DYNAMIC MERGING MUTUALLY-DISTILLED MULTI-VIEW NETWORKS FOR RGB-EVENT-BASED REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-modal fusion algorithms combining conventional frame-based cameras and event cameras have recently attracted substantial research attention. Typically, existing approaches convert event streams into frames, voxels, or tensors, and then learn the required spatiotemporal information from these transformed event representations for fusion with conventional frame-camera data. Obviously, such paradigms only exploit the limited representational capacity of event data. Moreover, standard training pipelines fail to fully leverage diverse model parameters obtained throughout the learning process. In this paper, we propose a novel mutual distillation method leveraging multi-view teacher network weights to better integrate multiple event representations and achieve model-parameter level fusion. Specifically, we employ multiple encoders to extract different types of event representations and perform mutual knowledge distillation at both the feature-level and logit-level in this process. Afterwards, we perform a dynamic hierarchical model merge strategy on the model parameters from the first stage to yield a unified multi-modal RGB-Event representation model. Experimental results across multiple event-based datasets demonstrate that our proposed framework achieves significant performance improvements. Source code of this paper will be released upon acceptance.

1 INTRODUCTION

RGB video offers rich appearance and scene context through regularly sampled frames, forming the basis of many visual recognition systems. However, its fixed temporal sampling can miss rapid motion and fine temporal changes. Event cameras instead record asynchronous brightness changes with high temporal resolution and wide dynamic range, offering complementary motion cues. Combining the two sensing modalities therefore provides a way to associate detailed appearance with temporally precise dynamics.

Event observations can be represented as accumulated event frames, asynchronous event streams, or spatio-temporal voxel grids. Event frames preserve spatial response patterns, streams retain precise timing, and voxels organize local spatial and temporal information. These representations have different structures and statistics, thus, selecting only one view or fusing views without accounting for their differences can discard useful event information. How can we learn a coherent RGB-Event representation while preserving the complementary cues carried by multiple event views?

Learning from multiple event views requires reconciling shared semantics with representation-specific characteristics. Their heterogeneous structures induce different feature distributions, so indiscriminate alignment may weaken the diversity that motivates multi-view learning. Cross-modal distillation offers a means of transferring supervision between modalities (Wang et al., 2021; Huo et al., 2024), while mutual learning allows peer networks to learn collaboratively without a fixed teacher (Zhang et al., 2018). Together, these perspectives motivate knowledge exchange that promotes semantic agreement while respecting the distinct information conveyed by each event view.

Complementarity also extends to the parameters acquired during learning. Models associated with different event views and different stages of optimization can encode distinct aspects of the training observations. Model merging provides a basis for integrating such knowledge through parameter

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

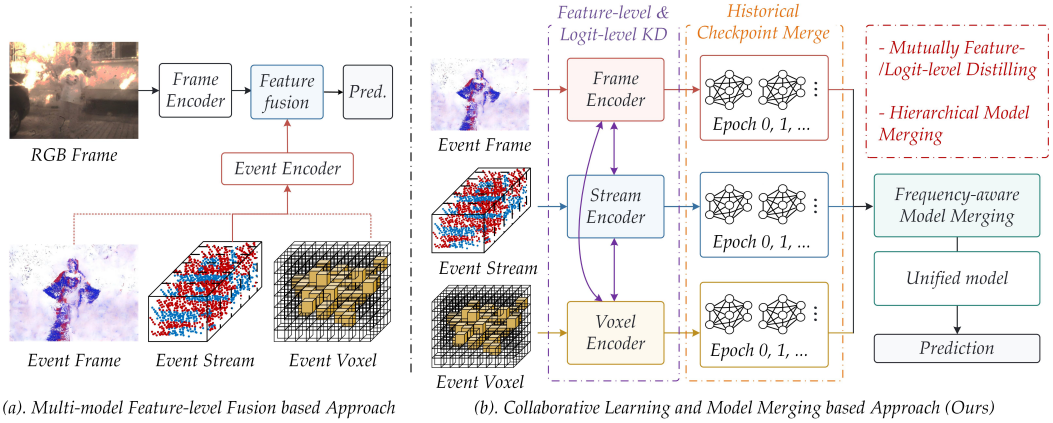


Figure 1: Comparison of RGB-Event representation-learning paradigms. (a) RGB-Event fusion with one event representation. (b) Our multi-view framework combines mutual distillation across event-frame, event-stream, and event-voxel branches with dynamic hierarchical model merging.

updates (Ilharco et al., 2022), with subsequent work addressing interference and the separation of shared and specialized components (Yadav et al., 2023; Lu et al., 2024; Zheng & Wang, 2025). Nevertheless, a common backbone architecture does not ensure that updates learned from heterogeneous event inputs are mutually compatible. The central challenge is to integrate knowledge across event representations and optimization histories without erasing the specialization that makes these sources complementary.

Motivated by this perspective, we propose a mutually distilled multi-view framework for RGB-Event representation learning. As illustrated in Fig. 1, the framework extends single-view learning and conventional cross-modal fusion by considering complementarity in both representation space and parameter space. Collaborative learning allows event-frame, event-stream, and event-voxel branches to exchange knowledge through feature- and logit-level mutual distillation. Dynamic hierarchical merging then integrates historical optimization knowledge with frequency-aware cross-view parameter consolidation. This formulation connects the acquisition of complementary representations with the preservation of shared and view-specific knowledge in a unified model.

To sum up, the main contributions of this paper can be summarized as follows:

1). We propose a novel collaborative multi-view event data representation (i.e., event voxels, frames, and streams) learning framework to tackle monotonous event representations, weak cross-modal feature coupling and inadequate cross-modal information mining. Both feature- and score-level bidirectional mutual distillation strategies are adopted to facilitate mutual supervision and collaborative optimization among the three fusion branches.

2). We propose a dynamic hierarchical model merge strategy for multi-modal and multi-view RGB-Event fusion. This strategy fully exploits model weights from different training epochs and performs frequency-aware model merging using the optimal weights of the three-branch model.

3). Extensive experiments on two widely used RGB-Event based recognition benchmark datasets (i.e., PokerEvent (Wang et al., 2025c), HARDVS (Wang et al., 2024)) fully validated the effectiveness of our proposed dynamic hierarchical model merge strategy for this task.

2 Related Works

Event-based Vision. Event cameras record asynchronous brightness changes with high temporal resolution and wide dynamic range, motivating RGB-Event methods that combine appearance with precise motion cues. Recent methods include hybrid spiking-transformer processing in SSTFormer (Wang et al., 2025c), CNN–Transformer interaction in TSCFormer (Wang et al., 2025b), multimodal heat-conduction fusion in MMHCO-HAR (Wang et al., 2025a), and hardware-oriented event processing in ViP-HMNN (Ji et al., 2025). Most existing methods, however, use a single

fixed event representation, such as frames or voxels. Because event frames, streams, and voxels encode different spatial and temporal information, our method uses parallel view-specific branches and mutual distillation to preserve and integrate their complementarity.

Model Merging. Model merging combines fine-tuned networks into one model without joint retraining. Task Arithmetic (Ilharco et al., 2022) introduced task vectors and weighted parameter addition, while later methods addressed interference through parameter competition (Du et al., 2024), global-basis projection (Liu et al., 2026), isotropic subspace alignment (Marczak et al., 2025), learnable singular-component selection (Lee et al., 2025), separation of shared and exclusive knowledge (Lu et al., 2024), and frequency-domain filtering (Zheng & Wang, 2025). These methods generally target homogeneous model instances. Our approach integrates branches with a shared backbone and heterogeneous event inputs, combining parameter consolidation with mutual distillation for RGB-Event representation learning.

Mutual Distillation. Knowledge distillation transfers supervision by aligning predictions or intermediate representations. Cross-modal approaches include bidirectional image-to-event distillation (Wang et al., 2021), object-centric feature distillation (Li et al., 2024), dynamic bidirectional distillation for modality imbalance (Huo et al., 2024), and frequency-decoupled feature transfer (Liu et al., 2024). Deep Mutual Learning (Zhang et al., 2018) and subsequent online and competitive variants (Guo et al., 2020; Shi et al., 2025) allow peer networks to learn without a fixed teacher. Most such methods use logits or fixed peer aggregation. Our method combines logit-level supervision with sample-adaptive feature distillation, using peer confidence and cross-view similarity to weight the contribution of each event view.

3 METHODOLOGY

3.1 OVERVIEW

As shown in Fig. 2, our framework consists of three sequential components: Multi-View Event Encoding, Collaborative Multi-View Learning, and Dynamic Hierarchical Model Merging. First, the Multi-View Event Encoding module converts the raw event sequence into three temporally aligned representations. These complementary representations capture the spatial, asynchronous, and spatio-temporal characteristics of the same event sequence, and are projected into a unified feature space through view-specific encoding networks. Second, the Collaborative Multi-View Learning module consists of three parallel RGB-Event branches, each taking the RGB frames together with one event representation as input. The three branches are jointly optimized through feature-level and logit-level mutual distillation, enabling complementary knowledge to be exchanged across different event views while preserving their view-specific representations. Finally, the Dynamic Hierarchical Model Merging module hierarchically integrates the knowledge learned by different branches. Specifically, historical checkpoints are first dynamically aggregated through trainable weight averaging to obtain robust branch models, after which the resulting models are merged in the frequency domain to produce a unified model that preserves both shared knowledge and branch-specific expertise.

3.2 MULTI-VIEW EVENT ENCODING

Given a paired RGB-Event video sample, we denote the sampled RGB frames as $\mathbf{R} = \{R_1, \ldots, R_T\} \in \mathbb{R}^{T\times 3\times H\times W}$, where T, H, and W are the temporal length, frame height, and frame width. The event camera records asynchronous events aligned with the RGB frames, where each event is represented as (x_i, y_i, t_i, p_i), and (x_i, y_i), t_i, and p_i denote the spatial coordinate, timestamp, and polarity. The event frame view accumulates events within each sampled RGB interval into a dense frame-like representation:

$$\mathbf{E}_f \in \mathbb{R}^{T\times C_f\times H\times W}, \tag{1}$$

where $C_f = 3$. This view preserves the spatial layout of event responses and can be processed with image-style spatial transformations. The event stream view keeps the asynchronous event form and partitions events according to the sampled RGB intervals:

$$\mathbf{E}_s = \{\mathcal{S}_t\}_{t=1}^{T}, \quad \mathcal{S}_t \quad = \{(x_i, y_i, \tau_i, p_i)\}_{i=1}^{N_t}, \tag{2}$$

where τ_i denotes the timestamp normalized within the t-th interval, and N_t is the number of events in this interval. In implementation, the events in each interval are discretized into 12 temporal

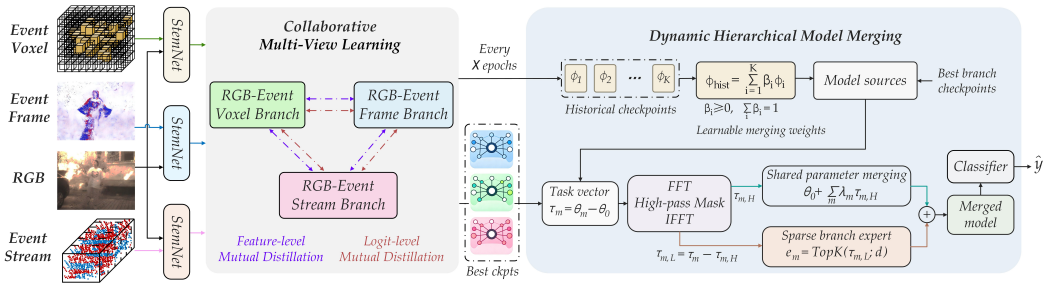


Figure 2: Overview of the proposed RGB-Event representation learning framework. Three RGB-Event branches encode event frames, streams, and voxels and exchange knowledge through feature- and logit-level mutual distillation. Historical checkpoint aggregation and frequency-aware parameter merging then produce a unified inference model.

bins. The event voxel view voxelizes events in the spatial-temporal space. We treat the normalized timestamp as the temporal coordinate and represent each RGB interval as a set of occupied voxel tokens:

$$\mathbf{E}_v = \{\mathcal{V}_t\}_{t=1}^{T}, \quad \mathcal{V}_t \quad = \{[V_x, V_y, V_\tau, \mathbf{V}_f]\} \in \mathbb{R}^{Q_t \times L}, \tag{3}$$

where Q_t is the number of occupied voxels in the t-th interval, (V_x, V_y, V_τ) denotes the voxel center coordinate in the x-y-τ space, and $\mathbf{V}_f$ is the voxel feature aggregated from events falling into the voxel. Each voxel token has $L = 19$ dimensions, consisting of three coordinate values and a 16-dimensional voxel feature. The three representations therefore describe different organizations of the event source rather than a single homogeneous event tensor.

Since the three event views exhibit heterogeneous data structures and feature dimensionalities, they cannot be directly processed by same backbones. To bridge this representation gap, we employ three view-specific stem networks to project the event frame, stream, and voxel into a unified feature space. Formally, the encoded representations are obtained as

$$\mathcal{E}m = \mathrm{SN}_m(E_m), \quad m \in \{f, s, v\}, \tag{4}$$

where $\mathrm{SN}_m(\cdot)$ denotes the stem network corresponding to the m event view, $\{f, s, v\}$ denote the event frame, stream and voxel views, respectively. The resulting representations, denoted as $\mathcal{E}_f$, $\mathcal{E}_s$, and $\mathcal{E}_v$, share a common embedding dimension while preserving the distinctive characteristics of their respective event views. These unified representations are subsequently fed into the collaborative multi-view learning module for joint representation learning.

3.3 Collaborative Multi-View Learning

The Collaborative Multi-View Learning module aims to jointly exploit the complementary information provided by heterogeneous event views while preserving their view-specific characteristics. Specifically, three parallel RGB-Event branches, denoted as Θ_f, Θ_s, and Θ_v, take the RGB frames $\mathbf{R}$ together with the encoded event representations $\mathcal{E}_f$, $\mathcal{E}_s$, and $\mathcal{E}_v$ as input, respectively. Although the three branches share an identical backbone architecture, each branch is independently optimized to learn representations tailored to its corresponding event view. During training, complementary knowledge is exchanged across branches through both logit-level and feature-level mutual distillation. For each view $m \in \{f, s, v\}$, the corresponding model outputs a high-level feature and a class logit:

$$\mathbf{h}_m, \mathbf{z}_m = N_{\Theta_m}(\mathbf{R}, \mathcal{E}_m). \tag{5}$$

$\mathbf{h}_m$ is used for feature-level mutual distillation, and $\mathbf{z}_m$ is used for logit-level mutual distillation and final prediction. At the logit level, each model receives online peer supervision from the other two branches. For each branch Θ_m, we define its peer set $\mathcal{P}_m$ as the collection of the remaining two branches corresponding to the other event views, i.e., $\mathcal{P}_m = \{\, n \mid n \in \{f, s, v\},\, n \neq m \,\}$. With distillation temperature T_d, the peer soft target is:

$$\tilde{\mathbf{p}}_{\mathrm{peer},m}^{T_d} = \mathrm{Softmax}\left(\frac{1}{T_d |\mathcal{P}_m|} \sum_{n \in \mathcal{P}_m} \mathrm{sg}(\mathbf{z}_n) \right), \tag{6}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. The peer soft target is used to supervise the prediction of branch Θ_m through knowledge distillation. The temperature-scaled prediction of the current branch is defined as $\mathbf{p}_m^{T_d} = \mathrm{Softmax}\left(\frac{\mathbf{z}_m}{T_d}\right)$, where the temperature scaling softens the predicted probability distribution to facilitate knowledge transfer. The logit-level mutual distillation loss for branch Θ_m is then formulated as

$$\mathcal{L}_D^m = T_d^2 D_{\mathrm{KL}}\left(\tilde{\mathbf{p}}_{\mathrm{peer},m}^{T_d}, \mathbf{p}_m^{T_d}\right), \tag{7}$$

where $D_{\mathrm{KL}}(\cdot,\cdot)$ denotes the Kullback–Leibler divergence. The factor T_d^2 follows the standard temperature-scaled knowledge distillation formulation to preserve the gradient magnitude during optimization.

At the feature level, each branch feature is projected into a shared semantic space, and a sample-adaptive peer teacher is built from the remaining branches:

$$\mathbf{q}_m = G_m(\mathbf{h}_m),\ \tilde{\mathbf{q}}_m = \sum_{n\in\mathcal{M},\, n\neq m} a_{m,n}\,\mathrm{sg}(\mathbf{q}_n), \tag{8}$$

where $G_m(\cdot)$ is the feature projector of branch m. The adaptive weight $a_{m,n}$ is predicted from the student feature, peer feature, peer confidence, and pairwise feature similarity, where the peer confidence is the maximum class probability of the peer branch and the pairwise similarity is computed in the projected feature space. Therefore, each model learns from a reliability-aware peer feature target instead of a fixed average of the other branches. The reliability-aware peer feature serves as the supervision target for feature-level knowledge distillation. To encourage semantic consistency across different event views, the projected feature of the current branch is aligned with its peer feature target using the cosine similarity loss:

$$\mathcal{L}_F^m = 1 - \mathrm{Cos}\left(\mathbf{q}_m, \tilde{\mathbf{q}}_m\right), \tag{9}$$

where $\mathbf{q}_m$ and $\tilde{\mathbf{q}}_m$ are the projected feature and the corresponding peer feature target, respectively, and $\mathrm{Cos}(\cdot,\cdot)$ denotes the cosine similarity. By maximizing the feature similarity, the proposed feature-level distillation encourages different branches to learn semantically consistent yet complementary representations across heterogeneous event views.

3.4 Dynamic Hierarchical Model Merging

Although collaborative multi-view learning enables the three branches to acquire complementary knowledge from different event representations, the resulting models remain independently optimized and therefore cannot fully exploit the shared knowledge accumulated during training. To obtain a unified and more robust model, we introduce a Dynamic Hierarchical Model Merging module, which progressively consolidates model parameters in a hierarchical manner.

Specifically, the proposed module consists of two stages. The first stage, referred to as dynamic historical model merging, dynamically integrates historical checkpoints within each branch through learnable weight averaging, producing a robust branch model that captures complementary optimization trajectories throughout training. The second stage, termed frequency-aware model merging, further integrates the aggregated models from different branches in the frequency domain, where shared transferable knowledge and branch-specific expertise are selectively preserved to construct the final unified model.

The dynamic historical model merging stage aggregates temporal checkpoints after collaborative multi-view training. It is optimized separately from the preceding Collaborative Multi-View Learning module. The historical merging coefficients are then learned while all checkpoint parameters remain fixed. Given K historical checkpoints $\{\phi_i\}_{i=1}^K$ sampled along the training trajectory, the dynamically aggregated model is defined as:

$$\phi_{\mathrm{hist}} = \sum_{i=1}^{K} \beta_i \phi_i, \quad \sum_{i=1}^{K} \beta_i = 1,\ \beta_i \geq 0. \tag{10}$$

The coefficients $\{\beta_i\}_{i=1}^K$ are trainable parameters.

The frequency-aware model merging stage then integrates the branch models in the parameter space. Specifically, for each branch $m \in \{f, s, v, hist\}$, let θ_0 denote the initial parameters and θ_m denote

the trained parameters of branch m. The task vector is defined as $\tau_m = \theta_m - \theta_0$. Each task vector is reshaped into a two-dimensional parameter map and transformed to the frequency domain. We denote the frequency-domain high-pass mask as $\mathbf{M}_{\text{hp}}$, which suppresses low-frequency components that tend to contain branch-specific interference. Then the inverse transform produces a high-frequency shared component:

$$\begin{aligned} \tau_{m,H} &= \mathcal{F}^{-1}\left(\mathbf{M}_{\text{hp}} \odot \mathcal{F}(\tau_m)\right), \\ \tau_{m,L} &= \tau_m - \tau_{m,H}, \end{aligned} \tag{11}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the two-dimensional Fourier transform and inverse transform, and $\odot$ denotes element-wise multiplication. After the inverse transform, the filtered map is reshaped back to the original parameter-vector form. Specifically, $\tau_{m,H}$ has the same dimensionality as τ_m and represents the high-frequency transferable component of branch m, which is used for shared merging. $\tau_{m,L}$ is the complementary low-frequency residual, which retains branch-specific information for expert extraction. The high-frequency components are merged as shared transferable knowledge:

$$\theta_{\text{merge}} = \theta_0 + \sum_{m \in \mathcal{M}} \lambda_m \tau_{m,H}, \tag{12}$$

where λ_m denotes the merging coefficient estimated from the normalized magnitude of the task vector. We extract a sparse branch expert from $\tau_{m,L}$ by retaining the top-magnitude entries.

$$\begin{aligned} e_m &= \text{TopK}(\tau_{m,L}; d), \\ \theta_{\text{out},m} &= \theta_{\text{merge}} + \mu_m e_m, \end{aligned} \tag{13}$$

where d is the expert density and μ_m is the rescaling factor for the branch expert.

3.5 Optimization Objective

Since the proposed framework is optimized in two sequential stages, the corresponding objective functions are formulated separately for each stage. Specifically, the Collaborative Multi-View Learning module is first optimized to learn complementary representations across heterogeneous event views through supervised classification and mutual distillation. Based on the resulting branch models, the Dynamic Hierarchical Model Merging module is then optimized to progressively aggregate historical checkpoints and merge multi-view model parameters into a unified model. The optimization objectives for the two stages are described as follows.

Optimization of Collaborative Multi-View Learning. Besides two mutual distillation losses $\mathcal{L}_D^m$ and $\mathcal{L}_F^m$, the three RGB-Event branches are trained jointly with supervised classification loss $\mathcal{L}_C^m$ and relation-preserving feature alignment loss $\mathcal{L}_R^m$. Specifically, for branch m, they are defined as

$$\mathcal{L}_C^m = -\sum_{c=1}^{C} y_c \log p_{m,c}, \quad \mathcal{L}_R^m = \left\|\mathbf{A}(\mathbf{q}_m) - \mathbf{A}(\tilde{\mathbf{q}}_m)\right\|_2^2, \tag{14}$$

where $\mathbf{A}(\cdot)$ denotes the normalized mini-batch feature similarity matrix. The overall objective of collaborative multi-view learning is

$$\mathcal{L}_{\text{CML}} = \sum_{m \in \mathcal{M}} \left(\lambda_C \mathcal{L}_C^m + \lambda_D \mathcal{L}_D^m + \lambda_F \mathcal{L}_F^m + \lambda_R \mathcal{L}_R^m\right). \tag{15}$$

In practice, the mutual distillation terms are enabled after warm-up epochs, reducing unstable peer supervision in early training.

Optimization of Dynamic Hierarchical Model Merging. In the second stage, the branch parameters obtained from collaborative multi-view learning are kept fixed. The dynamic historical model merging coefficients are optimized with the supervised classification loss:

$$\mathcal{L}_{\text{DHM}} = \sum_{m \in \mathcal{M}} \omega_m^{\text{hist}} \mathcal{L}_C^m(\phi_{\text{hist}}), \tag{16}$$

where ω_m^{hist} is the branch weight used for coefficient learning. The optimization searches for the coefficient values that minimize the above loss. The frequency-aware model merging stage is then performed in parameter space, without introducing an additional end-to-end training loss.

Table 1: Comparison of recognition performance and computational efficiency on the PokerEvent (Wang et al., 2025c) dataset. Efficiency metrics include the number of parameters, FLOPs, training throughput, and inference latency. "-" indicates that the corresponding result is not reported. The best three Top-1 results are highlighted in **red**, **blue**, and **bold**, respectively.

#No.	Methods	Source	Backbone	Params (M)	FLOPs (G)	Train (clip/s)	Infer. (ms)	Top-1
01	C3D	ICCV 2015	3D CNN	34.8	21.7	12.65	35.69	51.76
02	TSM	ICCV 2019	ResNet50	24.3	65.0	10.31	30.58	55.43
03	X3D	CVPR 2020	ResNet	11.1	48.4	27.29	16.34	51.75
04	TimeSformer	ICML 2021	ViT	121.4	53.6	2.83	114.78	55.69
05	ACTION-Net	CVPR 2021	ResNet50	27.9	17.3	3.30	75.34	54.29
06	MVIT	CVPR 2022	ViT	52.0	10.2	3.35	319.65	55.02
07	V-SwinTrans	CVPR 2022	Swin-Former	88.1	282.0	11.91	214.64	54.17
08	Vision Mamba	ICML 2024	Mamba	25.6	18.1	9.25	35.29	56.50
09	SAFE	PR 2024	VLM	243.2	943.9	2.03	181.09	57.64
10	TSCFormer	MIR 2025	TSCFormer	54.5	34.5	54.39	68.83	**57.70**
11	SSTFormer	TCDS 2025	SNN-Former	336.5	1104.8	2.38	138.74	54.74
12	EFV++	TMM 2025	Former-GNN	160.8	45.6	10.71	24.28	55.40
13	ViP-HMNN	Info. Fusion 2026	ResNet50	-	-	-	90.0	**58.27**
14	MMHCO-HAR	IJCV 2026	Multi-modal HCO	76.1	61.7	-	-	57.40
15	Ours	-	ResNet50	186.2	224.2	9.40	78.37	**58.73**

Overall, the proposed framework consists of three sequential components, where the Multi-View Event Encoding module provides a unified representation interface for heterogeneous event data, while the primary contributions lie in the subsequent Collaborative Multi-View Learning and Dynamic Hierarchical Model Merging modules. The former promotes complementary representation learning through collaborative mutual distillation across multiple event views, whereas the latter hierarchically aggregates both historical optimization knowledge and cross-view model parameters to obtain a unified and more generalizable model. By jointly exploring complementarity at both the representation level and the parameter level, the proposed framework effectively bridges heterogeneous event representations and achieves robust RGB-Event representation learning.

4 Experiments

We evaluate the proposed method on two public RGB-Event recognition benchmarks, PokerEvent (Wang et al., 2025c) and HARDVS (Wang et al., 2024).

4.1 Comparison on Public Benchmark Datasets

• **Results on PokerEvent Dataset (Wang et al., 2025c).** As illustrated in Table 1, we compare the proposed method with recent state-of-the-art (SOTA) approaches on the PokerEvent dataset. Our model achieves the highest Top-1 accuracy of 58.73%, outperforming the previous best method ViP-HMNN (Ji et al., 2025). Compared with other competitive recent methods, such as TSCFormer (Wang et al., 2025b), SAFE (Li et al., 2025), and MMHCO-HAR (Wang et al., 2025a), our approach also shows consistent performance improvements. Notably, these methods cover transformer-based, VLM-based, and multimodal HCO-based frameworks, indicating that our model remains competitive against diverse advanced recognition paradigms. These results demonstrate the effectiveness of the proposed modules and validate their generalization ability for human activity recognition in challenging poker-event scenarios.

• **Results on HARDVS Dataset (Wang et al., 2024).** As shown in Table 2, our method achieves the best overall performance on the HARDVS dataset, with a Top-1 accuracy of 53.28% and a Top-5 accuracy of 63.57%. The strong performance on both metrics demonstrates that the proposed method can effectively integrate complementary information from heterogeneous event representations and generate reliable predictions across multiple candidate categories. Since HARDVS contains diverse human activities captured under varying viewpoints, illumination conditions, and motion speeds,

Table 2: Results on the HARDVS (Wang et al., 2024) dataset. The best three results are highlighted in **red**, **blue** and **bold**.

#No.	Methods	Source	Backbone	Top-1	Top-5
01	C3D	ICCV 2015	3D CNN	50.88	56.51
02	SlowFast	ICCV 2019	ResNet50	46.54	54.76
03	TSM	ICCV 2019	ResNet50	52.58	62.12
04	X3D	CVPR 2020	ResNet	47.38	51.42
05	TimeSformer	ICML 2021	ViT	51.57	58.48
06	ACTION-Net	CVPR 2021	ResNet50	46.85	56.19
07	V-SwinTrans	CVPR 2022	Swin-Former	51.91	59.11
08	ESTF	AAAI 2024	ResNet18	49.93	55.77
09	SCNN-MST	TCDS 2025	SCNN-Former	49.40	55.55
10	TSCFormer	MIR 2025	TSCFormer	**53.04**	**62.67**
11	HoloEv-Net	arXiv 2026	MobileNetV3	52.69	**62.56**
12	EIS-HAR	CVPR 2026	Swin-Former	**53.21**	60.45
13	ViP-HMNN	Info. Fusion 2026	ResNet50	51.88	58.98
14	Ours	-	ResNet50	**53.28**	**63.57**

these results further validate the robustness and generalization ability of the proposed framework in challenging real-world scenarios.

4.2 Ablation Study

To systematically assess the contribution of each component in the proposed method, we performed a comprehensive ablation study on the PokerEvent dataset.

• **Component Analysis.** As shown in Table 3, we evaluate the contributions of the key components, including Collaborative Multi-View Learning (CML) and Dynamic Hierarchical Model Merging (DHMM), to the model accuracy on the PokerEvent dataset. The Baseline denotes the setting where the three event-view branches are trained independently without mutual distillation or model merging, and we report the best-performing branch. Compared with this Baseline, introducing CML improves the Top-1 accuracy from 56.48% to 58.28%, yielding a gain of 1.80%. This improvement indicates that online peer supervision across event frame, event stream, and event voxel views can transfer complementary information and produce more discriminative branch representations than isolated training. After further applying DHMM, the full model reaches 58.73%, bringing an additional gain of 0.45% over CML alone and a total gain of 2.25% over the Baseline. These results suggest that CML and DHMM play complementary roles: CML improves branch learning during training, while DHMM consolidates the transferable parameters after training.

• **Event-View Complementarity Analysis.** As illustrated in Table 3, we evaluate the contribution of each event view on the PokerEvent dataset. When each view is trained and evaluated independently, the event voxel view achieves the best Top-1 accuracy of 56.48%, outperforming the event frame view with 55.81% and the event stream view with 55.54%. This suggests that the voxel representation provides a stronger standalone spatio-temporal encoding, while the frame and stream views still preserve complementary spatial and asynchronous cues. After combining all three event views with CML, the accuracy increases to 58.28%, bringing a 1.80% gain over the strongest single-view result. These results demonstrate that the three event views are complementary and that mutual distillation effectively integrates their knowledge into more discriminative representations.

• **Class-wise Modality Complementarity Analysis.** Fig. 3 visualizes how the three event-modality branches behave across PokerEvent categories. Among the 57 displayed classes, the branches produce different accuracies on 55 classes. Moreover, 28 classes have a unique best-performing branch: the event-frame, event-stream, and event-voxel branches lead on 15, 8, and 5 classes, respectively. This distribution reveals class-dependent complementarity rather than consistent dominance by a single event representation.

Table 3: Ablation study on the PokerEvent dataset. CML and DHMM denote Collaborative Multi-View Learning and Dynamic Hierarchical Model Merging, respectively. DHIS and FA are the dynamic historical model merging and frequency-aware model merging components of DHMM. Symbols "×" and "✓" indicate exclusion and inclusion, respectively.

Configuration	Event View			CML	DHMM		Top-1
	Frame	Stream	Voxel		DHIS	FA	
Frame only	✓	×	×	×	×	×	55.81
Stream only	×	✓	×	×	×	×	55.54
Voxel only	×	×	✓	×	×	×	56.48
CML only	✓	✓	✓	✓	×	×	58.28
CML + DHIS	✓	✓	✓	✓	✓	×	58.43
CML + FA	✓	✓	✓	✓	×	✓	58.62
Ours	✓	✓	✓	✓	✓	✓	**58.73**

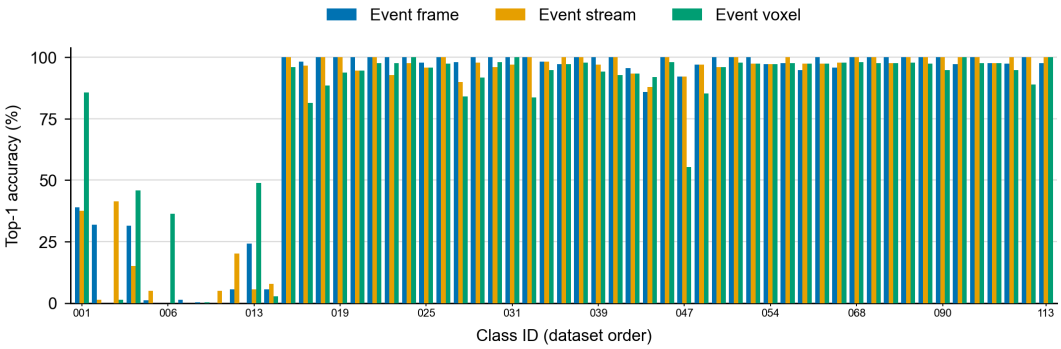


Figure 3: Class-wise Top-1 accuracy of the event-frame, event-stream, and event-voxel branches on PokerEvent in the original dataset order. For clarity, we omit 55 classes scored at 100% by all three branches and two scored at 0%, leaving 57 classes.

4.3 Efficiency Analysis

To evaluate the efficiency benefit of the model merging design, we compare the original inference setting without model merging and the merged model under the same testing configuration.

As shown in Table 4, model merging reduces the total GPU memory footprint from 5642 MB to 3442 MB, corresponding to a 39.0% reduction. More importantly, the peak model allocation decreases from 3970 MB to 1768 MB, achieving a 55.5% reduction. These results indicate that the proposed merging strategy preserves the multi-view knowledge learned from the three event representations while substantially lowering the memory requirement during deployment.

Setting	Total GPU Memory (MB)	Peak Allocation (MB)
Before	5642	3970
After	3442	1768

Figure 4: Memory usage analysis.

5 Conclusion

We presented a multi-view approach to RGB-Event representation learning that considers complementarity in both representation space and parameter space. Mutual distillation promotes knowledge exchange among event-frame, event-stream, and event-voxel representations, while dynamic hierarchical merging integrates historical optimization knowledge and cross-view model parameters. Results on PokerEvent and HARDVS demonstrate the effectiveness of this formulation for fine-grained recognition and human activity recognition. These findings support the joint consideration of collaborative representation learning and parameter consolidation.

AI USE STATEMENT

Generative AI tools (large language models) were employed in this work to assist with manuscript composition and language polishing. We did not use generative AI tools for other research activities, such as research conception, experimental implementation, coding, figure generation, or data analysis. We have carefully reviewed all AI-assisted text and bear full responsibility for the final content of this paper.

REFERENCES

Lan Chen, Dong Li, Xiao Wang, Pengpeng Shao, Wei Zhang, Yaowei Wang, Yonghong Tian, and Jin Tang. Retain, blend, and exchange: A quality-aware spatial-stereo fusion approach for event stream recognition. *IEEE Transactions on Multimedia*, 27:8926–8939, 2025.

MohammadReza Davari and Eugene Belilovsky. Model breadcrumbs: Scaling multi-task model merging with sparse masks. In *European Conference on Computer Vision*, pp. 270–287. Springer, 2024.

Guodong Du, Junlin Lee, Jing Li, Runhua Jiang, Yifei Guo, Shuyang Yu, Hanting Liu, Sim K Goh, Ho-Kin Tang, Daojing He, et al. Parameter competition balancing for model merging. *Advances in Neural Information Processing Systems*, 37:84746–84776, 2024.

Qiushan Guo, Xinjiang Wang, Yichao Wu, Zhipeng Yu, Ding Liang, Xiaolin Hu, and Ping Luo. Online knowledge distillation via collaborative learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11020–11029, 2020.

Chenyu Huang, Peng Ye, Tao Chen, Tong He, Xiangyu Yue, and Wanli Ouyang. Emr-merging: Tuning-free high-performance model merging. *Advances in Neural Information Processing Systems*, 37:122741–122769, 2024.

Fushuo Huo, Wenchao Xu, Jingcai Guo, Haozhao Wang, and Song Guo. C2kd: Bridging the modality gap for cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16006–16015, 2024.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

Xiaoyue Ji, Yifeng Han, Chun Sing Lai, Fangwen Yu, and Zhekang Dong. Vip-hmnn: A visual pathway-inspired hybrid neural network incorporated with in-memory computing for object recognition. *Information Fusion*, pp. 104086, 2025.

Chanhyuk Lee, Jiho Choi, Chanryeol Lee, Donggyun Kim, and Seunghoon Hong. Adarank: Adaptive rank pruning for enhanced model merging. *arXiv preprint arXiv:2503.22178*, 2025.

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. *Pattern Recognition*, 158:111080, 2025.

Lei Li, Alexander Linger, Mario Millhäusler, Vagia Tsiminaki, Yuanyou Li, and Dengxin Dai. Object-centric cross-modal feature distillation for event-based object detection. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 15440–15447. IEEE, 2024.

Junhong Liu, Yuan Zhang, Tao Huang, Haozhao Wang, Wenchao Xu, Shanghang Zhang, and Renyu Yang. Frequency-decoupled cross-modal knowledge distillation. 2024.

Zehao Liu, Kun Li, and Wei Zhou. Globa: Rethinking parameter conflicts in model merging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23954–23962, 2026.

Zhenyi Lu, Chenghao Fan, Wei Wei, Xiaoye Qu, Dangyang Chen, and Yu Cheng. Twin-merging: Dynamic integration of modular expertise in model merging. *Advances in Neural Information Processing Systems*, 37:78905–78935, 2024.

Daniel Marczak, Simone Magistri, Sebastian Cygert, Bartłomiej Twardowski, Andrew D Bagdanov, and Joost Van De Weijer. No task left behind: Isotropic model merging with common and task-specific subspaces. *arXiv preprint arXiv:2502.04959*, 2025.

Daqian Shi, Xiaolei Diao, Xu Chen, and Cédric M John. Competitive distillation: A simple learning strategy for improving visual classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2981–2990, 2025.

Lin Wang, Yujeong Chae, Sung-Hoon Yoon, Tae-Kyun Kim, and Kuk-Jin Yoon. Evdistill: Asynchronous events to end-task learning via bidirectional reconstruction-guided cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 608–619, 2021.

Shiao Wang, Xiao Wang, Bo Jiang, Lin Zhu, Guoqi Li, Yaowei Wang, Yonghong Tian, and Jin Tang. Human activity recognition using rgb-event based sensors: A multi-modal heat conduction model and a benchmark dataset. *arXiv preprint arXiv:2504.05830*, 2025a.

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*, volume 38, pp. 5615–5623, 2024.

Xiao Wang, Yao Rong, Shiao Wang, Yuan Chen, Zhe Wu, Bo Jiang, Yonghong Tian, and Jin Tang. Unleashing the power of cnn and transformer for balanced rgb-event video recognition. *Machine Intelligence Research*, 22(6):1031–1047, 2025b.

Xiao Wang, Yao Rong, Zongzhen Wu, Lin Zhu, Bo Jiang, Jin Tang, and Yonghong Tian. Sstformer: Bridging spiking neural network and memory support transformer for frame-event based recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 2025c.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. *Advances in neural information processing systems*, 36:7093–7115, 2023.

Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. Adamerging: Adaptive model merging for multi-task learning. In *International Conference on Learning Representations*, volume 2024, pp. 22743–22763, 2024.

Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4320–4328, 2018.

Shenghe Zheng and Hongzhi Wang. Free-merging: Fourier transform for efficient model merging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3863–3873, 2025.

A Experimental Setup

A.1 Datasets and Evaluation Metric

• **PokerEvent Dataset** (Wang et al., 2025c) is a dual-modality dataset focusing on the recognition of character patterns in poker cards. It contains 114 classes and includes 27,102 aligned RGB-Event pairs recorded using a DAVIS 346 event camera, which simultaneously outputs RGB frames and event streams. This benchmark requires fine-grained recognition of visually similar poker patterns under dynamic event responses. The dataset is divided into training, validation, and testing subsets, with 16,216, 2,687, and 8,199 samples, respectively.

• **HARDVS Dataset** (Wang et al., 2024) is a large-scale human activity recognition dataset collected using a DAVIS 346 event camera. It encompasses 300 categories of daily human activities, such as drinking, riding a bike, and sitting down, and includes 107,646 RGB-Event video sequences. The dataset covers challenging real-world factors, including multiple viewing angles, illumination conditions, and motion speeds, making it suitable for evaluating robustness in complex activity

scenarios. It is divided into training, validation, and testing subsets, with 64,526, 10,734, and 32,386 samples, respectively.

For evaluation, we follow the commonly used accuracy metrics of each benchmark. On PokerEvent (Wang et al., 2025c), we report **Top-1 accuracy**, which measures whether the highest-confidence prediction matches the ground-truth label. On HARDVS (Wang et al., 2024), we report both **Top-1 accuracy** and **Top-5 accuracy**, where Top-5 accuracy counts a prediction as correct if the ground-truth label appears among the five highest-confidence classes.

A.2 Implementation Details

We implement the proposed framework in PyTorch based on MMAction2 and conduct all experiments on NVIDIA RTX 3090 GPUs. Each RGB-Event branch adopts an ImageNet-pretrained ResNet50-based backbone. During training, the same multi-scale crop and horizontal flip are synchronously applied to the RGB input, while center cropping is used during evaluation. We use SGD with momentum 0.9, weight decay 5×10^{-5}, and gradient clipping with a maximum norm of 20. The distillation temperature is set to $T_d = 4$. The classification and logit-distillation weights are $\lambda_C = 0.6$ and $\lambda_D = 0.1$, respectively, and logit-level mutual distillation is activated from epoch 6. The model is trained for 30 epochs with a batch size of 5 and an initial learning rate of 8×10^{-4}. Feature-level distillation and relation-preserving distillation are activated from epoch 8, with $\lambda_F = 0.15$ and $\lambda_R = 0.15$. For model merging, we merge the RGB-Event fusion and classification-head parameter groups. The frequency cutoff and sparse-expert density are set to 0.005 and 0.01, respectively. We use the checkpoints from epochs $\{5, 10, 15, 20, 25, 30\}$ and optimize their coefficients on the training split.

B Ablation Study

• **Model-Merging Strategy Analysis.** As shown in Table 3, we evaluate the individual and joint contributions of Dynamic Historical model merging (DHIS) and frequency-aware model merging (FA) after CML training. Using CML alone achieves 58.28% Top-1 accuracy. When only DHIS is applied, the result increases to 58.43%, showing that dynamic historical model merging provides a stable complementary solution. When only the FA is used, the result improves to 58.62%, indicating that parameter merging can consolidate transferable knowledge from the multi-view networks. Combining DHIS with the FA achieves the best Top-1 accuracy of 58.73%, demonstrating that the strategy provides complementary benefits.

• **Comparison with Model Merging Methods.** Table 4 compares our method with representative model merging approaches. Weight Averaging yields only 21.17% Top-1 accuracy, suggesting that uniformly averaging heterogeneous branch parameters causes severe interference. Task Arithmetic(Ilharco et al., 2022) and AdaMerging (Yang et al., 2024) achieve Top-1 accuracies from 55.23% to 55.42%. In contrast, our method reaches 58.73%, exceeding the strongest competing result by 3.31%. This consistent advantage supports the effectiveness of hierarchically combining learned historical aggregation with frequency aware cross view merging for heterogeneous event representations.

Table 4: Comparison with different model merging methods on the PokerEvent dataset.

Methods	**Top-1**	**Methods**	**Top-1**
Weight Averaging	21.17	Ties-Merging (Yadav et al., 2023)	55.32
Task Arithmetic(Ilharco et al., 2022)	55.23	Breadcrumbs (Davari & Belilovsky, 2024)	55.40
EMR-Merging (Huang et al., 2024)	54.73	AdaMerging (Yang et al., 2024)	55.42
Twin-Merging (Lu et al., 2024)	55.42	Ours	**58.73**

• **Effect of the Number of Historical Epochs.** Table 5 studies the effect of the number of historical epochs used in DHIS. The model achieves its highest Top-1 accuracy of 58.73% when 5 epochs are used, whereas the other settings remain within the range from 58.43% to 58.53%. In particular, increasing the number of epochs beyond 5 does not lead to further improvement, indicating that simply incorporating more historical checkpoints is not necessarily beneficial because later checkpoints

may contain redundant or interfering optimization information. These results show that DHIS is generally stable across different numbers of historical epochs, while selecting 5 epochs provides the best balance between trajectory diversity and parameter compatibility.

Table 5: Effect of the number of epochs used in DHIS on the PokerEvent dataset.

#Epochs	Top-1	#Epochs	Top-1
2	58.53	6	58.43
4	58.51	8	58.43
5	**58.73**	10	58.46

• **Analysis of Hyperparameters in Model Merge.** Table 6 examines the sensitivity of model merging to the frequency cutoff and sparse expert density. The results remain close across all settings, with Top-1 accuracy staying around 58.6%. The smallest tested frequency cutoff yields the strongest result, suggesting that retaining a broader set of frequency components benefits shared knowledge integration. For sparse expert extraction, a moderate density performs best, while overly sparse or dense experts provide no further benefit. These observations support the selected configuration as an effective balance between shared parameter integration and the preservation of branch specific information.

Table 6: Analysis of the frequency cutoff and sparse expert density in model merging on the PokerEvent dataset.

Frequency Cutoff	Top-1	Expert Density	Top-1
0.005	**58.73**	0.001	58.68
0.050	58.68	0.010	**58.73**
0.200	58.63	0.100	58.63
0.500	58.64	0.500	58.65

• **Analysis of Distilling Strategies in Collaborative Multi-View Learning.** Table 7 evaluates how the three distillation objectives contribute to Collaborative Multi View Learning. Logit distillation provides stronger supervision than feature distillation when each objective is used alone. Adding relation preservation to feature distillation improves the result, which shows that maintaining pairwise feature structure complements direct feature alignment. The complete strategy combines prediction alignment, feature alignment, and relation preservation and achieves the best Top-1 accuracy of 58.28%. This result demonstrates that the three objectives provide complementary supervision at different representation levels and work most effectively when optimized together.

Table 7: Ablation of distillation strategies on the PokerEvent dataset.

Setting	Logit KD	Feature KD	Relation	Top-1
(a)	✓	×	×	56.99
(b)	×	✓	×	56.41
(c)	×	✓	✓	56.92
(d)	✓	✓	✓	**58.28**

C Efficiency Analysis

As shown in Table 1, we compare representative methods in terms of model size, computational cost, training throughput, inference latency, and precision on the PokerEvent dataset. Our method achieves the highest precision of 58.73%, outperforming all compared baselines, including recent RGB-Event and video recognition models such as TSCFormer (Wang et al., 2025b) and EFV++ (Chen et al., 2025). In terms of efficiency, our model requires 186.2M parameters and 224.2G FLOPs, which is higher than lightweight single-backbone methods. This overhead mainly comes from the multi-view design, where event frame, event stream, and event voxel representations

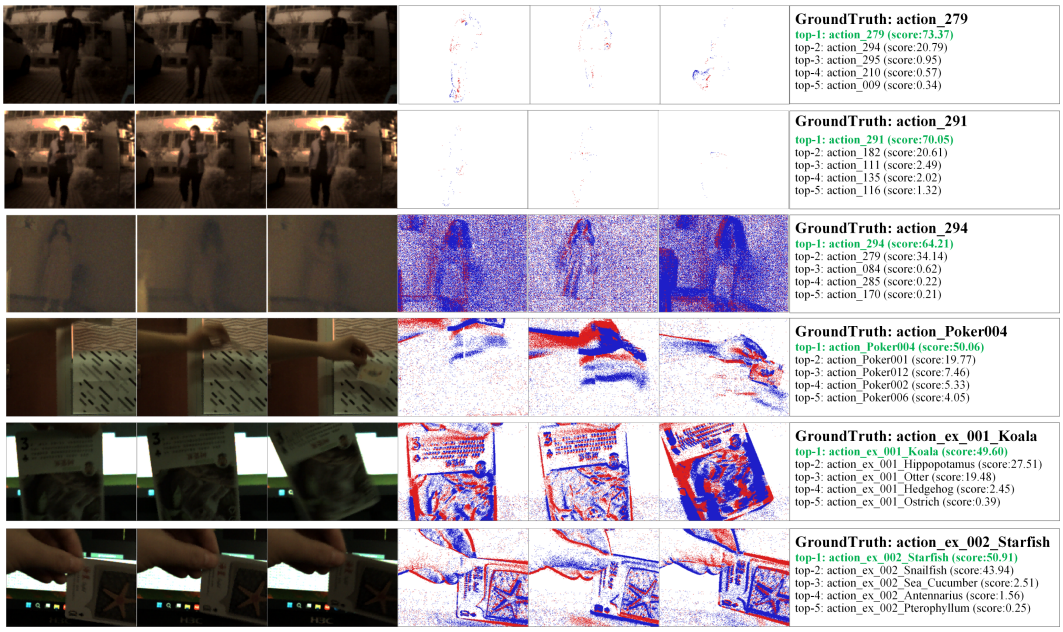


Figure 5: Visualization of Top-5 accuracy on the HARDVS and the PokerEvent dataset.

are jointly exploited to capture complementary spatio-temporal cues. Nevertheless, the proposed model still maintains a practical training throughput of 9.40 clips/s and an inference latency of 78.37 ms. Compared with high-capacity alternatives such as SAFE (Li et al., 2025) and SSTformer (Wang et al., 2025c), our method uses fewer FLOPs, runs faster during both training and inference, and achieves substantially better precision. These results indicate that the proposed mutually-distilled multi-view learning and model merging strategy offers a favorable accuracy-efficiency trade-off: it improves recognition performance by integrating complementary event views, while keeping the additional computational and deployment cost within a reasonable range.

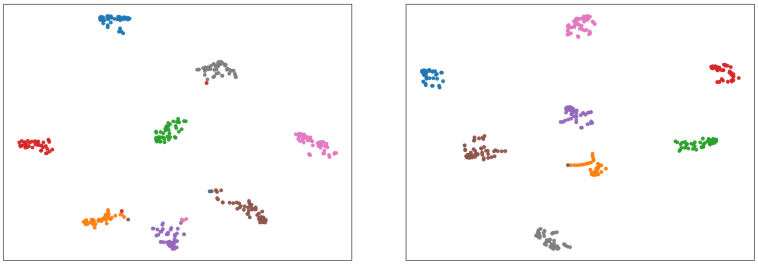
D Limitation Analysis

Although the proposed framework demonstrates the effectiveness of multi-view mutual distillation and model merging for RGB-Event representation learning, there remains room for further exploration. In this work, the distillation process mainly focuses on feature-level and logit-level knowledge exchange among the three event views, while more fine-grained distillation strategies, such as hierarchical feature alignment, uncertainty-aware peer teaching, and relation-aware cross-view transfer, are not fully investigated. Similarly, our model merging design adopts frequency-domain parameter decomposition and historical solution averaging, but more advanced merging mechanisms that explicitly model parameter conflicts, layer-wise reliability, or sample-adaptive merging coefficients could further improve the integration of multi-view knowledge.

E Visualization

• **Top-5 Predictions on HARDVS and PokerEvent.** As shown in Fig. 5, we visualize the top-5 predictions for six representative samples from the HARDVS and PokerEvent datasets. The HARDVS samples correspond to action_279, action_291, and action_294, while the PokerEvent samples correspond to Poker004, Koala, and Starfish. For all six samples, the ground truth category is ranked first. The results show that our method can reliably recognize human activities under challenging motion and illumination conditions and distinguish visually similar poker categories by effectively integrating complementary RGB and event information.

• **Distribution.** As shown in Fig. 6, we visualize the feature embeddings of the Baseline and our proposed method on the PokerEvent dataset, with different colors representing different test classes.



(a). Baseline *(b). Ours*

Figure 6: The feature distribution of Baseline (left sub-figure) and our newly proposed model (right sub-figure) on the PokerEvent dataset using T-SNE.

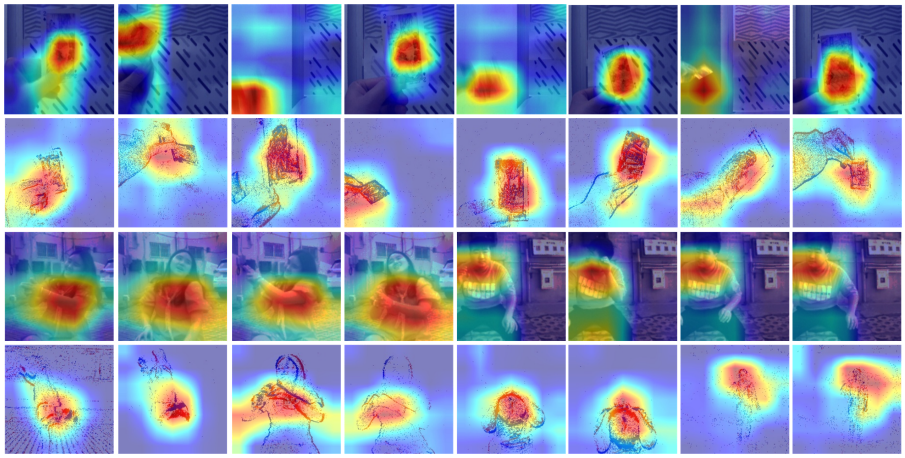


Figure 7: Visualization of feature maps obtained using our proposed method on the PokerEvent and HARDVS dataset.

Compared with the Baseline, which exhibits scattered feature distributions with weak intra-class compactness and limited inter-class separation, our method produces a more discriminative embedding space with compact class-wise clusters and clearer decision boundaries. These results indicate that the proposed module enhances feature discriminability, providing qualitative evidence for the performance improvements reported in our experiments.

• **Feature Maps.** As shown in Fig. 7, we visualize the feature activation maps produced by the proposed method on representative samples from the PokerEvent and HARDVS datasets. For PokerEvent, the model consistently concentrates its responses on the card patterns and hand–object interaction regions, which contain the most discriminative cues for fine-grained category recognition. For HARDVS, the activated regions mainly cover the human body and motion-related areas, while irrelevant background responses are effectively suppressed. Moreover, the attention remains spatially consistent across different temporal samples and event representations. These qualitative results demonstrate that the proposed method can effectively integrate complementary appearance and motion information from RGB frames and heterogeneous event views, thereby learning more discriminative and robust representations for both fine-grained pattern recognition and human activity recognition.