends on grounding spatial cues and geometric structure, and is essential for robust 3D spatial reasoning in dynamic event stream environments.

**Absolute Distance (AD).** This task measures a model's ability to estimate metric distances between scene entities using spatial cues inferred from the event stream (*"How far apart are A and B?"*). It evaluates quantitative spatial reasoning and geometric awareness, reflecting the model's ability to internalize depth cues from event representations.

**Object Counting (OC).** Estimating the number of objects in an event stream (*"How many instances are there?"*) requires managing cluttered scenes and leveraging fine-grained spatial cues. The task evaluates numerical reasoning coupled with event-based spatial perception.

Table 3. Evaluation results on EventBench, showing three categories of models: closed-source MLLMs, open-source MLLMs (SFT), and open-source event-based MLLMs. The 8 metrics correspond to: 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-5[41] | - | 2025-08-07 | 68.5 | **70.1** | 46.0 | 49.8 | 22.2 | **47.8** | 30.7 | **25.6** | **45.1** |
| GPT-4o[40] | - | 2024-11-20 | 62.5 | 65.8 | **55.3** | 48.6 | 19.1 | 45.5 | 17.8 | 21.5 | 42.0 |
| Gemini2.5-pro[14] | - | 2025-06-17 | 61.3 | 63.9 | 44.1 | **49.9** | 19.1 | 40.5 | 17.5 | 16.9 | 39.2 |
| Gemini2.5-flash[14] | - | 2025-06-17 | 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-VL[3] | 3B | Qwen2.5 | 59.8 | 65.6 | 52.9 | 35.2 | **45.6** | **44.9** | 31.9 | 59.9 | 49.5 |
| InternVL3.5[57] | 4B | Qwen3 | 73.1 | **75.0** | **69.3** | **45.2** | 38.9 | 40.5 | **46.8** | **61.1** | **56.2** |
| MiniCPM-V-4[51] | 4.1B | MiniCPM4 | **72.9** | 63.9 | 58.9 | 37.1 | 44.4 | 44.2 | 43.1 | 55.6 | 52.5 |
| MiMo-VL[52] | 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** | | | | | | | | | | | |
| EventGPT[37] | 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** |

### 3.4. 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 publicly released; (ii) Task diversity: it covers a wide range of understanding, recognition, and spatial reasoning tasks, enabling comprehensive evaluation of model capabilities; (iii) Spatial integration: it pioneers the inclusion of 3D spatial reasoning tasks for event-based MLLMs; and (iv) Data scale: it includes a training set of over one million event–text pairs, supporting large-scale training and evaluation.

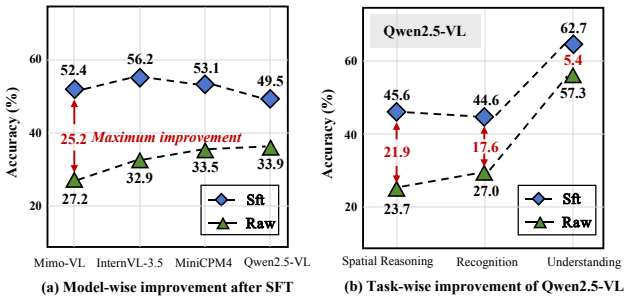



Figure 5. Overall performance improvement on EventBench after SFT training. (a) Model-wise improvement with a 25.2% maximum gain. (b) Task-wise improvement of Qwen2.5-VL across spatial reasoning, downstream, and open-domain QA tasks.

## 4. Experiments

### 4.1. 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

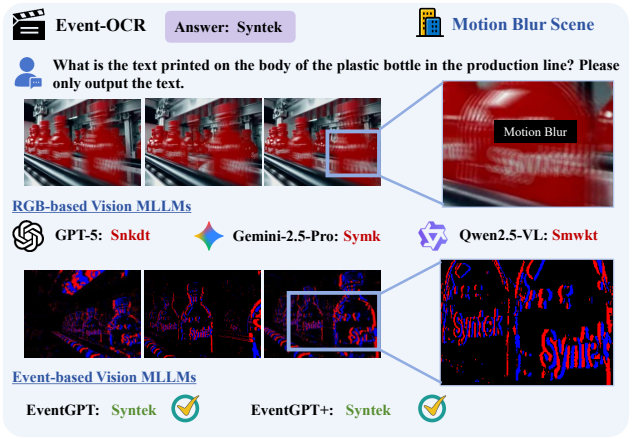



Figure 6. Comparison of frame-based and event-based MLLMs on the Event-OCR task under motion blur scenarios. Event-based models (i.e., EventGPT and EventGPT+) successfully recognize "Syntek", while RGB-based models fail due to motion blur.

Table 4. Performance of open-source MLLMs before and after SFT on EQA-1.4M, evaluated across on 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) |

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.

## 4.2. 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 3, 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. 8 presents a radar-style comparison of model performance across EventBench, highlighting the distinctions among commercial models,

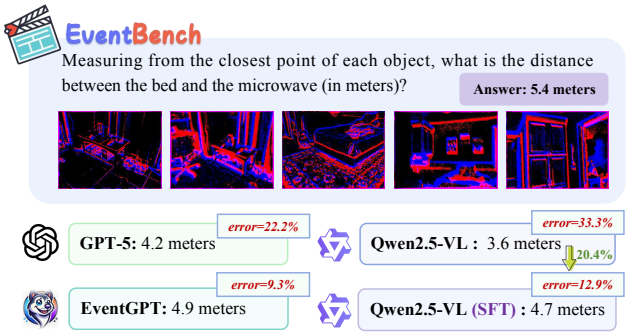



Figure 7. Comparison of model performance on the Absolute Distance task in EventBench. EventGPT reaches a 9.3% error, and Qwen2.5-VL reduces its error by 20.4% after SFT.

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.

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 anal-

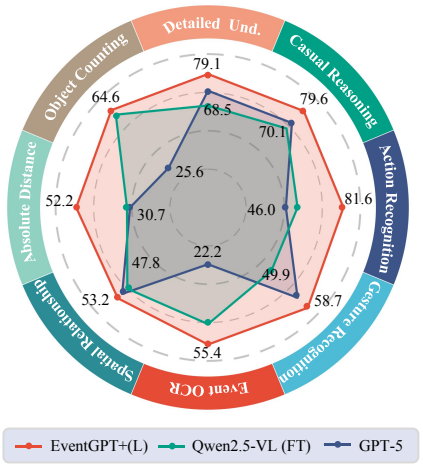



Figure 8. Comparison of EventGPT+(L), Qwen2.5-VL (FT), and GPT-5 on 8 evaluation tasks in EventBench. EventGPT+(L) exhibits the best overall performance in EventBench.

yses on challenging scenarios from EventBench. As illustrated in Fig. 6, 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. Moreover, Fig. 7 presents a comparison of the Absolute Distance task in EventBench. 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 4 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). In particular, Qwen2.5-VL attains improvements of 49.1%, 10.2%, and 6.1% for three spatial reasoning 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 5, varying $\Delta t$ 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.

Table 5. Performance comparison of open-source and event-based multimodal large language models (MLLMs) under different sampling intervals $\Delta t$ to assess the impact of temporal granularity.

| Model | Venue | Params | Sampling Interval ($\Delta t$) | | | |
|---|---|---|---|---|---|---|
| | | | $\Delta t = 0.5$ | $\Delta t = 1.0$ | $\Delta t = 1.5$ | $\Delta t = 2.0$ |
| **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 |

Table 6. 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 |

**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 6, 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.

## 5. 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


[1] Muhammad Aitsam, Sergio Davies, and Alessandro Di Nuovo. Event camera-based real-time gesture recognition for improved robotic guidance. In *Proceedings of the International Joint Conference on Neural Networks*, pages 1–8, 2024. 3, 4

[2] Gianluca Amprimo, Alberto Ancilotto, Alessandro Savino, Fabio Quazzolo, Claudia Ferraris, Gabriella Olmo, Elisabetta Farella, and Stefano Di Carlo. Ehwgesture–a dataset for multimodal understanding of clinical gestures. *arXiv*, 2025. 3, 4

[3] 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. 1, 2, 3, 6

[4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 3, 4

[5] 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. 4

[6] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv*, 2025. 3, 4

[7] 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*, pages 961–970, 2015. 3, 4

[8] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv*, 2019. 3, 4

[9] 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*, pages 5105–5113, 2025. 1

[10] Jiaben Chen and Huaizu Jiang. Sportsslomo: A new benchmark and baselines for human-centric video frame interpolation. *arXiv*, 2023. 3, 4

[11] 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*, 2025. 1

[12] Rui Chen, Xingyu Chen, Shaoan Wang, Shihan Kong, and Junzhi Yu. Let-us: Long event-text understanding of scenes. *arXiv preprint arXiv:2508.07401*, 2025. 1, 2, 3, 4

[13] Jang Hyun Cho, Andrea Madotto, Effrosyni Mavroudi, Triantafyllos Afouras, Tushar Nagarajan, Muhammad Maaz, Yale Song, Tengyu Ma, Shuming Hu, Suyog Jain, et al. Perceptionlm: Open-access data and models for detailed visual understanding. *arXiv*, 2025. 3, 4

[14] 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. 2, 6

[15] 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*, pages 5828–5839, 2017. 3, 4

[16] Pierre De Tournemire, Davide Nitti, Etienne Perot, Davide Migliore, and Amos Sironi. A large scale event-based detection dataset for automotive. *arXiv*, 2020. 2

[17] Yiting Dong, Yang Li, Dongcheng Zhao, Guobin Shen, and Yi Zeng. Bullying10k: A neuromorphic dataset towards privacy-preserving bullying recognition. *CoRR*, 2023. 3, 4

[18] Yipeng Du, Tiehan Fan, Kepan Nan, Rui Xie, Penghao Zhou, Xiang Li, Jian Yang, Zhenheng Yang, and Ying Tai. Motionsight: Boosting fine-grained motion understanding in multimodal llms. *arXiv*, 2025. 3, 4

[19] 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*, pages 24108–24118, 2025. 1

[20] 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. 1

[21] 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*, pages 3586–3595, 2020. 1

[22] 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. 3, 4

[23] Botao He, Ze Wang, Yuan Zhou, Jingxi Chen, Chahat Deep Singh, Haojia Li, Yuman Gao, Shaojie Shen, Kaiwei Wang, Yanjun Cao, et al. Microsaccade-inspired event camera for robotics. *Science Robotics*, 9(90):eadj8124, 2024. 1

[24] 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*, pages 8450–8460, 2025. 1, 3, 4

[25] 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*, pages 1–6, 2020. 2

[26] 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*, pages 1312–1321, 2021. 3

[27] Yixuan Hu, Yuxuan Xue, Simon Klenk, Daniel Cremers, and Gerard Pons-Moll. Controlevents: Controllable synthesis of event camera datawith foundational prior from image diffusion models. *arXiv*, 2025. 1

[28] 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*, pages 2146–2156, 2021. 2, 3, 4

[29] 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*, pages 2378–2388, 2024. 1

[30] 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. 2

[31] 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, 2025. 1

[32] 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. 1, 2

[33] 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*, pages 22195–22206, 2024. 1

[34] Pengteng Li, Yunfan Lu, Pinghao Song, Wuyang Li, Huizai Yao, and Hui Xiong. Eventvl: Understand event streams via multimodal large language model. *arXiv*, 2025. 1, 2, 4

[35] 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*, pages 8656–8667, 2025. 1

[36] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv*, 2023. 2

[37] 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*, pages 29139–29149, 2025. 1, 2, 3, 4, 6

[38] 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. 1

[39] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv*, 2023. 2

[40] OpenAI. Gpt-4o system card. https://openai.com/index/gpt-4o-system-card/, 2024. 2, 6

[41] OpenAI. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/, 2025. 2, 6

[42] 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. 2

[43] Joachim Ott, Zuowen Wang, and Shih-Chii Liu. Text-to-events: Synthetic event camera streams from conditional text input. In *Proceedings of the Neuro Inspired Computational Elements Conference*, pages 1–10, 2024. 1

[44] Xin Peng, Yifu Wang, Ling Gao, and Laurent Kneip. Globally-optimal event camera motion estimation. In *Proceedings of the European Conference on Computer Vision*, pages 51–67, 2020. 1

[45] 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. 2

[46] 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*, pages 16–30, 2025. 1

[47] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3857–3866, 2019. 1

[48] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale

dataset of paired third and first person videos. *arXiv*, 2018. 3, 4

[49] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv*, 2012. 4

[50] 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*, pages 7211–7219, 2025. 1

[51] 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. 1, 2, 6

[52] Xiaomi MiMo Team. Mimo-vl technical report. *arXiv*, 2025. 2, 6

[53] 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*, pages 22637–22646, 2024. 1

[54] Zengyu Wan, Ganchao Tan, Yang Wang, Wei Zhai, Yang Cao, and Zheng-Jun Zha. Event-based optical flow via transforming into motion-dependent view. *IEEE Transactions on Image Processing*, 33:5327–5339, 2024.

[55] Jin Wang, Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Unsupervised video deraining with an event camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10831–10840, 2023. 1

[56] 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*, pages 55–72, 2024. 3, 4

[57] 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*, 2025. 2, 6

[58] 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*, pages 22958–22967, 2025. 1

[59] 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*, pages 5615–5623, 2024. 3, 4

[60] 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*, 2025. 1, 4

[61] 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*, 2025. 1, 3

[62] Ziyi Wu, Xudong Liu, and Igor Gilitschenski. Eventclip: Adapting clip for event-based object recognition. *arXiv*, 2023. 2

[63] Yan Yang, Liyuan Pan, and Liu Liu. Event camera data pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10699–10709, 2023. 1

[64] Yan Yang, Liyuan Pan, and Liu Liu. Event camera data dense pre-training. In *Proceedings of the European Conference on Computer Vision*, pages 292–310, 2024. 1

[65] 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*, pages 4628–4638, 2025. 2

[66] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv*, 2023. 2

[67] Jiqing Zhang, Xin Yang, Yingkai Fu, Xiaopeng Wei, Baocai Yin, and Bo Dong. Object tracking by jointly exploiting frame and event domain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13043–13052, 2021. 1

[68] 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. 1

[69] 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*, pages 335–351, 2024. 1

[70] 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/, 2024. 1

[71] 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. 1

[72] 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. 1, 2, 3, 4

[73] Jiazhou Zhou, Xu Zheng, Yuanhuiyi Lyu, and Lin Wang. Eventbind: Learning a unified representation to bind them all for event-based open-world understanding. In *Proceedings of the European Conference on Computer Vision*, pages 477–494. Springer, 2024. 2

[74] 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*, pages 18633–18643, 2024. 1, 2

[75] 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*, pages 13691–13701, 2025. 1

[76] 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. 2

# EventBench: Towards Comprehensive Benchmarking of Event-based MLLMs

## Supplementary Material

## 1. Event Camera Sensing Mechanism

Event cameras, often referred to as dynamic vision sensors (DVS), operate by emitting asynchronous events in response to changes in pixel-wise log-intensity rather than capturing absolute brightness values. This sensing principle, which is reminiscent of the signal transduction mechanisms of biological retinas, stands in sharp contrast to conventional frame-based imaging, where measurements are acquired at fixed temporal intervals. An event $e_n$ is generated at pixel $\mathbf{u}_n = (x_n, y_n)$ once the variation in logarithmic luminance surpasses a contrast threshold $C$:

$$\log\left(\frac{I(x_n, y_n, t_n)}{I(x_n, y_n, t_n - \Delta t)}\right) = p_n C, \tag{1}$$

where $p_n \in \{+1, -1\}$ denotes the polarity of the brightness change and $\Delta t$ indicates the inter-event time at the same location. The resulting observation sequence can thus be written as

$$\mathcal{E} = \{(x_n, y_n, t_n, p_n)\}_{n=1}^{N}. \tag{2}$$

Due to their sparse output and microsecond-level temporal resolution, DVS sensors achieve extremely low latency and maintain reliable performance under rapid motion as well as challenging illumination conditions (e.g., low-light environments). These advantages are further enhanced by their inherently high dynamic range, which typically exceeds 120 dB. Despite these favorable sensing characteristics, existing evaluation resources provide only partial coverage of the capabilities required to assess event-based multimodal large language models (MLLMs). To address this gap, EventBench introduces a unified and systematically constructed benchmark that evaluates event-based MLLMs across multiple dimensions (i.e., understanding, recognition, and spatial reasoning) in diverse real-world scenarios.

## 2. 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.*.

**How Do Event-based MLLMs Perform Across Tasks?** To examine the overall performance of event-based MLLMs across diverse tasks, we evaluate a wide range of models as reported in Tab. 3. Our analysis reveals two major observations. (1) *Event-based MLLMs perform well on general understanding tasks.* As shown in Tab. 3, the state-of-the-art event-based model (i.e., EventGPT+(L)) achieves an average accuracy of 79.4% on understanding tasks, substantially outperforming strong closed-source and open-source MLLMs. This indicates that event-based MLLMs exhibit strong capability in generic scene understanding. (2) *Event-based MLLMs underperform on domain-specific tasks.* Despite their strong general abilities, event-based MLLMs still struggle with domain-specific tasks (e.g., gesture recognition, event-OCR, and spatial reasoning). This suggests that the world knowledge embedded in LLMs facilitates general event understanding and enables certain zero-shot capabilities, yet remains insufficient for specialized recognition and spatial reasoning tasks.

**What Are the Principal Performance Bottlenecks?** In Tab. 3, 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. 4, 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. 4. Our analysis yields two key observations. (1) *SFT on event stream data contributes to reducing the modality gap.* As shown in Tab. 4, 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. 3, 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,

Table 7. Evaluation results on EventBench for additional closed-source MLLMs (i.e., GPT-5-mini, GPT-4o-mini) and open-source RGB-based MLLMs (i.e., LLaVA-Next-Video), providing a broader comparison.

| Model | Params | Backbone | Understanding | | Recognition | | | Spatial Reasoning | | | Overall |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | DU | CR | AR | GR | E-OCR | SR | AD | OC | |
| **• Closed-Source MLLMs** | | | | | | | | | | | |
| GPT-5-mini | – | – | 67.2 | 67.8 | 51.5 | 49.2 | 8.0 | 42.9 | 6.0 | 30.2 | 40.4 |
| GPT-4o-mini | – | – | 60.5 | 61.7 | 50.1 | 47.4 | 15.5 | 46.2 | 22.4 | 16.2 | 40.0 |
| **• Open-Source MLLMs (SFT)** | | | | | | | | | | | |
| LLaVA-Next-Video | 7B | Qwen2 | 62.3 | 62.8 | 51.6 | 36.2 | 42.8 | 41.2 | 36.9 | 58.2 | 49.0 |

and that fine-tuning with million-scale event samples can mitigate the gap, though considerable disparity remains.

## 3. Dynamic Event Bin Selection

Event streams possess extremely high temporal resolution, making long-duration sequences easily exceed the context limits of MLLMs. Furthermore, the quality of motion perception depends critically on how the continuous stream is partitioned into temporal bins. To balance temporal coverage and information density under a fixed computational budget, we introduce a simple yet effective dynamic event bin selection mechanism. Instead of adopting the spatial-temporal aggregator used in EventGPT, we build a stronger baseline, termed EventGPT+, and release multiple model sizes, including EventGPT+(B) and EventGPT+(L).

To enable computationally scalable processing of the event stream, We consider its timestamp set $\mathcal{T} = \{t_1, \ldots, t_M\}$ spanning the observation interval $[t_{\min}, t_{\max}]$. We subsequently partition this interval into $N_b$ equal-duration event bins, with the $k$-th bin formally defined as:

$$\mathcal{I}_k = \big[\, t_{\min} + k\Delta,\ t_{\min} + (k+1)\Delta \,\big), \tag{3}$$

where $\Delta = (t_{\max} - t_{\min})/N_b$. Events assigned to $\mathcal{I}_k$ are further time-normalized by $\tilde{t}i^{(k)} = t_i - (t\min + k\Delta)$, ensuring a more stable and consistent temporal alignment across varying event densities and dynamic motion patterns, thereby achieving robust and adaptive normalization with enhanced temporal coherence consistent with Algorithm 1.

Within each bin, we build a histogram $h_k(j)$ over uniformly quantized sub-intervals and apply an overlapping sliding window of width $W$ (e.g., 10,ms with 5,ms stride). For each window start index $s$, the aggregated event count is computed as in a robust and temporally coherent manner:

$$C_k(s) = \sum_{j=s}^{s+W-1} h_k(j). \tag{4}$$

The window with the largest event count is chosen as the representative temporal slice. Low-activity bins (where $\max_s C_k(s)$ falls below a preset threshold) fall back to their temporal center to maintain temporal coverage.

This dynamic event bin selection strategy preserves the computational budget while adaptively focusing on motion-salient segments, leading to improved temporal adaptability.

**Algorithm 1** Dynamic Event Bin Selection

**Input:** Event timestamps $\mathcal{T} = \{t_1, \ldots, t_M\}$; number of bins $N_b$; window size $W$
**Output:** Representative event subsets $\{\hat{\mathcal{E}}_k\}_{k=0}^{N_b-1}$
$t_{\min} \leftarrow t_1,\ t_{\max} \leftarrow t_M$
$\Delta \leftarrow (t_{\max} - t_{\min})/N_b$
**for** $k = 0$ to $N_b - 1$ **do**
  $\mathcal{I}_k \leftarrow [\, t_{\min} + k\Delta,\ t_{\min} + (k+1)\Delta \,)$
  $\mathcal{E}_k \leftarrow \{\, t_i \in \mathcal{T} \mid t_i \in \mathcal{I}_k \,\}$
  $\tilde{t}_i^{(k)} \leftarrow t_i - (t_{\min} + k\Delta)$
  Build histogram $h_k(\cdot)$ over $\tilde{t}_i^{(k)}$
  **for** each window start $s$ **do**
    $C_k(s) \leftarrow \sum_{j=s}^{s+W-1} h_k(j)$
  **end for**
  $s_k^* \leftarrow \arg\max_s C_k(s)$
  $\hat{\mathcal{E}}_k \leftarrow \{\, t_i \in \mathcal{E}_k \mid \tilde{t}_i^{(k)} \in [\, s_k^*,\ s_k^* + W \,)\}$
**end for**
**return** $\{\hat{\mathcal{E}}_k\}_{k=0}^{N_b-1}$

## 4. Data Source Diversity

EventBench is constructed from a broad collection of over 20 real-world and synthetic data sources, as illustrated in Fig. 9. These sources cover real-world scenarios (i.e., human-centric, driving, object-centric, and text-centric) and synthetic environments (i.e., outdoor, sports, indoor-scanning, and action-centric). This extensive scene diversity stands in sharp contrast to existing event-stream datasets and benchmarks, which are often restricted to a single domain such as driving. By encompassing heterogeneous motion patterns, spatial structures, and semantic contexts, EventBench mitigates dataset-specific overfitting and

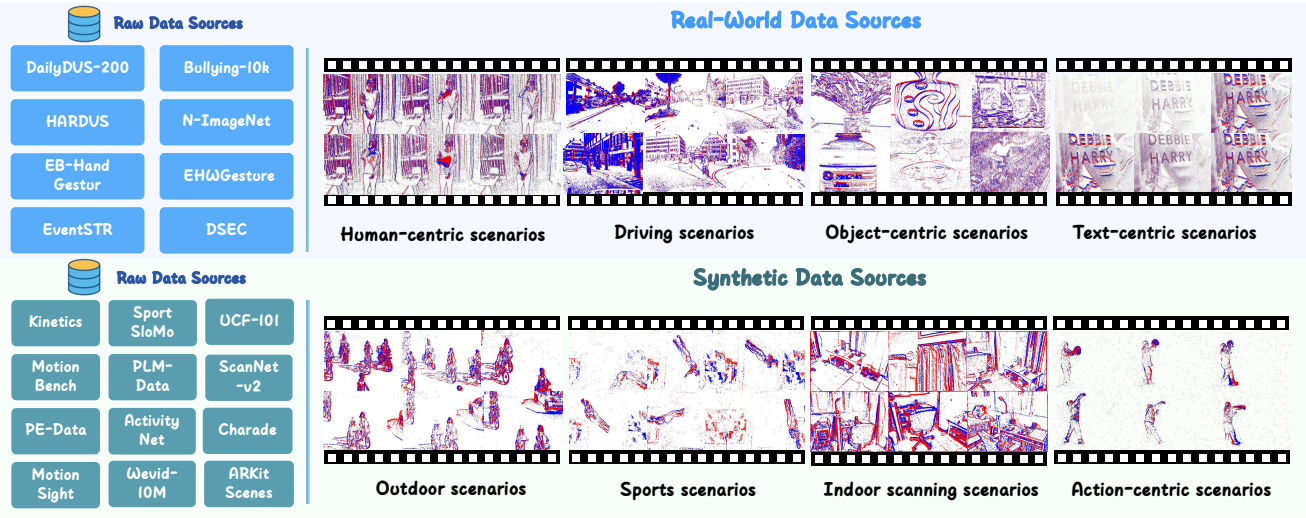



Figure 9. Diverse data sources of EventBench encompass real-world (i.e., human-centric, driving, object-centric, text-centric) and synthetic (i.e., outdoor, sports, indoor-scanning, action-centric) scenarios. This broad coverage provides rich motion patterns, spatial structures, and semantic contexts, supporting comprehensive evaluation of event-stream understanding, recognition, and spatial reasoning.

reduces evaluation bias, thereby supporting a more generalizable assessment of understanding, recognition, and spatial reasoning in event-based MLLMs.

## 5. More results

This section provides supplementary experiments, including extended evaluations over additional MLLMs in Sec. 5.1 and an analysis of inference efficiency under long-horizon event scenarios in Sec. 5.2.

### 5.1. Extended Evaluation

Due to space limitations in the main paper, we include additional evaluations here. To achieve a more comprehensive characterization of the capabilities of current MLLMs on event stream understanding, recognition, and spatial reasoning, we further broaden the evaluation scope. Specifically, we include additional closed-source MLLMs (i.e., GPT-5-mini, GPT-4o-mini) as well as open-source MLLMs (i.e., LLaVA-Next-Video). As shown in Tab. 7, we report extended results covering a wider set of architectures evaluated under the same protocol. These experiments not only offer a more thorough assessment of the practical performance of existing MLLMs on event-driven tasks but also establish informative reference baselines for future research on event-based MLLMs.

### 5.2. Inference Efficiency Analysis

We evaluate TTFT on the spatial reasoning subset of EventBench, where event streams reach minute-level duration. This setting imposes the strongest temporal burden and thus best reflects long-horizon inference efficiency.

Table 8. Token-to-First-Token (TTFT) latency comparison across open-source MLLMs and open-source event-based MLLMs on the spatial reasoning subset of EventBench.

| Model | Params | LLM Backbone | TTFT (s) | Average Acc |
|---|---|---|---|---|
| • Open-Source Video MLLMs | | | | |
| Qwen2.5-VL | 3B | Qwen2.5 | 2.76 | 49.5 |
| InternVL-3.5 | 4B | Qwen3 | 2.27 | 56.2 |
| MiniCPM4-V | 4.1B | MiniCPM4 | 4.22 | 52.5 |
| Mimo-VL | 7B | MiMO | 6.78 | 52.3 |
| LLaVA-Next-Video | 7B | Qwen2 | 6.89 | 49.0 |
| • Open-Source Event-Based MLLMs | | | | |
| EventGPT | 7B | Vicuna-v1.5 | 0.58 | 63.5 |
| EventGPT+(B) | 2B | Qwen2.5 | **0.49** | 63.8 |
| EventGPT+(L) | 7B | Qwen2.5 | 0.82 | **65.5** |

The TTFT comparison conducted on an NVIDIA A800 GPU (Tab. 8) reveals a clear and persistent efficiency gap between RGB-based MLLMs and event-based MLLMs. Open-source RGB models (i.e., Qwen2.5-VL, InternVL-3.5, MiniCPM4-V, MiMO-VL, and LLaVA-Next-Video) exhibit high latency, ranging from 2.27 s to 6.89 s, driven by dense frame tokenization and the computational load of long video sequences. In contrast, event-based MLLMs operate within 0.49 s to 0.82 s, benefiting from the sparse and asynchronous structure of event streams. Under this challenging condition, EventGPT+(B) achieves the lowest TTFT at 0.49 s while maintaining competitive accuracy, and EventGPT+(L) delivers the highest accuracy at 65.5%, yet remains an order of magnitude faster than most RGB-based models. These results show that event-based MLLMs achieve a more favorable performance–efficiency trade-off.