EXACT: Event-based Action Recognition via Conceptual Reasoning and Uncertainty Estimation with Language Guidance
–*Supplementary Material*–

Anonymous CVPR submission

Paper ID 8228

1. The AFE Representation

Algorithm 1 provides the pseudo-code of the Adaptive Fine-grained Event (AFE) presentation in a Python-like style. To implement the AFE representation recursively, we define a function named 'RECUSIVE FUNC' and finally obtain the 'FrameList' consisting of a series of fine-grained event frames I_M^T.

Besides, the denoising of the event count image is also crucial for the effectiveness and stability of the AFE presentation. The noise can significantly interfere with the accuracy of calculating the ratio of overlapped sub-actions (the difference rate R), particularly when the count of the event sub-stream is relatively low. Consequently, we employ the morphological open operation with 2×2 kernels [6] for denoising, which is very effective since the event noise is sparsely and randomly distributed over the spatial space [2].

Algorithm 1: Pseudo-code of the AFE representation a Python-like style.

Input : Event stream E_0^0, Minimum sample event number: N_{min}, Maximum sample threshold: Δ;
Output: FrameList consisting of a series of fine-grained event frames I_M^T;

```
FrameList = []
RECUSIVE FUNC(E_0^0, N_min, Δ, FrameList)
Function RECUSIVE FUNC(E_0^0, N_min, Δ, FrameList)
    Divide the event stream equally E_0^0 to obtain
      event stream parts E_1^0 and E_1^1;
    Calculate the different rate R;
    if R ≤ Δ then
        Generate event frame I_0^0 from E_0^0;
        return FrameList.append(I_0^0);
    end
    if len(E_1^0) ≤ N_min or len(E_1^1) ≤ N_min then
        Generate event frame I_1^0 from E_1^0;
        Generate event frame I_1^1 from E_1^1;
        return FrameList.append([I_1^0, I_1^1]) ;
    end
    RECUSIVE FUNC(E_1^0, N_min, Δ, FrameList)
    RECUSIVE FUNC(E_1^1, N_min, Δ, FrameList)
end
```

2. SeAct Dataset

Dataset details Tab. 1 displays 58 actions of our SeAct dataset, belonging to four themes: (1) Body-Motion; (2) Human-Object Interaction; (3) Health-care Monitoring; (4) Human-Human Interaction. For every action, there are 10 event stream recordings from different people (6 males and 4 females).

Action caption generation Tab. 3 shows the generated action captions of all 58 actions in our SeAct dataset. We utilize the following text prompt to generate the action captions by GPT-4 [5]: `'Please describe the meaning of human action for [CLASS] in a sentence of about 15 words.'`, where [*CLASS*] denotes the action name.

3. Supplemental Experiment

3.1. Additional Dataset and Experimental Settings

We follow the official split settings of HARDVS [7] where 60%, 10%, and 30% of each category for training, validating, and testing, respectively. For the PAF dataset [3] without an official dataset split, we randomly split 80% and 20% of the dataset for training and testing. The Adam optimizer is employed with four RTX 3090 GPUs, resulting in a mini-batch size of 16 event-text pairs. The PyTorch architecture [4] serves as the fundamental for conducting all experiments.

For the ablation study of AFE representation, considering the TBR [1] doesn't release the official code, we implemented it ourselves, with the aggregated time interval set to

Body-Motion Only	Human-Object Interaction
clap	catch a ball
circle	throw a ball
jumping jack	catch and throw a ball
squat down	walk with a ball
jump squat	circle the ball around the main body
push-up	circle the ball around the leg
sit down	open and close umbrella
salute	open the computer
bend forward	close the computer
hurdle start	use the phone
long jump	put on glasses
nod head	put off glasses
walking	tie shoelaces
running	take a photo
shake head	lift the box
circle head	put down the box
circle arm	drink water
raise the arm	twist the bottle cap
side kick	walk with an opened umbrella
forward kick	walk with a box
high leg lift	run with a box
waving hand	
punch straight forward	
Health-care Monitoring	**Human-Human Interaction**
falling down	hug
vomit	fight
staggering	wave hand to each other
walk with stomach pain	handshake
walk with headache	shoulder tapping
walk with back pain	clap hand
leg injury walking	handing box

Table 1. The category of actions in our SeAct dataset.

2000 ms. This yielded a total of 2758 frames, maintaining a similar order of magnitude for frame amount as other comparative representations. Note that the official aggregated time interval is 20 ms, leading to 289,477 frames in total.

3.2. Ablation Studies

Impact of different text prompts. Five text prompts are designed to investigate their impacts on model performance. According to the results presented in Tab. 2, the text prompt `'A series of photos recording human action for'` presents the highest recognition performance (94.83% Top-1 accuracy), outperforming other four text prompts. Hence, this hand-crafted text prompt is chosen as the input of the text encoder.

Hyper-parameter searching of the AFE representation For the hyper-parameter searching of the minimum aggregating event count number N_{min} and lowest sampling threshold Δ, we conduct experiments on PAF, HARDVS, and our SeAct dataset, training for 10, 1, and 10 epochs based on their dataset sizes. Fig. 1 and Fig. 2 present the hyper-parameter search results of N_{min} and Δ, respectively. Based on the hyper-parameter search, we set the N_{min} as 100000, 100000, and 150000 for the PAF, SeAct, and HARDVS and set the Δ as 40, 40, and 50 for the SeAct, HARDVS, and PAF datasets. We believe searching with smaller intervals may lead to enhanced performance.

Text Prompt	Accuracy	
	Top-1	Top-5
A series of photos for	91.07	93.28
A series of frames recording human action for	91.75	94.03
A series of sketch images recording human action for	92.86	95.41
A series of photos recording human action for	**94.83**	**98.28**

Table 2. Effect of different text prompts.

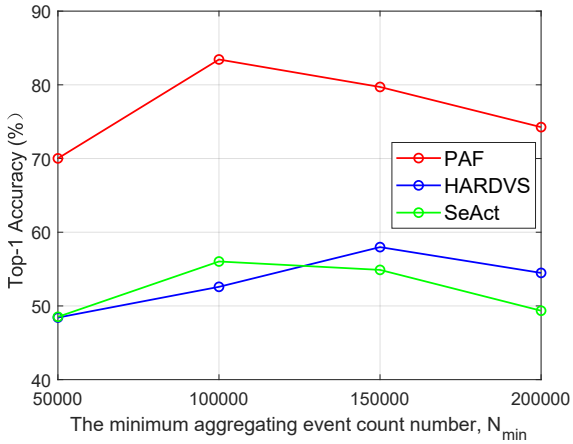


Figure 1. Hyper-parameter searching of the minimum aggregating event count number N_{min}.

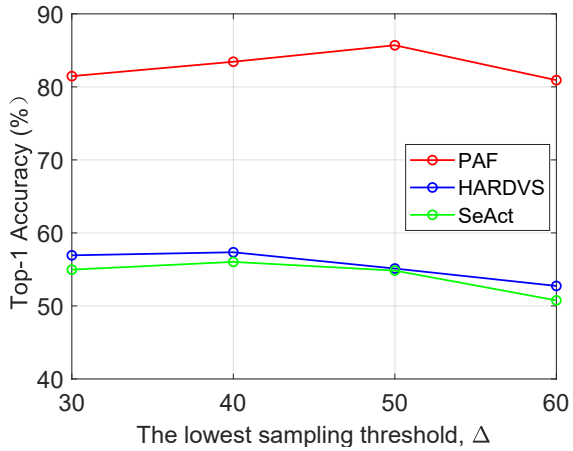


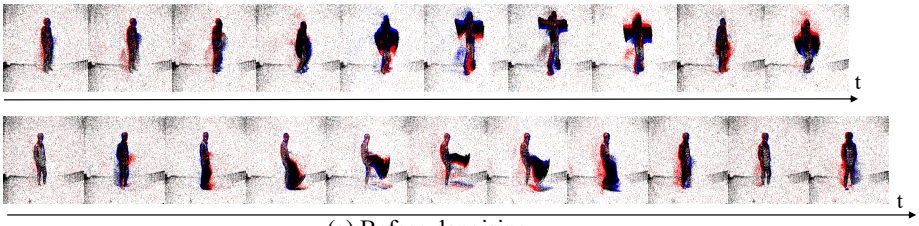
Figure 2. Hyper-parameter searching of the lowest sampling threshold Δ.

Visulization of event count image denoising results As shown in Fig. 3, we present visualization results of the denoising operations impact on the event count images mentioned in the AFE representation. Comparing the event count images before and after denoising utilizing the morphological open operation, we can observe that the noise is significantly suppressed. It proves the effectiveness of

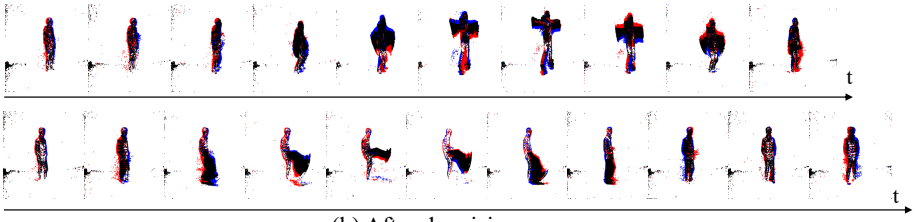
the denoising operation, which is important for ensuring the stability of the AFE representation.

3.3. Extension to Other Tasks

As shown in Fig. 4, we present more retrieval results for event-to-text and text-to-event tasks. All retrieved data exhibits a high degree of similarity to the input text or event query, proving the effectiveness of ExACT.

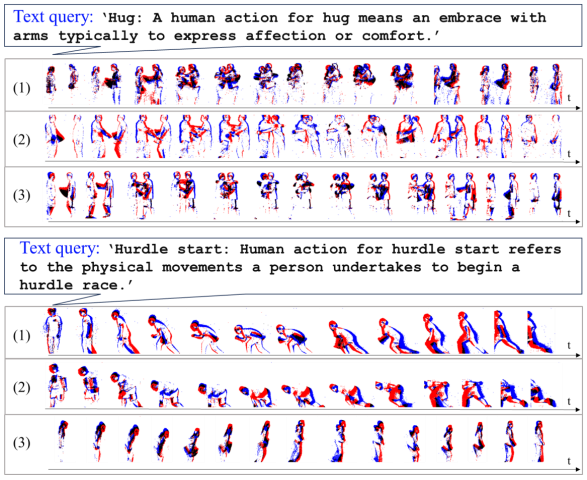


(a) Before denoising

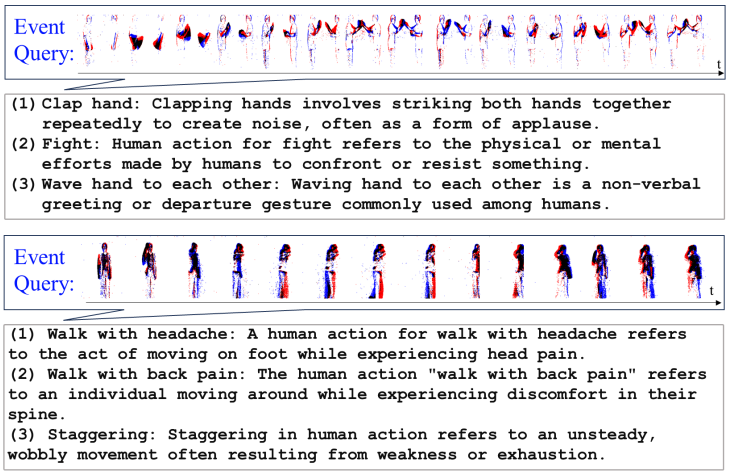


(b) After denoising

Figure 3. denosing results.



(a) Text-to-event



(b) Event-to-text

Figure 4. Event-to-text retrieval results.

1	”clap: Human action for clap refers to the act of striking one’s hands together to produce a sharp, loud sound.”
2	”circle: A human action for ”circle ” could mean drawing a circular shape or moving in a circular pattern.”
3	”jumping jack: A jumping jack is a physical exercise in which one jumps from a standing position with legs together and arms at the sides to a position with the legs apart and the arms above the head.”
4	”squat down: The human action ”squat down ” means to bend the knees and lower the body close to the ground.”
5	”jump squat: Jump squat is a human action involving a lower body exercise that combines a squat with a jump.”
6	”push-up: Human action for push-up refers to the body movement of raising and lowering oneself by arm strength.”
7	”sit down: Human action for sit down refers to a person lowering their body to rest in a seated position.”
8	”salute: A salute is a gesture of respect or acknowledgment, often used in military or ceremonial contexts.”
9	”bend forward: ”Bend forward” means to lean or incline the upper part of the body towards the front or ground.”
10	”hurdle start: Human action for hurdle start refers to the physical movements a person undertakes to begin a hurdle race.”
11	”long jump: Human action for long jump involves running, leaping, and landing to achieve maximum horizontal distance.”
12	”nod head: Nodding the head is a human action typically used to express agreement, affirmation or understanding.”
13	”walking: Human action for walking refers to the conscious, voluntary movement of legs for locomotion.”
14	”running: Running is a human action involving swift movement on foot where both feet leave the ground simultaneously.”
15	”shake head: Shake head as a human action means to move one’s head from side to side, typically indicating denial or disapproval.”
16	”circle head: Human action for circle head refers to the movements or behaviors performed by a character with a rounded face.”
17	”circle arm: Human action for circle arm refers to the movement where a person swings their arm in a circular motion.”
18	”raise the arm: Human action to raise the arm refers to the deliberate, volitional movement of the upper limb upwards.”
19	”side kick: A human action side kick involves someone performing a swift, sideways kicking movement, often in martial arts.”
20	”forward kick: Human action for forward kick involves projecting force from the leg to strike or propel something forward.”
21	”high leg lift: High leg lift in human action denotes the movement of raising one’s leg up to waist level or higher.”
22	”waving hand: Waving hand is a human action used as a gesture for greeting, attention-seeking or saying goodbye.”
23	”punch straight forward: Human action for punch straight forward refers to the physical movement of thrusting one’s fist directly ahead.”
24	”catch a ball: Human action to catch a ball refers to the conscious and coordinated physical effort to intercept and secure a thrown ball.”
25	”throw a ball: Human action for throwing a ball refers to the physical movement people perform to propel a ball.”
26	”catch and throw a ball: Human action for catching and throwing a ball involves coordination of motor skills, timing, and visual perception.”
27	”walk with a ball: Human action for ”walk with a ball” refers to a person moving while carrying or controlling a ball.”
28	”circle the ball around the main body: The person is moving the ball in a circular motion around their body.”
29	”circle the ball around the leg: Human action for circling the ball around the leg refers to maneuvering a ball around one’s limb.”
30	”open and close umbrella: Human action for open and close umbrella refers to the physical activity of expanding or folding an umbrella.”
31	”open the computer: The human action to open the computer refers to the act of power-on or initiating the computer system.”
32	”close the computer: Human action to close the computer refers to a person manually shutting down or turning off the PC.”
33	”use the phone: Human action refers to the steps taken by a person, such as dialing numbers on a phone.”
34	”put on glasses: The human action of putting on glasses involves lifting eyewear to the face and positioning them over the eyes.”
35	”put off glasses: The human action of putting off glasses refers to the act of removing eyewear from one’s face.”
36	”tie shoelaces: Human action for tying shoelaces involves intricately manipulating and knotting laces to secure shoes onto feet.”
37	”take a photo: Human action for taking a photo refers to someone using a camera or smartphone to capture images.”
38	”lift the box: Human action for lifting the box refers to a person using their strength and effort to elevate a container.”
39	”put down the box: The term refers to the deliberate movement carried out by a person to set a box on a surface.”
40	”drink water: Human action for drinking water refers to the voluntary activity of ingesting liquid H2O for hydration.”
41	”Twist the bottle cap: Human action for twisting the bottle cap involves applying force to rotate the cap for removal or tightening.”
42	”walk with an opened umbrella: A person strolling with an opened umbrella usually indicates protection from ongoing rain or harsh sun.”
43	”Walk with a box: Human action for walk with a box involves a person physically moving while carrying a container.”
44	”Run with a box: A human action for run with a box means the physical activity of a person jogging or sprinting while carrying a box.”
45	”falling down: Falling down refers to the involuntary action of losing balance and suddenly collapsing to the ground.”
46	”vomit: human action for vomit refers to the act of forcefully expelling stomach contents through the mouth.”
47	”staggering: Staggering in human action refers to an unsteady, wobbly movement often resulting from weakness or exhaustion.”
48	”walk with stomach pain: A human action of walking with stomach pain refers to someone moving while experiencing abdominal discomfort.”
49	”walk with headache: A human action for walk with headache refers to the act of moving on foot while experiencing head pain.”
50	”walk with back pain: The human action ”walk with back pain” refers to an individual moving around while experiencing discomfort in their spine.”
51	”leg injury walking: Human action for leg injury walking refers to the deliberate movements made by an individual to accommodate a leg injury while walking.”
52	”hug: A human action for hug means an embrace with arms typically to express affection or comfort.”
53	”fight: Human action for fight refers to the physical or mental efforts made by humans to confront or resist something.”
54	”wave hand to each other: Waving hand to each other is a non-verbal greeting or departure gesture commonly used among humans.”
55	”handshake: A handshake is a human action symbolizing agreement, friendship, respect, or conclusion of a deal.”
56	”shoulder tapping: Shoulder tapping is a human action signifying attention-seeking, alerting someone, or initiating communication.”
57	”clap hand: Clapping hands involves striking both hands together repeatedly to create noise, often as a form of applause.”
58	”handing box: Human action for handing box refers to the physical motion of a person giving or passing a box to someone else.”

Table 3. The generated action captions of our SeAct dataset.

References

[1] Simone Undri Innocenti, Federico Becattini, Federico Pernici, and Alberto Del Bimbo. Temporal binary representation for event-based action recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 10426–10432. IEEE, 2021. 1

[2] Ana I Maqueda, Antonio Loquercio, Guillermo Gallego, Narciso García, and Davide Scaramuzza. Event-based vision meets deep learning on steering prediction for self-driving cars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5419–5427, 2018. 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] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 1

[5] Partha Pratim Ray. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 2023. 1

[6] Luc Vincent. Morphological area openings and closings for grey-scale images. In *Shape in picture: mathematical description of shape in grey-level images*, pages 197–208. Springer, 1994. 1

[7] 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. *arXiv preprint arXiv:2211.09648*, 2022. 1