Exploiting Frequency Dynamics for Enhanced Multimodal Event-based Action Recognition

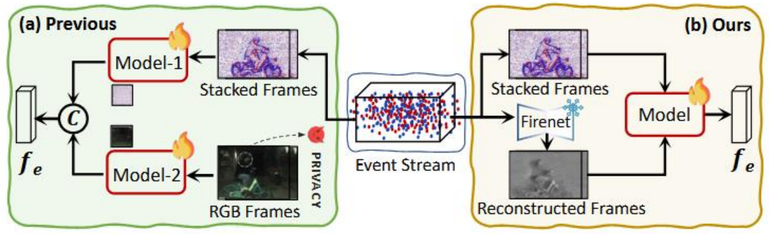
Meiqi Cao[1], Xiangbo Shu[1*], Xin Jiang[1], Rui Yan[1,2], Yazhou Yao[1] and Jinhui Tang[3]
1 Nanjing University of Science and Technology. 2 Nanjing University. 3 Nanjing Forestry University.
Email: {cmq123,shuxb,xinjiang,ruiyan,yazhou.yao}@njust.edu.cn, tangjh@njfu.edu.cn.

ICCV OCT 19-23, 2025 HONOLULU HAWAII

➢ Motivation

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.

Ours: takes a lightful reconstruct network to substitute RGB data while narrowing the modality divergence with events, thereby enabling enhanced unified multimodal perception.



➢ Overview

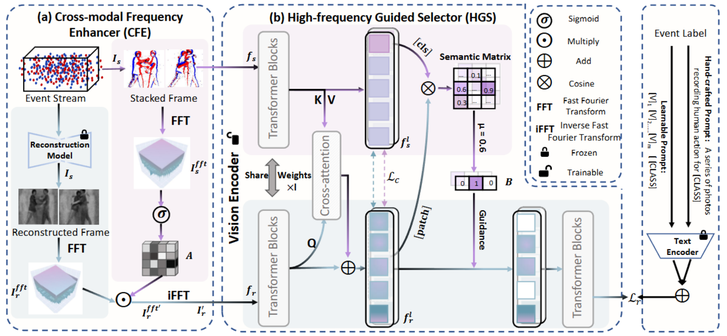
(i) Complementary event modalities (reconstructed and stacked frames) undergo frequency-domain refinement via CFE, which amplifies edge dynamics while preserving textural in reconstructed counterparts;

(ii) During the unified visual encoding process of CLIP, the HGS is employed to capture semantic-consistency tokens across modalities, which effectively eliminates spatiotemporally redundant features;

(iii) Event features and text features derived from the text/vision encoder are contrasted for activity recognition.

➢ Contribution

- 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.
- An Enhanced-Multimodal Perceptual framework that hierarchically explores intrinsic multi-modal cues from raw event streams through dual-stage innovations at representation and feature levels.



- Comparative performance for EAR

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
Ours	✓	-	**99.80** (+3.31)	**75.00** (+2.59)	98.86 (-0.15)

➢ Method

- The class probability p for the action recognition are obtained as:

$$p = \mathrm{softmax}(\boldsymbol{f}^{\mathrm{e}}(\boldsymbol{f}^{\mathrm{t}})^{\top}) \qquad c = \mathrm{argmax}(p)$$

- Cross-modal Frequency Enhancer: strategically leverages frequency traits across modalities to refine event representation:

$$\boldsymbol{I}_{\mathrm{s}}^{\mathrm{fft}} = \mathrm{FFT}(\boldsymbol{I}_{\mathrm{s}}), \quad \boldsymbol{I}_{\mathrm{r}}^{\mathrm{fft}} = \mathrm{FFT}(\boldsymbol{I}_{\mathrm{r}})$$

$$\mathrm{A} = \mathrm{sigmoid}(\boldsymbol{I}_{\mathrm{s}}^{\mathrm{fft}}), \quad \boldsymbol{I}_{\mathrm{r}}^{\mathrm{fft}'} = \boldsymbol{I}_{\mathrm{r}}^{\mathrm{fft}} \odot \mathrm{A}, \quad \boldsymbol{I}_{\mathrm{r}}^{'} = \mathrm{iFFT}(\boldsymbol{I}_{\mathrm{r}}^{\mathrm{fft}'}).$$

- High-frequency Guided Selector: strategically leverages 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:

$$Q = W_{\mathrm{q}}\boldsymbol{f}_{\mathrm{r}}^{l}, \quad K = W_{\mathrm{k}}\boldsymbol{f}_{\mathrm{s}}^{l}, \quad V = W_{\mathrm{v}}\boldsymbol{f}_{\mathrm{s}}^{l},$$

$$\mathrm{sim}(\boldsymbol{f}_{\mathrm{s}}^{\mathrm{cls}}, \boldsymbol{f}_{\mathrm{s}}^{l}) = \frac{\boldsymbol{f}_{\mathrm{s}}^{\mathrm{cls}\top}\boldsymbol{f}_{\mathrm{r}}^{l}}{\|\boldsymbol{f}_{\mathrm{s}}^{\mathrm{cls}}\|\|\boldsymbol{f}_{\mathrm{r}}^{l}\|} \qquad S = \{\mathrm{Top}_K(B)\}$$

- Visualization of cross-modal consistent tokens selected by HGS