ExACT: Language-guided Conceptual Reasoning and Uncertainty Estimation for Event-based Action Recognition and More

Jiazhou Zhou[1] Xu Zheng[1] Yuanhuiyi Lyu[1] Lin Wang[1,2*]
[1]AI Thrust, HKUST(GZ) [2]Dept. of CSE, HKUST
{jiazhouzhou,yuanhuiyilv}@hkust-gz.edu.cn, zhengxu128@gmail.com, linwang@ust.hk
Project Page: https://vlislab22.github.io/ExACT/

Abstract

Event cameras have recently been shown beneficial for practical vision tasks, such as action recognition, thanks to their high temporal resolution, power efficiency, and reduced privacy concerns. However, current research is hindered by 1) the difficulty in processing events because of their prolonged duration and dynamic actions with complex and ambiguous semantics and 2) the redundant action depiction of the event frame representation with fixed stacks. We find language naturally conveys abundant semantic information, rendering it stunningly superior in reducing semantic uncertainty. In light of this, we propose ExACT, a novel approach that, for the first time, tackles event-based action recognition from a cross-modal conceptualizing perspective. Our ExACT brings two technical contributions. Firstly, we propose an adaptive fine-grained event (AFE) representation to adaptively filter out the repeated events for the stationary objects while preserving dynamic ones. This subtly enhances the performance of ExACT without extra computational cost. Then, we propose a conceptual reasoning-based uncertainty estimation module, which simulates the recognition process to enrich the semantic representation. In particular, conceptual reasoning builds the temporal relation based on the action semantics, and uncertainty estimation tackles the semantic uncertainty of actions based on the distributional representation. Experiments show that our ExACT achieves superior recognition accuracy of 94.83%(+2.23%), 90.10%(+37.47%) and 67.24% on PAF, HARDVS and our SeAct datasets respectively.

1. Introduction

Action recognition is a crucial vision task with many applications, such as robot navigation [33, 43], and abnormal human behavior recognition [25, 35]. Many frame-based learning approaches have been presented, leading to impressive performance improvements [24, 43]. However, these methods may not be ideal solutions for power-constrained scenarios, *e.g.*, surveillance [3, 24]. RGB cameras also degrade due to environment bias [36] like motion blur and lighting variations. Moreover, frame-based cameras trigger considerable privacy concerns as they directly capture users' appearance.

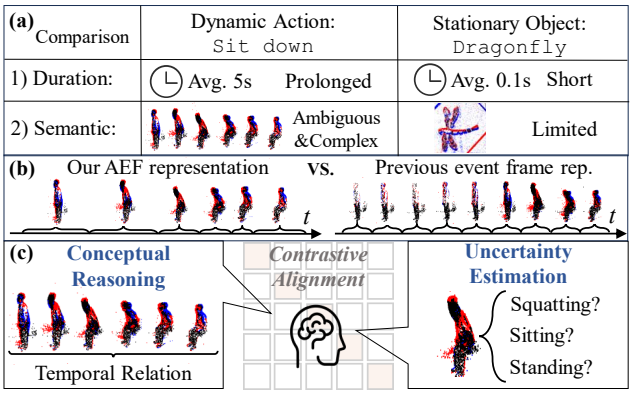


Figure 1. **(a)** Unlike stationary objects, *e.g.*, `'Dragonfly'` with short duration (0.1s) and limited semantics, dynamic actions like `'Sit down'` have the prolonged duration (5s) with ambiguous and complex semantics. **(b)** Compared with previous event representation, stacking events with fixed counts, we adaptively filter out events recording stationary actions while preserving dynamic ones; **(c)** We introduce language guidance to stimulate the recognition process, particularly focusing on conceptually reasoning temporal relations and estimating uncertain semantics.

Recently, bio-inspired event cameras are gaining popularity [1, 14, 27], which ignore the background and only record moving objects. This leads to sensing efficiency and resilience to rapid motion and illumination changes with low power consumption. Also, event cameras mostly reflect objects' edges, which alleviates users' privacy concerns like skin color and gender. Owing to these advantages, event-based action recognition offers more pragmatic solutions for real-world implementations. This has inspired research endeavors [14, 27, 36, 39, 49, 54, 55] in event-based action recognition areas with plausible performance.

*Corresponding author.

However, the above methods are deficient for two reasons: 1) They have a limited capacity for recognizing a large number of different human actions, as demonstrated by experiments on the HARDVS [48] dataset with 300 categories in Tab. 1. This probably arises from the complex and ambiguous semantics induced by the dynamic actions and prolonged duration (around 5s [48]), compared to objects with short duration (around 0.1s [48]) and limited semantics. For example, as shown in Fig. 1 (a), `'Dragonfly'` *vs.* `'Sit down'`. 2) The lack of *tailored event representation* as the raw events are directly stacked into event frames with fixed stacks, resulting in event frames with either overlapped or vague edge information depicting the same action, see Fig. 3 and Fig. 1 (b).

Recent advancements in vision-language models (VLMs) [6, 11, 37] have pioneered ideas that incorporate semantic concepts across text and vision modalities, aiming at simulating the human processes of conceptualization and reasoning [19, 41, 42]. *The key insight is that language naturally conveys inherent semantic richness, which can be beneficial for modeling semantic uncertainty and establishing complex semantic relations*. Inspired by this, we introduce language as guidance for event-based action recognition. As the first exploration, the research hurdles include: 1) How to represent events to depict dynamic actions in detail without redundant event frames? 2) How to integrate text embeddings with event embeddings to help reason complex semantics of dynamic actions and reduce semantic uncertainty?

To this end, we propose a novel ExACT framework to tackle event-based action recognition from a cross-modal conceptual reasoning perspective, as depicted in Fig. 1 (b) and (c). To address the first challenge, an Adaptive Fine-grained Event (AFE) representation (Sec. 3.1) is inspired by the 'overlapped action regions' observed in Fig. 3. These regions indicate an excessive stack of events in one frame, which is inevitable for previous frame-based representations with fixed stacks. Differently, our AFE recursively and offline find the dividing line of different actions based on the overlapped regions. It eliminates repeated events and preserves dynamic actions, thus enhancing model performance without extra computational costs (Tab. 2).

For the second challenge, we propose a novel Conceptual Reasoning-based Uncertainty Estimation (CRUE) module (Sec. 3.3) to simulate the action recognition process of humans. Concretely, CRUE initially establishes the temporal relation of event frames by leveraging the text embeddings to reason each frame's semantics and then obtain the fused event embeddings. Subsequently, CRUE converts event and text embeddings from discrete representation to distributional representation, where the distribution variance quantifies semantic uncertainty. In this way, our proposed CRUE module establishes a semantic-abundant and uncertainty-aware embedding space to enhance model performance (Tab. 3).

Meanwhile, as existing datasets solely provide category-level labels, we propose the SeAct dataset, consisting of 58 categories of actions, with semantic-abundant caption-level labels. Our dataset serves as the first dataset for event-text action recognition (67.24% accuracy). We also conduct extensive experiments to show that our ExACT outperforms previous methods , *e.g.*, [14, 26] on the public datasets, PAF [34] 94.83% accuracy (+2.23%) and especially on HARDVS [48] 90.10% accuracy (+37.47%) by a large margin. Beyond action recognition, our ExACT can be flexibly applied to the event-text retrieval task.

In summary, our main contributions are: (I) We propose the ExACT– the first framework utilizing language guidance for event-based action recognition; (II) We propose the CRUE module to mimic human action recognition, creating a rich, uncertainty-aware cross-modal embedding space for action recognition. Also, our AFE representation adaptively filters redundant events, yielding effective representation for dynamic actions. (III) We introduce the SeAct dataset with detailed text captions for evaluating the recognizing ability of actions composed of multiple sub-actions with different semantics. Extensive experiments demonstrate the superiority of our ExACT framework on our SeAct dataset and public datasets.

2. Related Works

Event-based Action Recognition methods can be categorized into the event-only and event-other-modality methods. For event-only frameworks, the most widespread techniques involve stacking the event stream into compact frames, followed by the utilization of off-the-shelf Convolutional Neural Networks (CNNs) [14] or Vision Transformers (ViTs) [39, 55] for effective feature extraction. This approach currently exhibits state-of-the-art performance owing to the excellent CNN/ViT backbones performance. Meanwhile, considering the asynchronous characteristics of event data, the research community has explored the applicability of bio-inspired Spiking Neural Networks (SNNs) [27] and the spatiotemporal capabilities of Graph Convolutional Networks (GCNs) [54] to more aptly resonate with the unique structure of event data. However, these approaches have yielded suboptimal performance and exhibited limited adaptability, partly due to the specialized hardware requisites inherent to SNNs.

Concurrently, there have been endeavors to integrate event data with additional modalities. For example, incorporating the abundance of color and texture information in RGB [49] data with event information or utilizing the supplementary motion knowledge in optical flow [36]. In summary, most of these approaches rely on dense consecutive event frames, inevitably resulting in redundant frames with

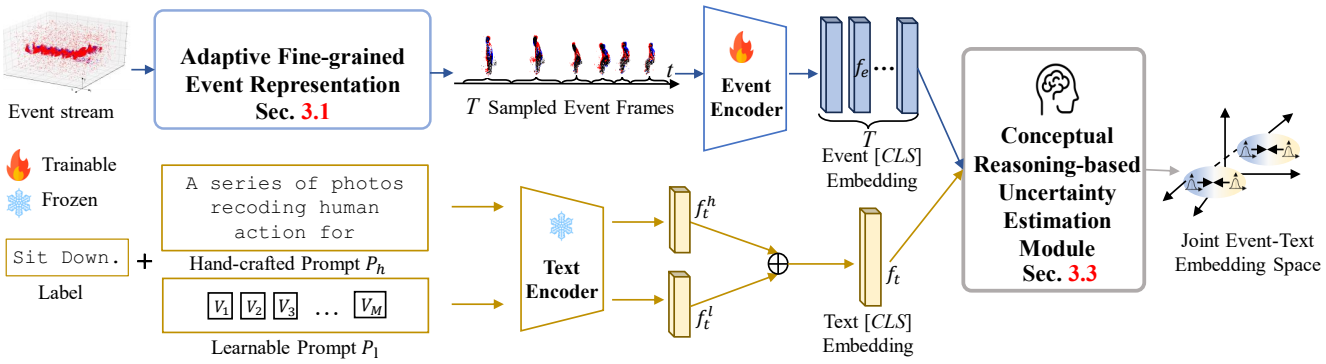


Figure 2. **Overall framework of our proposed ExACT framework.** It consists of four components: (1) the AFE representation recursively eliminates repeated events and generates event frames depicting dynamic actions (Sec. 3.1); (2) the event encoder and (3) the text encoder, responsible for the event and text embedding, respectively (Sec. 3.2); (4) the CRUE module simulates the action recognition process to establish the complex semantic relations for sub-actions and reduce the semantic uncertainty. (Sec. 3.3)

overlapped actions and uncertain semantics. To represent events to depict detailed dynamic actions, the AFE representation is proposed to sample event frames adaptively without introducing extra computation costs.

Vision-laguage Models (VLMs) Recently, there has been a growing interest in large-scale pre-trained VLMs [6, 11] for multimodal representations. Inspired by it, several pioneering works [7, 52, 58] have investigated the potential for transferring VLMs' capability to the event modality, thereby revitalizing the best performance of objection recognition. In addition, the remarkable zero-shot capability of VLMs has motivated researchers to explore event-based label-free [7] or zero (few)-shot applications [58], thus addressing the scarcity of high-quality event datasets. Nevertheless, prior event-text methods focus on recognizing objects with limited semantics but fail to recognize events recording actions with prolonged time duration and complex and ambiguous semantics. Therefore, Our ExACT aims at enhancing event-based action recognition from a cross-modal conceptual reasoning perspective.

3. The Proposed ExACT Framework

Overview. An overview of our ExACT framework is depicted in Fig. 2. *The key idea of ExACT is to introduce language as guidance for estimating semantic uncertainty and establishing semantic relations for event-based action recognition.* The following subsections explain the technical details of 1) the proposed Adaptive Fine-grained Event (AFE) representation (Sec. 3.1); 2) the event encoder and the text encoder (Sec. 3.2); 3) the Conceptual Reasoning-based Uncertainty Estimation (CRUE) module (Sec. 3.3). Besides, in Sec. 3.5, we introduce our proposed **s**emantic-abundant **e**vent-based **act**ion recognition (**SeAct**) dataset as the first dataset for event-text action recognition.

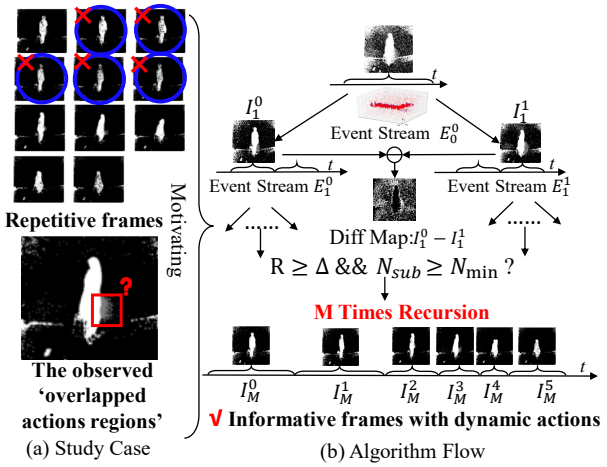


Figure 3. **(a)** Unlike existing methods often lead to repetitive event frames, our AFE representation adaptively filters out repetitive events for the same action based on the observed overlapped action regions; **(b)** Illustration of the AFE representation.

3.1. The AFE Representation

Most event-based action recognition models [14, 39, 55] primarily rely on event frame representations [31, 56], which is compatible with off-the-shelf CNN/ViT backbones. For these models, the event stream is spatially integrated into frames with fixed event counts or time duration [57], as shown in Fig. 1 (b). However, the high temporal resolution of event data inevitably leads to a profusion of repetitive event frames displaying the same stationary objects (refer to Fig. 3 (a) blue circles). Consequently, it is arduous for such representation to depict dynamic actions. To this end, we seek to answer the following question: *Can we adaptively filter out repetitive events for stationary objects while preserving events recording the dynamic actions?*

Accordingly, we visualize the previous event frame representation [57] in Fig. 3 (a). A salient observation is the 'overlapped action region', as marked by the red square,

which is a byproduct of the frame transformation process where consecutive actions overlap due to excessive event stacks. This 'overlapped action region' thus serves as a pivotal indicator for inappropriate event sampling intervals. In light of this, we propose the AFE representation.

Specifically, to find the most appropriate dividing line of different actions based on the 'overlapped action regions', we adopt the classic *binary search* and implement it *offline and recursively* with efficient $O(logn)$ algorithm complexity. As illustrated in Fig. 3 (b), the search algorithm can be seen as finding leaf nodes of a binary tree. To begin with, we equally divide the event stream E_0^0 (root) into two sub-streams E_1^0 (node) and E_1^1 (node) and thus generate their corresponding event count images I_1^0 and I_1^1. Then, we subtract I_1^0 with I_1^1 to obtain the difference map. Next, to measure the proportion of overlapped actions based on the difference map, we define a factor named deference rate R, given by:

$$R = sum(abs(I_1^0 - I_1^1))/(sum(E_0^0)/2), \qquad (1)$$

where the $sum(.)$ and $abs(.)$ functions indicate the event counts and absolute value operations, respectively. Intuitively, the high value of R indicates a high probability of stacking events recording two different actions into one frame. In this case, we need to divide the event sub-stream recursively.

For the recursive algorithm, the boundary conditions are important. In our cases, if the difference rate R is higher than the lowest sampling threshold Δ, we repeat the above division process until it's lower than Δ or the event count number of the sub-stream N_{sub} is less than the minimum aggregating event count number N_{min}. Note that hyper-parameters N_{min} and Δ vary with different datasets. *More discussions about selecting N_{min} and Δ refer to the Supplmat.* After the above searching process with M times recursion, we finally obtain a series of fine-grained event frames I_M^T with T event frames (all leaf nodes).

3.2. Feature Encoding

With the AFE representation, the event stream is processed into a series of fine-grained event frames I_M^T. Then, the event encoder and text encoder from the pre-trained event-text model of [58] are utilized to establish a unified event-text embedding space.

Event Encoder. As shown in Fig. 2, it inputs an RGB event frame $I_M^i \in \mathbb{R}^{H\times W\times 3}, i = 1, 2, ..., T$ of the spatial size $H \times W$ and outputs the event embedding f_e^i. Given T event frames, the event encoder processes T times and generates the event [*CLS*] embeddings $f_e \in \mathbb{R}^{T\times H_e\times W_e\times N_e}$.

Text Encoder. It takes two different kinds of text prompts as input: **1)** the hand-crafted text prompt `'A series of photos recording action for [CLASS].'`, where [*CLASS*] represents the category name. After encoding, each word is converted into the D_p-dimension word embedding and formed into the final text token P_h **2)** the learnable text prompt $P_l = [P_1, P_2, ..., P_n, P_{CLASS}.]$, where $P_i, i = 1, 2, ..., n_l$, is a random initialized parameter with D_p dimensions; n_l denotes the number of the learnable text prompts; P_{CLASS} represents the encoded word embeddings of [*CLASS*] and [.] means the concatenation operation. Then, the text encoder transforms hand-crafted text prompt P_h and learnable text prompt P_l into corresponding text embeddings f_t^h and f_t^l. We finally obtain the text [*CLS*] embeddings f_t by averaging f_t^h and f_t^l.

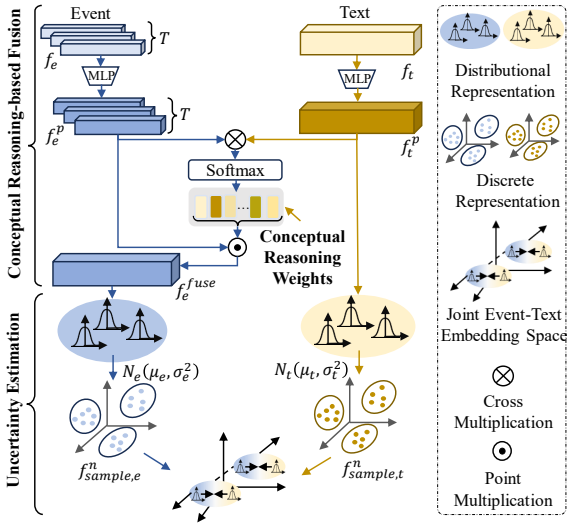


Figure 4. **The proposed CRUE module** consists of 1) conceptual reasoning for frame fusion based on the temporal relation among events and 2) uncertainty estimation of sub-actions for both text and event embeddings utilizing distributional representation.

3.3. CRUE Module

Previous event-based action recognition methods [14, 26] fail to consider the following aspects: **1) Temporal Relation**: Unlike stationary objects, dynamic actions unfold over time. The event data's temporal information is vital for understanding the meaning of an action. For instance, in Fig. 1, `'Sit down'` and `'Stand up'` involve similar sub-actions with the converse temporal occurrence, namely `'Standing'` → `'Squatting'` → `'Sitting'` *vs.* `'Sitting'` → `'Squatting'` → `'Standing'`, thus resulting in different semantics. **2) Semantic Uncertainty**: Actions, comprising various sub-actions, present greater semantic complexity than that of stationary objects. Take `'Sit down'` as an example. It includes sub-actions: `'Standing'`, `'Squatting'`, and `'Sitting'`. Each sub-action owns a specific meaning. Thus, using any sub-action is *insufficient* and *uncertain* to express the meaning of entire action `'Sit down'`. Motivated by these two aspects, we propose the CRUE module to emulate the action recognition process of humans through the proposed conceptual reasoning-based fusion and uncertainty estimation.

Conceptual Reasoning-based Fusion: Unlike the simple average fusion of event frames [14, 36, 39, 55], the proposed CRUE module leverages the text embeddings to guide the event frame fusion. Specifically, as depicted in Fig. 4, given the event [*CLS*] embeddings f_e and text [*CLS*] embeddings f_t, a two-layer MLP network is employed for dimension projection. In this way, we can obtain the projected event embeddings f_e^p and text embeddings f_t^p. Then, we multiply f_e^p and f_t^p followed by the softmax function to generate the conceptual reasoning weights. Next, the conceptual reasoning weights are multiplied with the original projected event embeddings f_e^p to obtain the fused event embeddings f_e^{fuse}. Intuitively, the conceptual reasoning weights are used as semantic weights generated based on event frames' temporal sequence for the frame fusion.

Uncertainty Estimation: Semantic uncertainty refers to the obtained messages that tend to present multiple targets [19]. To model the semantic uncertainty of actions, we borrow the idea of distributional representation applied both in Natural Language Processing (NLP) [2, 8, 47] and Computer Vision (CV) [5, 41, 42]. Unlike the methods that extract features as the discrete representation, we utilize the probability distribution encoder [19] for events and text embedding. Thus, the semantic uncertainty can be quantified by the variance of the probability distribution.

Specifically, as shown in Fig. 4, the fused event [*CLS*] embeddings $f_e^{fuse} \in \mathbb{R}^{H_e \times W_e \times N_e}$ are equally split into $f_{e,1}^{fuse}$ and $f_{e,2}^{fuse}$ at the channel dimension [19]. Then, $f_{e,1}^{fuse}$ and $f_{e,2}^{fuse}$ are fed into two standard self-attention modules [46]. Next, we can predict a mean vector $\mu_e \in \mathbb{R}^{H_e \times W_e \times N_e}$ and a variance vector $\sigma_e \in \mathbb{R}^{H_e \times W_e \times N_e}$. Here, μ_e and σ_e are the estimated parameters of the multivariate Gaussian distribution $f_e^{fuse} \sim N_e(\mu_e, \sigma_e^2)$. We conduct the same operation on the text *CLS* embeddings f_t to estimate its corresponding multivariate Gaussian distribution $f_t^p \sim N_t(\mu_t, \sigma_t^2)$. Overall, the above operations can be formulated as follows:

$$\mu_i = Att_1(f_{i,1}), \sigma_i = Att_2(f_{i,2}), \tag{2}$$

$$f_i = [f_{i,2}, f_{i,1}] \sim N_i(\mu_i, \sigma_i^2), \tag{3}$$

where $i = e, t$ denotes the event and text embeddings respectively; Att represents the self-attention module and $[.]$ indicates the concatenation operation.

With estimated distributional representation $N_i(\mu_i, \sigma_i^2)$, $i = t, e$, we now quantify the semantic uncertainty of event and text embeddings. Then, we sample arbitrary discrete representation by employing the re-parameterization method [23] to ensure smooth back-propagation. That is, we first sample a random noise $\delta \sim N(0, I)$ from the standard normal distribution rather than sample directly from $f_i \sim N_i(\mu_i, \sigma_i^2), i = t, e$. Next, we obtain the sampled discrete embeddings $f_{sample,i}^n$ by $f_{sample,i}^n = \mu_i + \delta\sigma_i$, where $n = 1, 2, ..., N$ and N is a hyper-parameter denoting the number of sampled discrete embeddings. (More discussion about N refer to Sec. 4.3.) The obtained $f_{sample,i}^n$ observes the estimated distribution N_i, and thus it can be used for estimating semantic uncertainty.

However, the above random sampling process increases the complexity of training, especially given the spatial sparsity of event data. To accelerate model convergence, we introduce the *Smooth L1* loss [15] between the normalized *CLS* embeddings $f_i, i = t, e$ and the sampled embeddings $f_{sample,i}^n, i = t, e$:

$$L_{smoothL1} = SmoothL1(f_{sample,i}^n, \frac{f_i - mean(f_i)}{std(f_i)}), \tag{4}$$

where $n = 1, 2, ..., N$ and N is a hyper-parameter denoting the number of sampled discrete embeddings, $mean(.)$ and $std(.)$ denote the mean and variance of the input embeddings. Besides, to reduce the semantic uncertainty of distributional representation, we introduce regularization loss:

$$L_{reg} = sum(\sigma_i^2) + sum(\sigma_i^2), i = t, e, \tag{5}$$

where $sum(.)$ and $abs(.)$ indicate the summation of input embeddings and absolute value operation, respectively. Note that the experiment results show the final L_{reg} is higher than zero when the model converges. This indicates the model doesn't degenerate from distributional representation into discrete representation as the variances are greater than zero.

3.4. Training Objectives

To establish a joint event-text representation space for action recognition, we utilize the contrastive loss $L(f_b^1, f_b^2)$ between two modal embeddings f_b^1 and f_b^2 as follows:

$$L_{contrastive}(f_b^1, f_b^2) = -\frac{1}{B}\sum_{b \in B} \log \frac{\exp\left(f_b^1 \times f_b^2/\tau\right)}{\exp\left(f_b^1 \times f_b^2/\tau\right) + \sum_{\bar{b} \neq b} \exp\left(f_b^1 \times f_{\bar{b}}^2/\tau\right)}, \tag{6}$$

where τ is the temperature coefficient, B represents the size of the mini-batch, b and $\bar{b}$ denote the b-th and the $\bar{b}$-th data among the mini-batch.

We calculate the contrastive loss among all sampled event embeddings $f_{sample,e}^n$ and text embeddings $f_{sample,t}^n$. Finally, the whole training objective is composed of the contrastive loss, the *Smooth L1* loss, and regularization loss combined with different rate hyperparameters:

$$\begin{aligned} L_{final} =& \alpha \times L_{contrastive}(f_{sample,e}^n, f_{sample,t}^n) \\ &+ \beta \times L_{smoothL1} + \theta \times L_{reg}, \end{aligned} \tag{7}$$

where we set the default values of α, β and θ to 1 after considering their numerical range.

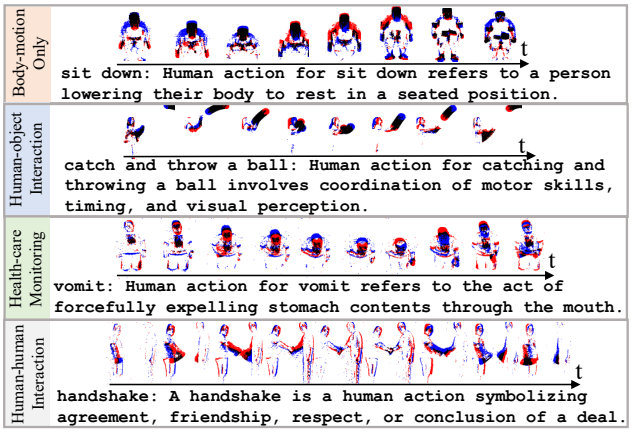


Figure 5. Examples of our SeAct dataset.

3.5. SeAct Dataset

Since the category-level labels provided in the previous event action dataset [1, 4, 34, 48] only use several words to describe each action, they fail to stimulate the ability of our ExACT framework to process complex language information. To this end, we propose the **first** semantic-abundant SeAct dataset for event-text action recognition, where the detailed caption-level label of each action is provided. SeAct is collected with a DAVIS346 event camera whose resolution is 346 × 260. It contains 58 actions under four themes, as presented in Fig. 5. Each action is accompanied by an action caption of less than 30 words generated by GPT-4 [38] to enrich the semantic space of the original action labels. We split 80% and 20% of each category for training and testing (validating), respectively. We believe our SeAct dataset will be a better evaluation platform for event-text action recognition and inspire more relevant research in the future. *Please refer to the Supplmat. for more dataset introduction.*

4. Experiments

4.1. Dataset and Experimental Settings

Dataset In this work, four datasets are adopted for the evaluation of our proposed model, including PAF [34], HARDVS [48], DVS128 Gesture [1] and our newly proposed SeAct. PAF [34], also named Action Recognition, is a dataset recorded indoors, containing ten action categories with 450 recordings. HARDVS [48] is a recently released dataset for event-based action recognition, currently having the largest action classes, namely 107,646 recordings for 300 action categories. Both of the above two datasets have an average time duration of 5 seconds with 346 × 260 resolution [48]. DVS128 Gesture [1] is collected using a DVS128 camera with 128 × 128 resolution, dividing into 11 hand and arm gestures. *Refer to Sec. 3.5 for the introduction of our SeAct dataset.*

Experimental Settings In the AFE representation, the lowest sampling rate Δ is set as 50%, 40%, and 40% while the minimum aggregating event number N_{min} is chosen at 100,000, 150,000, and 100,000 on the PAF, HARDVS our SeAct datasets, respectively. The event encoder and text encoder of the event-image-text model ECLIP [58] are utilized for feature encoding. The number of sampled discrete embeddings N is set to 5 based on the hyperparameter search. The initial learning rates are set to $1e-5$ with Adam optimizer [22] and weight decay equal to $2e-4$. CosineAnnealingLR [30] learning rate schedule is employed with a minimum learning rate of 1e-6. Our model is trained for 100 epochs for PAF and SeAct datasets, and 25 epochs for HARDVS. *Please refer to the Supplmat. for additional experimental settings.*

Dataset	Model	Accuracy (%)	
		Top-1	Top-5
PAF	HMAX SNN [53]	55.00	-
	STCA [16]	71.20	
	Motion SNN [27]	78.10	-
	MST [51]	88.21	-
	Swin-T (BN) [51]	90.14	-
	EV-ACT [14]	<u>92.60</u>	-
	ExACT (Ours)	**94.83**	**98.28**
HARDVS	X3D [12]	45.82	52.33
	SlowFast [13]	46.54	54.76
	ACTION-Net [50]	46.85	56.19
	R2Plus1D [45]	49.06	56.43
	ResNet18 [17]	49.20	56.09
	TAM [28]	50.41	57.99
	C3D [44]	50.52	56.14
	ESTF [48]	51.22	57.53
	Video-SwinTrans [29]	51.91	59.11
	TSM [26]	<u>52.63</u>	<u>60.56</u>
	ExACT (Ours)	**90.10**	**96.69**
DVS128 Gesture	Time-surfaces [32]	90.62	-
	SNN eRBP [20]	92.70	-
	Slayer [40]	93.64	-
	DECOLLE [21]	95.54	-
	EvT [39]	96.20	-
	TBR [18]	97.73	-
	EventTransAct [9]	<u>97.92</u>	-
	ExACT (Ours)	**98.86**	**98.86**
SeAct	EventTransAct [9]	57.81	64.22
	EvT [39]	61.30	67.81
	ExACT-category	<u>66.07</u>	<u>70.54</u>
	ExACT-caption	**67.24**	**75.00**

Table 1. An overall comparison with SoTA models for event-based action recognition task on the PAF, HARDVS, DVS128 Gesture, and our proposed SeAct dataset. The best scores are in bold, and the second scores are underlined. 'ExACT-category' and 'ExACT-caption' represent that ExACT is trained with category-level and caption-level labels, respectively.

4.2. Comparison with SOTA Methods

As shown in Tab. 1, our proposed ExACT demonstrates superior performance on the PAF and HARDVS datasets. Specifically, ExACT brings +2.23% improvements on the PAF dataset with ten classes compared with SOTA re-

Sample strategy	Frame number	Aggregated method	Accuracy	
			Top-1	Top-5
Event histogram	3122	every 80,000 events	89.29	91.07
	2720	every 90,000 events	94.21	97.14
	2405	every 100,000 events	93.10	95.24
Event voxel [10]	3122	every 80,000 events	88.88	90.46
	2720	every 90,000 events	92.45	94.89
	2405	every 100,000 events	90.85	92.31
TBR [18]	2758	every 2000 ms events	92.79	95.16
AFE (Ours)	2894	adaptive events number	**94.83**	**98.28**

Table 2. The comparison of AFE representation with previous event frame representation.

Method	Accuracy			
	Top-1	Δ	Top-5	Δ
Contrastive	92.86	-	94.64	-
+ CR	93.64	+0.78	96.43	+1.79
+ CR, UE	**94.83**	+1.97	**98.28**	+3.64

Table 3. The ablation study of the CRUE module, where Contrastive, CR, and UE denote the contrastive learning loss function, proposed conceptual reasoning-based fusion and uncertainty estimation, respectively.

Method	Accuracy			
	Top-1	Δ	Top-5	Δ
Sum	89.14	-	90.97	-
Mean Pooling	92.52	+3.38	95.04	+5.07
CR	**94.83**	+5.69	**98.28**	+8.31

Table 4. Impact of CR operation proposed in the CRUE module.

sults. Notably, ExACT brings remarkable +37.47% improvements on the HARDVS with 300 classes, showcasing ExACT's excellent potential in classifying complex and diverse actions. Moreover, upon evaluation using our proposed SeAct dataset, ExACT exhibits a 67.24% Top-1 and 75.00% Top-5 recognition accuracy in real-world scenarios involving 58 dynamic actions with caption-level labels. These results demonstrate the effectiveness of our ExACT framework in action recognition.

4.3. Ablation Studies

In this section, we ablate the key components of ExACT, training objectives, and important hyper-parameters to explore their effectiveness. Unless otherwise stated, experiments are performed on the PAF dataset.

Effectiveness of the AFE representation. Tab. 2 shows that, with a comparatively lower sampled number of 2816, our AFE representation obtains the best accuracy of 94.83%. This result indicates that our AFE representation can achieve better performance by filtering out the repetitive frames portraying the same action.

CRUE module *vs.* Contrastive learning. As shown in Tab. 3, the baseline model is trained using the contrastive learning loss, which is widely adopted in previous methods [52, 58]. Results indicate that the Conceptual Reasoning-based fusion (CR) and CR with Uncertainty Estimation (UE) lead to an increase in accuracy of +0.78% and +1.97%, respectively. This implies that the CRUE module enhances the model's ability to comprehend actions by conceptual reasoning-based fusing event frames based on their temporal relations and estimating the uncertainty of action semantics during training, thus achieving better performance than simply employing contrastive learning.

Training Method	Accuracy			
	Top-1	Δ	Top-5	Δ
$L_{contrastive}$	92.86	-	94.64	-
+ L_{reg}	93.96	+1.10	95.82	+1.18
+ $L_{smoothL1}$	94.41	+1.55	97.89	+3.25
+ L_{reg}, $L_{smoothL1}$	**94.83**	+1.97	**98.28**	+3.46

Table 5. Impact of different training objectives.

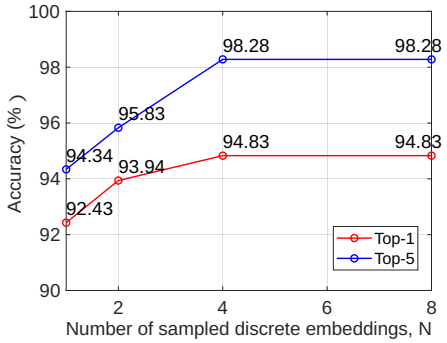


Figure 6. Impact of different numbers of sampled discrete embeddings proposed in the CRUE Module.

Conceptual Reasoning-based fusion (CR) *vs.* Other frame fusion methods To evaluate the effectiveness of the CR. We compare the CR with two other designs, namely, the sum and average pooling of all event frames, as shown in Tab. 3. The results demonstrate the effectiveness of our proposed CR, as it improves recognition accuracy by +5.69% and +2.31% when compared to the sum operation and the mean pooling operation, respectively.

Performance comparison of different training objectives. Tab. 5 shows the impact of different training objectives on model performance. Model trained using the contrastive learning loss $L_{contrastive}$ (Eq. 6) exhibits the lowest performance compared to other combinations of training objectives. Both the L_{reg} (Eq. 5) and $L_{smoothL1}$ (Eq. 4) enhance the model performance, bringing +1.10% and +1.55% accuracy improvements respectively. The combination of all training objectives brings the largest performance improvement of +1.97%, which further validates the effectiveness of the proposed CRUE module.

Impact of the different number of point sampling. As shown in Fig. 6, we explore the effect of the number of sampled discrete embeddings N. We find that as N increases from 1 to 4, the accuracy increases from 92.43% to 94.83%. When N increases from 4 to 8, the accuracy re-

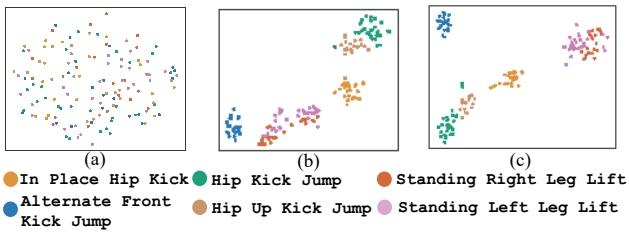


Figure 7. t-SNE visualization of event embeddings on the HARDVS dataset. (a) Before training; (b) Training without CRUE module; (c) Training with CRUE module.

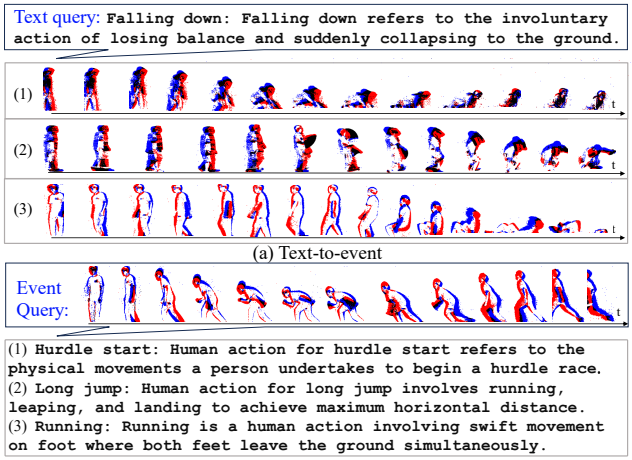
mains constant, indicating that increasing N yields diminishing performance accuracy improvement. Intuitively, this comes from the fact that distributional representation introduces disturbance during training while more sampled discrete embeddings mitigate this disturbance. Consequently, we set the hyperparameter N as 5 during training.

The t-SNE visualization of event embeddings for the CRUE module. We select 144 event examples belonging to six categories from the HARDVS dataset. These categories include a simple pair with different semantics, as well as two challenging pairs with similar semantics, namely `'Hip Kick Jump'` *vs.* `'Hip Up Kick Jump'` and `'Standing Right Leg Lift'` *vs.* `'"Standing Left Leg Lift"'`. Fig. 7 displays the alteration of event embedding distribution before training and training with/without the CRUE module. Comparing those hard pairs in Fig. 7 (b) and (c), we can find that event embeddings in different categories are widely distributed while our proposed CRUE module advances in distinguishing those hard pairs with similar semantic meanings. Intuitively, visualization results prove that the CRUE module enhances the model's performance by expressing the uncertainty of semantics, especially for those with similar semantics. *See more results in Supplmat.*

4.4. Extension to Other Tasks

In this section, we transfer our EXACT model to both text-to-event and event-to-text retrieval tasks. Retrieval refers to searching and retrieving required data that matches a given query. The query can be of different modalities like events or texts. It has several practical applications, such as matching a real-world abnormal action with its corresponding event data in surveillance scenarios. For action retrieval using EXACT, we feed the query and events/texts into corresponding encoders to obtain their embeddings. Next, we calculate the similarity score between the query embedding and event/text embeddings, then select those events/texts with the highest similarity scores as the output. All retrieval experiments are conducted on our SeAct dataset. *Refer to Supplmat. for more retrieving results.*

Text-to-event action retrieval In Fig. 8 (a), we present the Top-3 retrieved event streams utilizing the action



(a) Text-to-event

(1) Hurdle start: Human action for hurdle start refers to the physical movements a person undertakes to begin a hurdle race.
(2) Long jump: Human action for long jump involves running, leaping, and landing to achieve maximum horizontal distance.
(3) Running: Running is a human action involving swift movement on foot where both feet leave the ground simultaneously.

(b) Event-to-text

Figure 8. Action retrieval results.

`'Falling down'` with captions. All retrieved event data exhibit a high degree of similarity to the input text query, providing further evidence of EXACT's effectiveness.

Event-to-text action retrieval In Fig. 8 (b), we display the Top-3 retrieved text captions queried by the event stream recording action `'Hurdle start'`. The retrieved captions include actions of `'Hurdle start'`, `'Long jump'` and `'Running'`, which share several sub-actions of similar semantics, proving the effectiveness of ExACT.

5. Conclusion and Future Work

We presented the ExACT, as the first exploration utilizing language guidance for event-based action recognition. We proposed a CRUE module to simulate the action recognition of human beings, especially focusing on conceptual reasoning of the temporal relations among event frames and estimating the uncertainty of actions. Besides, we proposed the AFE representation, which adaptively eliminated repetitive events to generate detailed event frames for dynamic actions. To evaluate models' understanding of the complex semantics of actions involving multiple sub-actions with different semantics, we presented the SeAct dataset with semantic-abundant action captions as the first benchmark for event-text action recognition. Our ExACT framework achieved SOTA results on the PAF and HARDVA datasets and achieved plausible performance on our SeAct dataset. Furthermore, we extended ExACT to event-text retrieval tasks, proving its flexible transferability.

Future Work: In the future, we will enhance the conceptual reasoning and uncertainty estimation module in various event vision tasks and experiments on larger event action datasets with semantic-abundant caption-level labels.

Acknowledgement This paper is supported by the National Natural Science Foundation of China (NSF) under Grant No. NSFC22FYT45 and the Guangzhou City, University and Enterprise Joint Fund under Grant No.SL2022A03J01278.

References

[1] Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, et al. A low power, fully event-based gesture recognition system. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7243–7252, 2017. 1, 6

[2] Ben Athiwaratkun and Andrew Gordon Wilson. Multimodal word distributions. *arXiv preprint arXiv:1704.08424*, 2017. 5

[3] Djamila Romaissa Beddiar, Brahim Nini, Mohammad Sabokrou, and Abdenour Hadid. Vision-based human activity recognition: a survey. *Multimedia Tools and Applications*, 79(41-42):30509–30555, 2020. 1

[4] Yin Bi, Aaron Chadha, Alhabib Abbas, Eirina Bourtsoulatze, and Yiannis Andreopoulos. Graph-based spatio-temporal feature learning for neuromorphic vision sensing. *IEEE Transactions on Image Processing*, 29:9084–9098, 2020. 6

[5] Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5710–5719, 2020. 5

[6] Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1):38–56, 2023. 2, 3

[7] Hoonhee Cho, Hyeonseong Kim, Yujeong Chae, and Kuk-Jin Yoon. Label-free event-based object recognition via joint learning with image reconstruction from events. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19866–19877, 2023. 3

[8] Shib Dasgupta, Michael Boratko, Dongxu Zhang, Luke Vilnis, Xiang Li, and Andrew McCallum. Improving local identifiability in probabilistic box embeddings. *Advances in Neural Information Processing Systems*, 33:182–192, 2020. 5

[9] Tristan de Blegiers, Ishan Rajendrakumar Dave, Adeel Yousaf, and Mubarak Shah. Eventtransact: A video transformer-based framework for event-camera based action recognition. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–7. IEEE, 2023. 6

[10] Yongjian Deng, Hao Chen, Hai Liu, and Youfu Li. A voxel graph cnn for object classification with event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1172–1181, 2022. 7

[11] Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of vision-language pre-trained models. *arXiv preprint arXiv:2202.10936*, 2022. 2, 3

[12] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–213, 2020. 6

[13] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 6

[14] Yue Gao, Jiaxuan Lu, Siqi Li, Nan Ma, Shaoyi Du, Yipeng Li, and Qionghai Dai. Action recognition and benchmark using event cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1, 2, 3, 4, 5, 6

[15] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 5

[16] Pengjie Gu, Rong Xiao, Gang Pan, and Huajin Tang. Stca: Spatio-temporal credit assignment with delayed feedback in deep spiking neural networks. In *IJCAI*, pages 1366–1372, 2019. 6

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[18] 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. 6, 7

[19] Yatai Ji, Junjie Wang, Yuan Gong, Lin Zhang, Yanru Zhu, Hongfa Wang, Jiaxing Zhang, Tetsuya Sakai, and Yujiu Yang. Map: Multimodal uncertainty-aware vision-language pre-training model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23262–23271, 2023. 2, 5

[20] Jacques Kaiser, Alexander Friedrich, J Tieck, Daniel Reichard, Arne Roennau, Emre Neftci, and Rüdiger Dillmann. Embodied neuromorphic vision with event-driven random backpropagation. *arXiv preprint arXiv:1904.04805*, 2019. 6

[21] Jacques Kaiser, Hesham Mostafa, and Emre Neftci. Synaptic plasticity dynamics for deep continuous local learning (decolle). *Frontiers in Neuroscience*, 14:515306, 2020. 6

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[23] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 5

[24] Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401, 2022. 1

[25] Athanasios Lentzas and Dimitris Vrakas. Non-intrusive human activity recognition and abnormal behavior detection on elderly people: A review. *Artificial Intelligence Review*, 53 (3):1975–2021, 2020. 1

[26] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7083–7093, 2019. 2, 4, 6

[27] Qianhui Liu, Dong Xing, Huajin Tang, De Ma, and Gang Pan. Event-based action recognition using motion information and spiking neural networks. In *IJCAI*, pages 1743–1749, 2021. 1, 2, 6

[28] Zhaoyang Liu, Limin Wang, Wayne Wu, Chen Qian, and Tong Lu. Tam: Temporal adaptive module for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13708–13718, 2021. 6

[29] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 6

[30] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[31] Jacques Manderscheid, Amos Sironi, Nicolas Bourdis, Davide Migliore, and Vincent Lepetit. Speed invariant time surface for learning to detect corner points with event-based cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10245–10254, 2019. 3

[32] Jean-Matthieu Maro, Sio-Hoi Ieng, and Ryad Benosman. Event-based gesture recognition with dynamic background suppression using smartphone computational capabilities. *Frontiers in neuroscience*, 14:501775, 2020. 6

[33] Christoforos Mavrogiannis, Francesca Baldini, Allan Wang, Dapeng Zhao, Pete Trautman, Aaron Steinfeld, and Jean Oh. Core challenges of social robot navigation: A survey. *ACM Transactions on Human-Robot Interaction*, 12(3):1–39, 2023. 1

[34] 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. 2, 6

[35] Preksha Pareek and Ankit Thakkar. A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. *Artificial Intelligence Review*, 54: 2259–2322, 2021. 1

[36] Chiara Plizzari, Mirco Planamente, Gabriele Goletto, Marco Cannici, Emanuele Gusso, Matteo Matteucci, and Barbara Caputo. E2 (go) motion: Motion augmented event stream for egocentric action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19935–19947, 2022. 1, 2, 5

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[38] 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. 6

[39] Alberto Sabater, Luis Montesano, and Ana C Murillo. Event transformer. a sparse-aware solution for efficient event data processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2677–2686, 2022. 1, 2, 3, 5, 6

[40] Sumit B Shrestha and Garrick Orchard. Slayer: Spike layer error reassignment in time. *Advances in neural information processing systems*, 31, 2018. 6

[41] Yukun Su, Guosheng Lin, and Qingyao Wu. Self-supervised 3d skeleton action representation learning with motion consistency and continuity. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13328–13338, 2021. 2, 5

[42] Jennifer J Sun, Jiaping Zhao, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, and Ting Liu. View-invariant probabilistic embedding for human pose. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 53–70. Springer, 2020. 2, 5

[43] Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. Human action recognition from various data modalities: A review. *IEEE transactions on pattern analysis and machine intelligence*, 2022. 1

[44] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 6

[45] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 6

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[47] Luke Vilnis and Andrew McCallum. Word representations via gaussian embedding. *arXiv preprint arXiv:1412.6623*, 2014. 5

[48] 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. 2, 6

[49] Xiao Wang, Zongzhen Wu, Yao Rong, Lin Zhu, Bo Jiang, Jin Tang, and Yonghong Tian. Sstformer: Bridging spiking neural network and memory support transformer for frame-event based recognition. *arXiv preprint arXiv:2308.04369*, 2023. 1, 2

[50] Zhengwei Wang, Qi She, and Aljosa Smolic. Action-net: Multipath excitation for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13214–13223, 2021. 6

[51] Ziqing Wang, Yuetong Fang, Jiahang Cao, Qiang Zhang, Zhongrui Wang, and Renjing Xu. Masked spiking transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1761–1771, 2023. 6

[52] Ziyi Wu, Xudong Liu, and Igor Gilitschenski. Eventclip: Adapting clip for event-based object recognition. *arXiv preprint arXiv:2306.06354*, 2023. 3, 7

[53] Rong Xiao, Huajin Tang, Yuhao Ma, Rui Yan, and Garrick Orchard. An event-driven categorization model for aer image sensors using multispike encoding and learning. *IEEE transactions on neural networks and learning systems*, 31 (9):3649–3657, 2019. 6

[54] Bochen Xie, Yongjian Deng, Zhanpeng Shao, Hai Liu, and Youfu Li. Vmv-gcn: Volumetric multi-view based graph cnn

for event stream classification. *IEEE Robotics and Automation Letters*, 7(2):1976–1983, 2022. 1, 2

[55] Bochen Xie, Yongjian Deng, Zhanpeng Shao, Hai Liu, Qingsong Xu, and Youfu Li. Event voxel set transformer for spatiotemporal representation learning on event streams. *arXiv preprint arXiv:2303.03856*, 2023. 1, 2, 3, 5

[56] Fang Xu, Lei Yu, Bishan Wang, Wen Yang, Gui-Song Xia, Xu Jia, Zhendong Qiao, and Jianzhuang Liu. Motion deblurring with real events. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2583–2592, 2021. 3

[57] 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 preprint arXiv:2302.08890*, 2023. 3

[58] Jiazhou Zhou, Xu Zheng, Yuanhuiyi Lyu, and Lin Wang. E-clip: Towards label-efficient event-based open-world understanding by clip. *arXiv preprint arXiv:2308.03135*, 2023. 3, 4, 6, 7