Seeing Motion Through Polarity for Event-based Action Recognition

Supplementary Material

This document provides more details of experimental setting, visualization example and prompt setting. The contents are meticulously organized in a manner that facilitates ease of understanding and reference, as detailed in the subsequent sections :

- *Experimental Setting (cf. §A)*;
- *Visualization Example (cf. §B)*.
- *Prompt Setting (cf. §C)*.

A. Experimental Setting

We evaluate our research on three benchmarks, including SeAct [5], DVS Action [3] and THU$^{\text{E-ACT}}$-50-CHL [2].

SeAct dataset is the first semantic-abundant dataset for event-text action recognition. It is collected with a DAVIS346 event camera, whose resolution is 346 × 260. It contains 58 actions under four themes, each action is accompanied by an action caption of less than 30 words generated by GPT-4 to enrich the semantic space of the original action labels.

DVS Action dataset contains 10 action categories with 450 recordings, all captured using the DAVIS346 camera. The action was conducted in an unoccupied office environment, involving the participation of 15 subjects who performed a total of 10 different actions.

THU$^{\text{E-ACT}}$-50-CHL dataset is under challenging scenarios. It contains 58 actions and comprises 2,330 recordings featuring 18 students, captured from varied perspectives utilizing a DAVIS346 event camera in two distinct scenarios: a long corridor and an open hall. Every recording lasts from 2 to 5 seconds and maintains a spatial resolution of 346 × 260.

For event-to-image reconstruction, we follow EMP [1] to adopt the most representative lightwork network-Firenet [4]. For the event representation, we adopt the method from [5] to reformat the event stream into three-channel event frames. For the text prompts, we utilize a combination of the hand-crafted text prompt and the learnable text prompt, maintaining consistency with ExACT [5].

B. Visualization Example

Example for Zero-Shot EAR. As shown in 1, we illustrate the complete prompting scheme utilized when testing the Multi-modal Large Language Models (MLLMs) for event-based action recognition. The prompt structure is kept consistent across all tested MLLMs, comprising a System Prompt and a User Prompt. The MLLM's output is the predicted label category, which is then compared against the ground-truth label to calculate the accuracy.

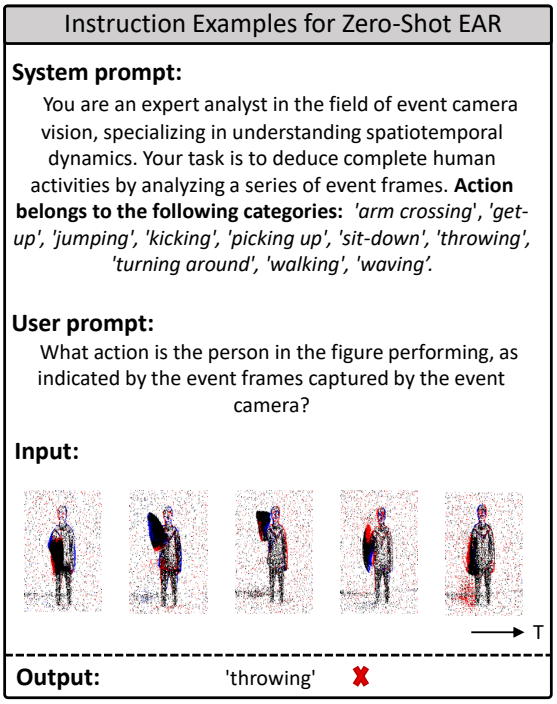


Figure 1. Visualization of instruction examples for zero-shot action recognition.

C. Prompt Setting

Complete Progressive Prompting in PMR. As shown in 2, we illustrate the complete progressive prompting scheme employed in our proposed Polarity Motion Reasoner (PMR). This scheme is designed to explicitly emphasise that positive polarity represents the forward tendency of motion, while negative polarity signifies the trailing persistence of movement. The generation of the final textual description for polarity-enhanced motion is guided by a three-stage progressive instruction pipeline: (1) *Observation* (focusing on raw input cues), (2) *Thinking* (inferring motion characteristics), and (3) *Synthesis and Prediction* (selecting the optimal motion tendency description).

References

[1] Meiqi Cao, Xiangbo Shu, Xin Jiang, Rui Yan, Yazhou Yao, and Jinhui Tang. Exploiting frequency dynamics for enhanced multimodal event-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5969–5979, 2025. 1

[2] Yue Gao, Jiaxuan Lu, Siqi Li, Nan Ma, Shaoyi Du, Yipeng Li, and Qionghai Dai. Action recognition and benchmark using

Instruction Examples for PMR

System prompt:

You are an expert analyst in the field of event camera vision, specializing in understanding spatiotemporal dynamics. Your task is to deduce complete human activities by analyzing a series of event frames. In this context, polarities indicate motion: positive polarity (red) events denote the leading edge of movement, and negative polarity (blue) events denote the trailing edge of movement. For each new key frame, you must use the following structured reasoning process to update and expand your ongoing analysis:

Step 1: Observation

Locate the person in the current frame and **identify** their status.

Step 2: Thinking

a) Analyze Instantaneous Motion: Determine the immediate direction and nature of movement (e.g., rotation, swinging, pushing).

b) Integrate Temporal Context: Link the just observed instantaneous movement with the broader action narrative built from previous frames.

(Example reasoning: "In the previous frame, we observed a dense cluster of red moving upward (arm raising). In the current frame, we see a red arc moving downward. This indicates the motion entering a second phase: a downward swing, directly resulting from the initial arm lift.')

Step 3: Synthesis and Prediction

Summarize the complete narrative motion (including body areas): Briefly restate the entire event sequence you've observed from the first frame to the current frame. For example: 'First, the object was lifted, remained stationary, and is now swinging forward with the arm.'

Adjust your prediction: Based on the comprehensive narrative so far, provide your most plausible prediction of the executed complete action. As more frames appear, this prediction should become more confident or be adjusted.

User prompt:

Please analyze the provided sequence of event frames. Determine the final output format:

Your entire response **MUST** be organized according to the following structure:

Thinking (Frame 1→T):

Instantaneous Motion: [analysis from Step 2a for the current frame]

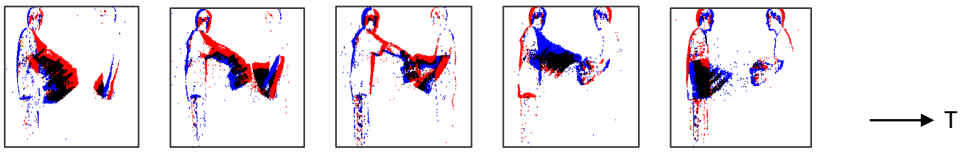
Temporal Integration: [key analysis from Step 2b, clearly linking current frame with past frames]

Overall Motion Narrative: [only describe the critical entire movement with motion regions, formatted as: {Movement Region, Movement Trend}->{Movement Region, Movement Trend}->...

Prediction: [final, concise action derived from the complete narrative]

!Check!: As the sequence progresses, your prediction should become more specific, but if evidence is lacking, always choose a general action (e.g., 'swinging') over a possibly incorrect specific action (e.g., 'playing tennis').

Input:



Output:

{Left Person's Arm, Extends Forward}->{Object, Transfers from Left to Right Person}->{Left Person's Arm, Retracts}

Figure 2. Visualization of instruction examples for PMR.

event cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):14081–14097, 2023. 1

[3] Shu Miao, Guang Chen, Xiangyu Ning, Yang Zi, Kejia Ren, Zhenshan Bing, and Alois Knoll. Neuromorphic vision datasets for pedestrian detection, action recognition, and fall detection. *Frontiers in Neurorobotics*, 13:38, 2019. 1

[4] Cedric Scheerlinck, Henri Rebecq, Daniel Gehrig, Nick Barnes, Robert Mahony, and Davide Scaramuzza. Fast image reconstruction with an event camera. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer*

Vision, pages 156–163, 2020. 1

[5] 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