

Exploiting Frequency Dynamics for Enhanced Multimodal Event-based Action Recognition

Meiqi Cao¹, Xiangbo Shu^{1*}, Xin Jiang¹, Rui Yan^{1,2}, Yazhou Yao¹, Jinhui Tang³

¹Nanjing University of Science and Technology

²Nanjing University

³Nanjing Forestry University

{cmq123, shuxb, xinjiang, ruiyan, yazhou.yao}@njjust.edu.cn, tangjh@njfu.edu.cn

Abstract

While event cameras excel in capturing microsecond temporal dynamics, they suffer from sparse spatial representations compared to traditional RGB data. Thus, multimodal event-based action recognition approaches aim to synergize complementary strengths by independently extracting and integrating paired RGB-Event features. However, this paradigm inevitably introduces additional data acquisition costs, while eroding the inherent privacy advantages of event-based sensing. Drawing inspiration from event-to-image reconstruction, texture-enriched visual representation directly reconstructed from asynchronous event streams is a promising solution. In response, we propose an Enhanced Multimodal Perceptual (EMP) framework that hierarchically explores multimodal cues (e.g., edges and textures) from raw event streams through two synergistic innovations spanning representation to feature levels. Specifically, we introduce Cross-Modal Frequency Enhancer (CFE) that leverages complementary frequency characteristics between reconstructed frames and stacked frames to refine event representations. Furthermore, to achieve unified feature encoding across modalities, we develop High-Frequency Guided Selector (HGS) for semantic consistency token selection guided by dynamic edge features while suppressing redundant multimodal information interference adaptively. Extensive experiments on four benchmark datasets demonstrate the superior effectiveness of our proposed framework. The code is available at <https://github.com/caomq123/EMP>.

1. Introduction

Traditional frame-based cameras often encounter performance limitations when dealing with rapid motion or complex illumination conditions [22, 25, 42, 43]. In contrast,

*Corresponding author

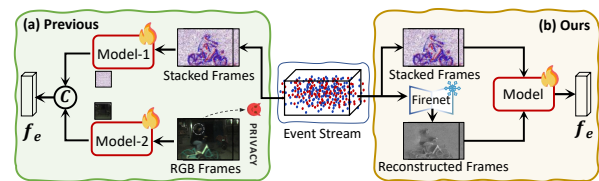


Figure 1. (a) Previous multi-modal input approaches independently extract features from paired RGB-Event modalities, both introducing excessive RGB data dependencies and disrupting the privacy-preserving property unique to event cameras. (b) Our approach takes a lightful reconstruct network (Firenet [40]) to substitute RGB data while narrowing the modality divergence with events, thereby enabling enhanced unified multimodal perception.

event-driven neuromorphic visual sensors, also known as event cameras [13, 15], operate by asynchronously capturing per-pixel intensity changes only when a significant light change occurs. This unique functionality endows event cameras with several prominent advantages, including high dynamic range, low power consumption, and high temporal resolution, which enable them to excel in challenging scenarios such as fast motion, extreme lighting conditions, and privacy-sensitive applications. These attributes have led to their widespread adoption in various computer vision tasks, including object detection and tracking [16, 26, 52], motion deblurring [5, 27], and image reconstruction [31, 40, 53]. Notably, the inherent sensitivity of event cameras to dynamic scenes provides a unique advantage in capturing motion-related actions. Leveraging these strengths, the task of Event-based Action Recognition (EAR) has emerged, aiming to effectively extract action features from event streams to enhance recognition performance.

Currently, the EAR methods faces significant challenges in effectively utilizing the spatially sparse and asynchronous event streams. Traditional single-modal input approaches

adopt two primary strategies: (i) Processing asynchronous event points through point-based encoders (GCNs [11, 54]) or spiking neural networks (SNNs [3, 8, 37]), and (ii) Applying CNN- [32, 44], Transformer- [7, 57] or Mamba- [38, 49] based visual encoders to frames stacked from asynchronous event streams. While point-based methods accommodate the asynchronous nature of events effectively, they exhibit suboptimal recognition performance. Simultaneously, frame-based methods suffer from the lack of critical texture information due to the focus on dynamic edge capture inherent in event cameras.

In recent years, there has been a notable shift toward multi-modal input strategies [9, 24, 48, 50], leveraging diverse data sources to improve feature extraction robustness for practical applications. As depicted in Fig. 1(a), conventional multi-modal methods typically process RGB images and event streams in parallel through independent encoders, followed by late fusion. Although such approaches improve robustness by aggregating complementary features, two critical limitations persist: (i) The incorporation of RGB images not only introduces additional data acquisition costs but also compromises the intrinsic privacy-preserving nature of event cameras by capturing detailed individual visuals. (ii) Independent encoders disregarding inter-modal correlations introduce redundancy that critically undermines the acquisition of discriminative action features.

To address the challenges above, we seek a privacy-preserving alternative to RGB data that eliminates additional data dependency and narrows the representational gap between surrogate modal with event modal. As shown in Fig. 1(b), inspired by the event-to-image [40] techniques, we employ a lightweight event reconstruction network to generate texture-enriched surrogate frames directly from event streams. This approach effectively emulates RGB semantics while filtering privacy-sensitive details. Building upon this, we introduce a unified multi-modal encoding architecture that harmonizes edge-focused event representations with texture-augmented surrogates, preserving modality-specific discriminative strengths while suppressing cross-modal redundant information interference.

Formally, we present an Enhanced Multi-modal Perceptual (EMP) framework designed to hierarchically exploit intrinsic multi-modal cues from the representation level to the feature level. Specifically, our framework consists of two components: Cross-modal Frequency Enhancer (CFE) and High-frequency Guided Selector (HGS). The CFE leverages the distinct frequency traits inherent in different modalities, utilizing the high-frequency information (*e.g.*, edges) from stacked event frames to enhance the reconstructed event frames including mid-frequency information (*e.g.*, texture). This process effectively refines the event representations for better feature extraction. In Addition, considering the discrepancy between modalities,

the HGS adaptively captures semantic consistency tokens across different modalities, thereby realizing unified encoding of multimodal features.

The main contributions are summarized as follows,

- A privacy-preserving paradigm inspired by event-to-image generates texture-enriched surrogate frames as alternatives to RGB images, eliminating auxiliary data acquisition while bridging the modality gap with events.
- A novel Enhanced-Multimodal Perceptual (EMP) framework that hierarchically explores intrinsic multi-modal cues from raw event streams through dual-stage innovations at representation and feature levels.
- A multi-modal unified perception strategy equipped with collaborative Cross-modal Frequency Enhancer (CFE) and High-frequency Guided Selector (HGS), enables progressive exploitation of intrinsic multi-modal cues from the representation level to the feature level.

2. Related Work

2.1. Single-modal Event-based Action Recognition

Existing single-modal approaches for event-based action recognition (EAR) focus on extracting discriminative features from event streams [8, 52]. These methods leverage the sparse spatial nature of event data by stacking asynchronous event streams into compact frames, which are processed using powerful frame-specific architectures, *e.g.* convolutional neural networks (CNNs) [32, 44] or visual transformers (ViTs) [7, 57]. For instance, Event-TransAct [10] introduce video transformers and event contrastive learning to extract fine-grained features from the event stream. Similarly, ExACT [57] incorporate rich semantic information to effectively handle event frames from a cross-modal perspective. Beyond frame-based methods, some works [37, 54] explore networks tailored to the unique properties of asynchronous events, such as spiking neural networks (SNNs) and graph convolutional networks (GCNs). For example, SpikePoint [37] proposed an end-to-end SNN architecture that processes sparse event streams by simultaneously extracting global and local features. VMv-GCN [54] introduced a novel volumetric multi-view framework to capture key spatial and temporal information from event streams. However, these approaches primarily fail to fully exploit the complementary information inherent in raw event streams. In contrast, our proposed Enhanced Multi-modal Perceptual framework synergistically integrates features from different modalities.

2.2. Multi-modal Event-based Action Recognition

Recent advances in robust representation for multi-modal data garner significant attention across various visual domains [7, 48]. For EAR, previous research explore two primary paradigms for multi-modal input: (i) Integrating

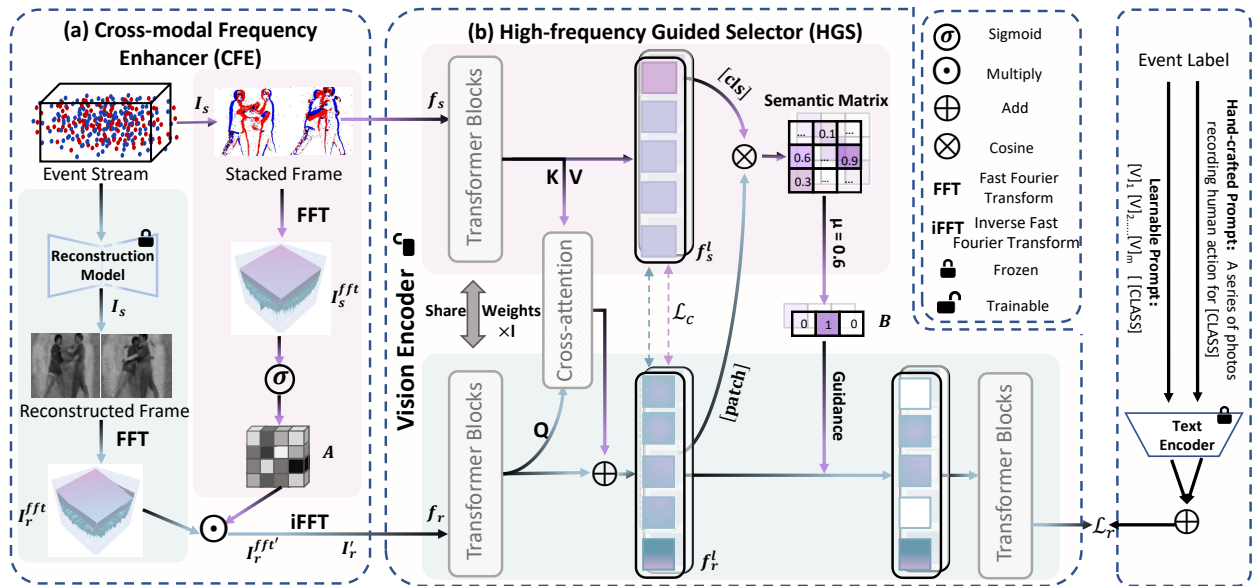


Figure 2. The overall framework of our proposed EMP. First, complementary event modalities (stacked and reconstructed frames) undergo frequency-domain refinement via the Cross-modal Frequency Enhancer (CFE), which amplifies edge dynamics while preserving textural in reconstructed counterparts. After that, during the unified visual encoding process of CLIP, the High-frequency Guided Selector (HGS) is employed to capture semantic-consistency tokens across modalities, which effectively eliminates spatiotemporally redundant features. Finally, event features and text features derived from the text/vision encoder are contrasted for activity recognition.

different representations of event data (*e.g.*, event frames, event points, and event voxels); and (ii) Fusing RGB images with event data. For instance, AGCN [19] introduce a dual point-voxel absorbing graph representation to exploit complementary information between event points and event voxels. EventCrab [6] propose a framework that synergistically processes event frames and event points balancing efficiency and effectiveness. However, while these methods fully leverage the rich temporal information within event streams, they overlook the inherent lack of texture information in events. To address this limitation, SAFE [24] leverages pre-trained large-scale vision-language models to fuse semantic labels, RGB frames, and event streams, improving multi-modal recognition accuracy. Similarly, sst-former [48] formally propose to recognize patterns by fusing RGB frames and event streams simultaneously. Despite these advances, existing methods require additional resources for RGB images and fail to fully exploit the consistency between multi-modal data. In contrast, our proposed framework eliminates dependency on RGB while explicitly addressing the impact of irrelevant information on multi-modal feature selection. Our approach provides a promising solution for practical deployment in real-world scenarios.

2.3. Token Selection in Transformer

With the widespread adoption of Transformer across various domains [6, 20, 21, 34, 35], token selection has gar-

nered increasing attention for its ability to leverage inherent redundancies between image tokens to focus on key objects and reduce computational costs [12, 36, 47, 55]. This is particularly crucial in visual tasks [36], where the extraction of fine-grained features is essential for visual understanding. DynamicViT [18, 36] inserts prediction modules into Transformer blocks to predict and select tokens with higher information content. TransFG [18] directly employs attention mechanisms to locate important tokens, demonstrating strong performance in fine-grained classification. Building on this, STA [12] proposes a framework that simultaneously considers temporal redundancy and semantic importance for selecting spatiotemporal tokens. Magic [55] extends the token selection to multi-modal tasks, enabling the model to dynamically choose object-centric tokens based on different input modalities. Unlike these methods, we propose a more flexible multi-modal token selection scheme that utilizes high-frequency modal-guided strategies to capture semantic consistency tokens, thereby enhancing the robustness of our approach.

3. Method

3.1. Joint Formulation

We introduce a jointly formulated framework for event-to-image reconstruction and CLIP-driven action recognition, without the need for additionally captured RGB images.

Event-to-Image Reconstruction. Considering the event stream $\mathbf{E} = \{(u_i, p_i, t_i)\}_{i \in (0, N)}$, where u_i denotes the pixel position, p_i indicates the polarity, t_i denotes the time of each event point, and N is the number of event points. We first generate two complementary event representations from raw asynchronous event streams: (i) Edge-dominant stacked frames \mathbf{I}_s obtained by temporally aggregating events; (ii) Texture-enriched reconstructed frames \mathbf{I}_r obtained by the image reconstruction network H as:

$$\mathbf{I}_r = H(\mathbf{E}), \quad (1)$$

where $\mathbf{I}_s, \mathbf{I}_r \in \mathbb{R}^{T \times H \times W \times 3}$.

CLIP-driven Action Recognition. In terms of C action classes, we follow [57] to adopt two different text prompts: hand-crafted prompt “A series of photos recording action for [class]” and learnable prompt “[V_1], [V_2], ..., [V_n], [class]”, where [class] represents the class name and [V_n] is the random initialized parameter. The CLIP-based text encoder [56] first encodes the two prompts to obtain D -dimensional text features $\mathbf{f}_1^t \in \mathbb{R}^{C \times D}$ and $\mathbf{f}_2^t \in \mathbb{R}^{C \times D}$, and the final text feature $\mathbf{f}^t \in \mathbb{R}^{C \times D}$ is obtained by averaging \mathbf{f}_1^t and \mathbf{f}_2^t . Meanwhile, we use the CLIP-based vision encoder [56] for event feature extraction, fully exploiting the potential of Visual Language Models (VLMs) in event-based recognition. The stacked frames \mathbf{I}_s and reconstructed frames \mathbf{I}_r after being enhanced for event representation, are fed into the vision encoder to obtain the event feature $\mathbf{f}^e \in \mathbb{R}^{1 \times D}$. Eventually, the class probability $p \in \mathbb{R}^{1 \times C}$ for the action recognition are obtained as,

$$\begin{aligned} p &= \text{softmax}(\mathbf{f}^e (\mathbf{f}^t)^\top), \\ c &= \text{argmax}(p), \end{aligned} \quad (2)$$

where $\text{softmax}(\cdot)$ denotes the softmax function. By taking the category of maximum probability, we obtain the final action prediction class c for the event stream.

3.2. Enhanced-Multimodal Perceptual (EMP)

To preserve diverse information within and across modalities while eliminating the influence of irrelevant information, we propose our Enhanced Multimodal Perceptual (EMP) framework for robust multi-modal unified perception. Fig. 2 presents the architecture of EMP. It is equipped with Cross-modal Frequency Enhancer (CFE) and High-frequency Guided Selector (HGS), enabling progressive exploitation of intrinsic multi-modal cues from the representation level to the feature level.

Cross-modal Frequency Enhancer (CFE). Event frames from distinct modalities demonstrate complementary spectral characteristics [55]. Specifically, stacked frames from asynchronous event streams predominantly focus on capturing high-frequency edge dynamics of transient actions,

whereas reconstructed frames preserve mid-frequency texture inherent to activity semantics. Both modalities are crucial for event action recognition, yet effectively harnessing their complementary strengths remains a critical challenge. The direct combination of these heterogeneous modalities yields suboptimal solutions because of their inherent discrepancies. Thus, we propose a Cross-modal Frequency Enhancer (CFE) that strategically leverages frequency traits across modalities to refine event representation. It enables the model to focus on more comprehensive features through frequency-domain enhancement, thereby optimizing the synergy between modality-specific information.

Whitin the CFE, we first transform stacked frames \mathbf{I}_s and reconstructed frames \mathbf{I}_r into the frequency domain via 3D Fast Fourier Transform (FFT) [29] as:

$$\begin{aligned} \mathbf{I}_s^{\text{fft}} &= \text{FFT}(\mathbf{I}_s), \\ \mathbf{I}_r^{\text{fft}} &= \text{FFT}(\mathbf{I}_r). \end{aligned} \quad (3)$$

The spectral conversion enables the explicit decomposition of modality-specific frequency components, where edge-dominant stacked frames predominantly occupy high-frequency bands while texture-abundant reconstructed frames exhibit concentrated energy in mid-frequency regions. The frequency-aware representation establishes a unified foundation for cross-modal information enhancement. Subsequently, to selectively amplify edge-dynamic information in stacked frames from the frequency spectrum, we apply a sigmoid layer to $\mathbf{I}_s^{\text{fft}}$, generating a frequency filter A . The proposed filtering is calculated as follows:

$$\begin{aligned} A &= \text{sigmoid}(\mathbf{I}_s^{\text{fft}}), \\ \mathbf{I}_r^{\text{fft}'} &= \mathbf{I}_r^{\text{fft}} \odot A, \end{aligned} \quad (4)$$

where \odot denotes element-wise multiplication. Finally, we transfer the $\mathbf{I}_r^{\text{fft}'}$ in the frequency space into the spatial domain to acquire enhanced reconstructed frames \mathbf{I}_r' using an inverse Fast Fourier Transform (iFFT) [29] as:

$$\mathbf{I}_r' = \text{iFFT}(\mathbf{I}_r^{\text{fft}'}). \quad (5)$$

High-frequency Guided Selector (HGS). Given the inherent semantic coherence across modalities derived from the same event stream, existing methods [48] that sequentially extract and aggregate distinct modal features may introduce redundant feature interference. Building on cross-modal frequency enhancement, we further propose a high-frequency guided selection strategy to adaptively prioritize inter-modal critical regions that jointly characterize event semantics. Specifically, both stacked frames and enhanced reconstructed frames are processed through the shared vision encoder for visual feature learning. During event feature encoding, the [cls] token inherently aggregates global semantic representations through self-attention

mechanisms [17]. We leverage high-frequency [cls] token from stacked frames to steer the attention mechanisms of reconstructed frames, enabling selective focus on cross-modal consistent semantic regions while suppressing modality-specific interference.

For stacked frames I_s and enhanced reconstructed frames I_r^l , we first obtain the corresponding non-overlapping patches $f_s, f_r \in \mathbb{R}^{T \times (N+1) \times D}$ including the special [cls] token $f_s^{\text{cls}}, f_r^{\text{cls}} \in \mathbb{R}^{1 \times D}$ [2] with D -dimension through a linear projection, where $N = H \times W / P^2$ and (P, P) is the size of the patch. Between l -th and $(l+1)$ -th layer of the vision encoder, we deploy a cross-attention mechanism to facilitate inter-modal interaction. Here, the reconstructed tokens serve as query feature Q , while the stacked tokens are transformed into key and value features, *i.e.* K and V . The detailed procedure can be formulated as:

$$Q = W_q f_r^l, \quad K = W_k f_s^l, \quad V = W_v f_s^l, \\ f_s^l = \text{CrossAtten}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \quad (6)$$

where $W_q, W_k, W_v \in \mathbb{R}^{D \times D}$ are the weights of the linear transformations corresponding to Q, K, V , respectively.

Subsequently, we extract the [cls] token f_s^{cls} that encapsulates global semantics from the stacked features f_s^l and compute its semantic affinity with the patch tokens in the augmented reconstructed features f_r^l via the cosine similarity. It can be formulated as:

$$\text{sim}(f_s^{\text{cls}}, f_r^l) = \frac{f_s^{\text{cls}\top} f_r^l}{\|f_s^{\text{cls}}\| \|f_r^l\|}. \quad (7)$$

By averaging the aforementioned similarity matrix $\text{sim}(f_s^{\text{cls}}, f_r^l)$ along the temporal dimension T , the final attention score matrix $B \in \mathbb{R}^{T \times N}$ is obtained. To maintain semantic consistency between the two modalities, we use B to incorporate dynamic edge features, which guide the reconstruction branch in adaptively selecting tokens. The selection is formalized as:

$$S = \{\text{Top}_K(B)\}, \quad (8)$$

where Top_K is a function that sorts scores in a set in descending order and outputs the indices corresponding to the K largest values, K is obtained by $\mu \cdot N$ and μ is a selection parameter. Using index set S , we select the corresponding critical tokens from f_r^l and merge them with the f_r^{cls} token before inputting them into the $(l+1)$ -th visual encoder.

3.3. Training Objectives

Modality Consistency Loss. Our model takes the edge-stacked frames and the texture-reconstructed frames as input, which are crucial for capturing the discriminative features of the event stream. The HGS we proposed above focuses on critical token selection guided by high-frequency

modality, based on the principle that different modalities maintain semantic consistency. Therefore, we introduce a loss to reduce the discrepancy between the two features of the same event in the feature space. The semantic consistency cross modalities (f_s/f_r) are constrained by the Mean Square Error (MSE) loss:

$$\mathcal{L}_c = \sum_{i=1}^N \|f_s^i - f_r^i\|_2^2, \quad (9)$$

where N is the number of tokens.

Contractive Recognition Loss. The event-text consistency \mathcal{L}_r is constrained by the contrastive loss between the event feature and the text feature as follows:

$$\mathcal{L}_r = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp((f_i^e)^\top \cdot f_i^t / \tau)}{\sum_{j=1}^B \exp((f_i^e)^\top \cdot f_j^t / \tau)}, \quad (10)$$

where τ is the temperature coefficient.

The final loss is composed of the modality consistency loss \mathcal{L}_c and the contractive recognition loss \mathcal{L}_r as

$$\mathcal{L} = \mathcal{L}_c + \lambda \mathcal{L}_r, \quad (11)$$

where λ denotes the parameter used to balance the loss optimization.

4. Experiments

4.1. Datasets and Settings

Datasets. To evaluate the performance of our approach, we conduct research on three datasets, including PAF [30], SeAct [57], HARDVS [50] and DVS128 Gesture [1]. DVS Action, also known as PAF, contains 10 action categories with 450 recordings, conducted in an unoccupied office environment. SeAct is the first semantic-abundant dataset for event-text action recognition, containing 58 actions under four themes. HARDVS is currently the largest dataset for event-based action recognition, currently having 107,646 recordings for 300 action categories. Both of the above three datasets have an average time duration of 5 seconds with 346×260 resolution. DVS128 Gesture is collected using a DAVIS128 camera with 128×128 resolution, dividing into 11 distinct classes of gestures.

Settings. The implementation of the overall framework is carried out on PyTorch in a Linux environment with two NVIDIA GeForce 4090 GPUs. We use the Adam [23] optimizer with the initial learning rate of 1×10^{-5} and weight decay of 2×10^{-4} . CosineAnnealingLR learning rate schedule is employed with a minimum learning rate of 1×10^{-6} . The representative lightweight FireNet [40] (only 38k parameters) for event-to-image reconstruction, which has low latency (1ms for 240×180).

Method	Event Representation		Top-1 Accuracy (%)		
	Frame-based	Point-based	PAF	SeAct	DVS128 Gesture
Motion SNN [28]	-	✓	78.10	-	92.70
Slayer [41]	-	✓	-	-	93.64
MST [51]	✓	-	88.21	-	-
EV-ACT [14]	✓	-	92.60	-	-
EventTransAct [10]	✓	-	-	57.81	97.92
EvT [39]	✓	-	-	61.30	96.20
SpikePoint [37]	-	✓	90.60	-	98.74
SpikMamba [8]	✓	-	96.28	71.02	99.01
ExACT [57]	✓	-	94.83	67.24	98.86
EventCrab [7]	✓	✓	96.49	72.41	98.80
EventCLIP [56] + Recon	✓	-	96.55	71.05	96.59
Ours	✓	-	99.80(+3.31)	75.00(+2.59)	<u>98.86(-0.15)</u>

Table 1. Comparative performance for EAR on the PAF, SeAct and DVS128 Gesture datasets. The best results are in **bold** and the second-best ones are in underlined.

4.2. Comparison with SOTA Methods

Comparison on Datasets without Paired RGB. As shown in Tab. 1, we compare the proposed method EMP with several representative event-based action recognition methods on three event-only datasets (PAF, SeAct, DVS128 Gesture), where paired RGB data is unavailable. All compared methods exclusively utilize event-driven representations, categorized into event frame-based and event point-based paradigms. As demonstrated in Tab. 1, point-based methods exhibit marginally inferior performance compared to frame-based approaches on the PAF and DVS128 Gesture datasets. This discrepancy stems from the latter’s enhanced capability in extracting discriminative spatiotemporal features through visual architectural designs. EventCrab [7] achieves impressive results by collaboratively mining complementary information from both event points and frames, thereby leveraging multi-modal event representations. However, unlike the paradigm of primarily exploring the sparse spatial characteristics of event streams from EventCrab, our method introduces texture-rich reconstructed frames and holistically exploits intrinsic correlations and extrinsic specificities between event frames and reconstructed frames. This unified perceptual framework yields significantly improved recognition efficacy, attaining 3.31% (Top-1) and 2.59% (Top-1) accuracy gains on the PAF and SeAct datasets respectively, while maintaining competitive performance on the DVS128 Gesture dataset.

Comparison on Datasets with Paired RGB. As shown in Tab. 2, we compare the proposed EMP with several representative EAR methods on the HARDVS dataset, which includes 300-class paired RGB-Event data. Comparative methods are categorized into two types: (i) those utilizing RGB modality and (ii) those solely relying on event streams. We can find that combining RGB frames with

Method	RGB	Param.	Top-1 Accuracy (%)
R2Plus1D [46]	✗	63.5M	49.06
C3D [45]	✗	147.2M	49.94
TimeSformer [4]	✗	121.2M	50.77
ESTF [50]	✗	46.1M	51.22
ExACT [57]	✗	155.1M	<u>90.10</u>
ESTF* [50]	✓	76.8M	49.93
SAFE [24]	✓	-	50.17
C3D* [45]	✓	245.3M	50.88
TimeSformer* [4]	✓	202.3M	51.57
SSTFormer [48]	✓	336.4M	52.97
Ours	✗	182.1M	97.68

Table 2. Comparative performance (%) on the HARDVS datasets. The best results are in **bold**. (·)* denotes the combination between RGB frames and event streams.

event streams yields marginal improvements over the single modality-based version. For instance, C3D [45] achieves an increase from 49.94% to 50.88% (Top-1), while TimeSformer improves from 50.77% to 51.57% (Top-1). However, the significant semantic discrepancy between RGB and event modalities limits performance gains from superficial cross-modal fusion. Different from conventional fusion paradigms, EMP adopts the text-aligned adaptation strategy to bridge the modality gap, while leveraging frequency-domain characteristics to extract discriminative event features. The results in Tab. 2 validate the efficacy of our approach, demonstrating superior robustness in handling heterogeneous modality.

4.3. Ablation Studies

Effectiveness of Each Component. As detailed in Tab. 3, we conduct an ablation study to evaluate the contributions of Cross-modal Frequency Enhancer (CFE) and High-frequency Guided Selector (HGS) within the pro-

CFE	HGS	SeAct (%)	DVS128 Gesture (%)
✗	✗	71.25	97.34
✓	✗	72.41	97.72
✗	✓	73.27	98.10
✓	✓	75.00	98.86

Table 3. Ablation study for effectiveness of each component in the proposed method EMP.

Selection ratio	Top-1 (%)	Top-5 (%)
0.2	70.68	90.51
0.4	71.55	92.24
0.6	75.00	95.68
0.8	73.27	93.96
1.0	72.41	93.10

Table 4. Different selection ratio on the SeAct dataset.

posed EMP. The baseline model trained only on event-reconstructed frames without incorporating the CFE and HGS achieves 71.55% accuracy on SeAct and 97.34% on DVS128 Gesture. When adding the baseline with CFE, we can find the performance improvements of 0.86% on SeAct and 0.38% on DVS128 Gesture, respectively. These performance improvements demonstrate the effectiveness of cross-modal representation enhancement in the frequency domain. Furthermore, the introduction of HGS achieves dynamic token selection across modalities, resulting in a 1.72% improvement on SeAct and a 0.76% improvement on DVS128 Gesture. By integrating all components, our model achieves the optimal performance. These results validate the effectiveness of our EMP in different scenarios.

4.4. Diagnose Studies

Impact of Different Token Selection Ratio. As illustrated in Tab. 4, we investigate the impact of varying selection thresholds μ on the SeAct dataset, where μ governs the proportion of semantically consistent tokens selected across modalities in the HGS framework. The experimental results reveal that the recognition accuracy improves as μ increases, reaching its peak at $\mu = 0.6$, beyond which further threshold escalation correlates with performance degradation. This trend underscores the inherent divergence between modalities, suggesting that indiscriminate cross-modal fusion is suboptimal. Critically, calibrating μ to an appropriate level optimally harnesses intermodal complementarity while suppressing shared redundancies.

Impact of Different Selection Layer. As shown in Tab. 5, we investigate the impact of integrating token selection at distinct layers of the vision encoder on the SeAct dataset, with the optimal selection ratio fixed at $\mu = 0.6$ (Tab. 4). We implement modality-guided token selection across indi-

Insertion Layer	Top-1 (%)	Top-5 (%)
[4]	73.27	93.96
[6]	73.96	92.41
[8]	75.00	95.68
[10]	74.13	93.10

Table 5. Different insertion layer of token selection strategy on the SeAct dataset.

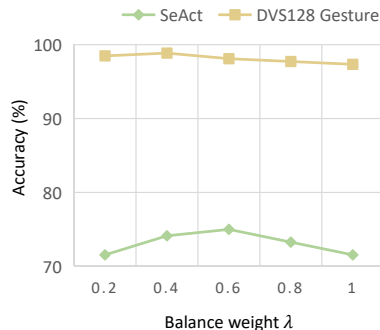


Figure 3. Effect of different values of λ on SeAct and DVS 128 Gesture datasets.

vidual layers and observe that shallow and deep layers yield marginal performance gains. In contrast, layer 8 achieves peak Top-1, validating that appropriately positioned layers within vision encoders optimally encode cross-modal feature representations, thereby enabling effective modality-guided selection of semantically consistent tokens.

Balance Between Consistency Loss and Recognition Loss. Our empirical validation substantiates that achieving equilibrium between modality consistency loss and contractive recognition (Eq. 11) loss constitutes a pivotal design criterion for robust multimodal learning. The contribution of \mathcal{L}_r is modified via λ . As shown in Fig. 3, the variation of λ affects the accuracy performance. The best performance is achieved with weights of 0.6, 0.4 on SeAct and DVS128 Gesture datasets respectively. It is worth noting that our findings highlight the potential benefits of leveraging consistency across modalities to enhance the representational capacity and recognition robustness of our EMP.

4.5. Qualitative Analysis

Visualization of Cross-modal Frequency Enhancer (CFE). To investigate the efficacy of the designed CFE, we visualize the four variants of event frames on the SeAct dataset: raw stacked frames, reconstructed frames, enhanced frames by sharpening operators, and enhanced frames by our CFE. As shown in fig. 4, conventional edge-enhancement techniques [33] yield marginal improvements over raw reconstructed frames. In contrast, event frames processed by CFE consistently highlight the critical regions

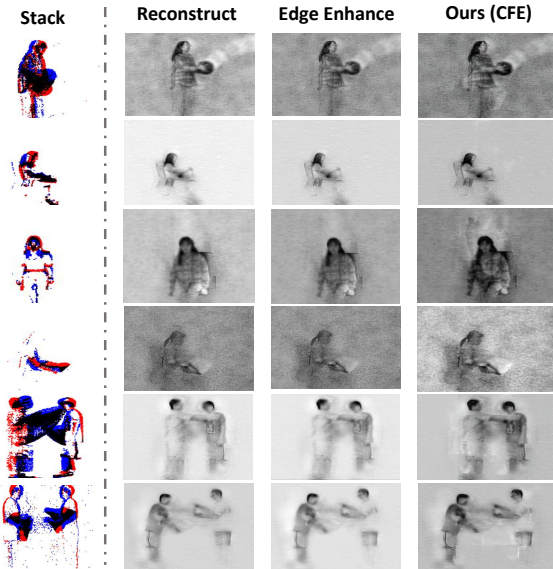


Figure 4. Visualization of event frames. “Stack” and “Reconstruct” denote stacked and reconstructed frames, respectively; “Edge Enhance” indicates enhanced reconstructed frames employing sharpening operators; “CFE” represents enhanced reconstructed frames by our Cross-modal Frequency Enhancer.

across both individual actions and interaction scenarios. Our qualitative analysis confirms that the proposed CFE selectively refines the high-frequency components of the reconstructed frames through spectral coordination with the edge dynamics of the stacked frames, thereby generating complementary event representations with synergistic spatial characteristics.

Visualization of Guidanced Selection. Fig. 5 visualizes the important tokens selected by the High-frequency Guided Selector (HGS). As shown in Fig. 5, the HGS demonstrates the adaptive selection of action-salient tokens across sequential event frames. This demonstrates the intrinsic consistency among multi-modal frames originating from the identical event stream, where HGS effectively identifies semantically coherent tokens while suppressing interference from spatiotemporal redundancies. The results highlight the ability of HGS to focus on critical regions of interest, thereby enhancing the discriminative power of EMP. This selection is crucial for improving the robustness and accuracy of action recognition in dynamic scenes.

Visualization of Event Features Learned by EMP. As illustrated in Fig. 6, we visualize the feature distributions of our proposed EMP against the baseline on the SeAct dataset. Seven action classes are randomly selected to demonstrate inter-class feature separability. Empirical results demonstrate the capability of our EMP in achieving both intra-class compactness and inter-class dispersion, further substantiating its efficacy in preserving semantic dis-

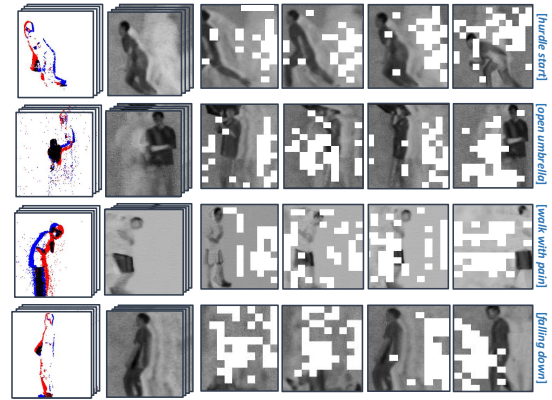


Figure 5. Visualization of cross-modal consistent tokens selected by the HGS, with unselected regions masked in white.

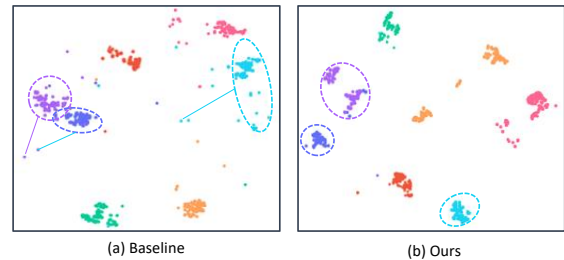


Figure 6. The t-SNE visualization of 7 randomly selected actions. Different color indicates different actions.

tinctiveness across classes.

5. Conclusion and Discussion

Conclusion. In this work, we propose a privacy-preserving paradigm that eliminates RGB dependencies by generating texture-enriched reconstructed frames directly from raw event streams, reducing multi-sensor acquisition costs while bridging the modality gap with event frames. To preserve diverse information across reconstructed and stacked frames while eliminating the influence of irrelevant information, we propose our Enhanced Multimodal Perceptual (EMP) framework for robust multi-modal unified perception. It consists of Cross-modal Frequency Enhancer (CFE) and High-frequency Guided Selector (HGS), which enables the progressive exploitation of intrinsic multi-modal cues from the representation level to the feature level to acquire discriminative event features. Extensive experimental results validate the effectiveness of the proposed method.

Discussion. While our method achieves promising performance in EAR and demonstrates the potential for scalability, its ability to handle more challenging scenarios remains to be explored. Future work will focus on developing a robust multi-modal unified perceptual approach to cope with diverse and variable scenes.

Acknowledgements. The work is supported by the National Natural Science Foundation of China (Grant No. 62222207, 62427808, 62472208), and the Nanjing University Integrated Research Platform of the Ministry of Education-Top Talents Program (Grant No. 2024300482).

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. 5
- [2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021. 5
- [3] Sami Barchid, Benjamin Allaert, Amel Aissaoui, José Mennesson, and Chaabane C Djeraba. Spiking-fer: spiking neural network for facial expression recognition with event cameras. In *Proceedings of the 20th International Conference on Content-based Multimedia Indexing*, pages 1–7, 2023. 2
- [4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning*, page 4, 2021. 6
- [5] Marco Cannici and Davide Scaramuzza. Mitigating motion blur in neural radiance fields with events and frames. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9286–9296, 2024. 1
- [6] Meiqi Cao, Rui Yan, Xiangbo Shu, Jiachao Zhang, Jinpeng Wang, and Guo-Sen Xie. Mup: multi-granularity unified perception for panoramic activity recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7666–7675, 2023. 3
- [7] Meiqi Cao, Xiangbo Shu, Jiachao Zhang, Rui Yan, Zechao Li, and Jinhui Tang. Eventcrab: Harnessing frame and point synergy for event-based action recognition and beyond. *arXiv preprint arXiv:2411.18328*, 2024. 2, 6
- [8] Jiaqi Chen, Yan Yang, Shizhuo Deng, Da Teng, and Liyuan Pan. Spikmamba: When snn meets mamba in event-based human action recognition. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia*, pages 1–8, 2024. 2, 6
- [9] Lan Chen, Haoxiang Yang, Pengpeng Shao, Haoyu Song, Xiao Wang, Zhicheng Zhao, Yaowei Wang, and Yonghong Tian. Velora: A low-rank adaptation approach for efficient rgb-event based recognition. *arXiv preprint arXiv:2412.20064*, 2024. 2
- [10] 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*, pages 1–7, 2023. 2, 6
- [11] Yongjian Deng, Hao Chen, and Youfu Li. A dynamic gcn with cross-representation distillation for event-based learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1492–1500, 2024. 2
- [12] Shuangrui Ding, Peisen Zhao, Xiaopeng Zhang, Rui Qian, Hongkai Xiong, and Qi Tian. Prune spatio-temporal tokens by semantic-aware temporal accumulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16945–16956, 2023. 3
- [13] 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
- [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*, 45(12):14081–14097, 2023. 6
- [15] 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
- [16] Mathias Gehrig and Davide Scaramuzza. Recurrent vision transformers for object detection with event cameras. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 13884–13893, 2023. 1
- [17] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *Advances in Neural Information Processing Systems*, 34:15908–15919, 2021. 5
- [18] Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, and Changhu Wang. Transfg: A transformer architecture for fine-grained recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 852–860, 2022. 3
- [19] Bo Jiang, Chengguo Yuan, Xiao Wang, Zhimin Bao, Lin Zhu, Yonghong Tian, and Jin Tang. Point-voxel absorbing graph representation learning for event stream based recognition. *arXiv preprint arXiv:2306.05239*, 2023. 3
- [20] Xin Jiang, Hao Tang, Junyao Gao, Xiaoyu Du, Shengfeng He, and Zechao Li. Delving into multimodal prompting for fine-grained visual classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2570–2578, 2024. 3
- [21] Xin Jiang, Hao Tang, and Zechao Li. Global meets local: Dual activation hashing network for large-scale fine-grained image retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 2024. 3
- [22] Xin Jiang, Hao Tang, Rui Yan, Jinhui Tang, and Zechao Li. Dvf: Advancing robust and accurate fine-grained image retrieval with retrieval guidelines. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2379–2388, 2024. 1
- [23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5
- [24] 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. *arXiv preprint arXiv:2311.18592*, 2023. 2, 3, 6
- [25] Pengpeng Li, Xiangbo Shu, Chun-Mei Feng, Yifei Feng, Wangmeng Zuo, and Jinhui Tang. Surgical video workflow analysis via visual-language learning. *npj Health Systems*, 2(1):5, 2025. 1
- [26] Bingde Liu, Chang Xu, Wen Yang, Huai Yu, and Lei Yu. Motion robust high-speed light-weighted object detection with event camera. *IEEE Transactions on Instrumentation and Measurement*, 72:1–13, 2023. 1
- [27] Haoyue Liu, Shihan Peng, Lin Zhu, Yi Chang, Hanyu Zhou, and Luxin Yan. Seeing motion at nighttime with an event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25648–25658, 2024. 1
- [28] Qianhui Liu, Dong Xing, Huajin Tang, De Ma, and Gang Pan. Event-based action recognition using motion information and spiking neural networks. In *International Joint Conference on Artificial Intelligence*, pages 1743–1749, 2021. 6
- [29] Shuangbiao Liu, Qian Wang, and Geng Liu. A versatile method of discrete convolution and fft (dc-fft) for contact analyses. *Wear*, 243(1-2):101–111, 2000. 4
- [30] 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. 5
- [31] Federico Paredes-Vallés and Guido CHE De Croon. Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3446–3455, 2021. 1
- [32] 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. 2
- [33] R Pushpavalli and G Sivaradje. Switching median filter for image enhancement. *International Journal of Scientific & Engineering Research*, 3(2):1–5, 2012. 7
- [34] Hongyu Qu, Rui Yan, Xiangbo Shu, Hailiang Gao, Peng Huang, and Guo-Sen Xie. Mvp-shot: Multi-velocity progressive-alignment framework for few-shot action recognition. *arXiv preprint arXiv:2405.02077*, 2024. 3
- [35] Hongyu Qu, Jianan Wei, Xiangbo Shu, and Wenguan Wang. Learning clustering-based prototypes for compositional zero-shot learning. In *Proceedings of the International Conference on Learning Representations*, 2025. 3
- [36] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in Neural Information Processing Systems*, 34:13937–13949, 2021. 3
- [37] Hongwei Ren, Yue Zhou, Yulong Huang, Haotian Fu, Xiaopeng Lin, Jie Song, and Bojun Cheng. Spikepoint: An efficient point-based spiking neural network for event cameras action recognition. *arXiv preprint arXiv:2310.07189*, 2023. 2, 6
- [38] Hongwei Ren, Yue Zhou, Jiadong Zhu, Xiaopeng Lin, Haotian Fu, Yulong Huang, Yuetong Fang, Fei Ma, Hao Yu, and Bojun Cheng. Rethinking efficient and effective point-based networks for event camera classification and regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2
- [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. 6
- [40] 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, 2, 5
- [41] Sumit B Shrestha and Garrick Orchard. Slayer: Spike layer error reassignment in time. *Advances in Neural Information Processing Systems*, 31, 2018. 6
- [42] Xiangbo Shu, Binqian Xu, Liyan Zhang, and Jinhui Tang. Multi-granularity anchor-contrastive representation learning for semi-supervised skeleton-based action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7559–7576, 2022. 1
- [43] Xiangbo Shu, Jiawen Yang, Rui Yan, and Yan Song. Expansion-squeeze-excitation fusion network for elderly activity recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8):5281–5292, 2022. 1
- [44] Zehua Sun, Qihong 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*, 45(3):3200–3225, 2022. 2
- [45] 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
- [46] 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/CVF conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 6
- [47] Hongjie Wang, Bishma Dedhia, and Niraj K Jha. Zero-truncate: Zero-shot token pruning through leveraging of the attention graph in pre-trained transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16070–16079, 2024. 3
- [48] 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. 2, 3, 4, 6
- [49] Xiao Wang, Shiao Wang, Pengpeng Shao, Bo Jiang, Lin Zhu, and Yonghong Tian. Event stream based human action

- recognition: a high-definition benchmark dataset and algorithms. *arXiv preprint arXiv:2408.09764*, 2024. [2](#)
- [50] 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. [2](#), [5](#), [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] Ziming Wang, Ziling Wang, Huaning Li, Lang Qin, Runhao Jiang, De Ma, and Huajin Tang. Eas-snn: End-to-end adaptive sampling and representation for event-based detection with recurrent spiking neural networks. In *European Conference on Computer Vision*, pages 310–328, 2024. [1](#), [2](#)
- [53] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based video reconstruction using transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2563–2572, 2021. [1](#)
- [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. [2](#)
- [55] Pingping Zhang, Yuhao Wang, Yang Liu, Zhengzheng Tu, and Huchuan Lu. Magic tokens: Select diverse tokens for multi-modal object re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17117–17126, 2024. [3](#), [4](#)
- [56] 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 *European Conference on Computer Vision*, pages 477–494, 2024. [4](#), [6](#)
- [57] 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. [2](#), [4](#), [5](#), [6](#)