SymMotion: Symmetric Motion Fusion with Wi-Fi Doppler and Visual Optical Flow for Robust Action Recognition

Anonymous ICME submission

Abstract—**Multi-modal Human Action Recognition (HAR) aims to leverage the complementary strengths of vision and wireless sensing. However, existing fusion paradigms suffer from physical inconsistency, often forcing the alignment of static visual textures with dynamic wireless fluctuations, or combining asymmetric representations that lead to modality dominance. To remedy this, we propose SymMotion, a novel framework based on the symmetric motion fusion paradigm. Our core insight is that WiFi Doppler shifts and visual optical flow are physically homologous: both are pure motion derivatives that naturally filter out environmental noise. We proposed the Feature-level Symmetric Gating (FSG) module that purifies modality-specific motion features by suppressing visual background clutter and wireless static components. We also introduced a Conceptual Fusion Decoder (CFD) that utilizes learnable queries to actively decode high-level motion semantics from heterogeneous streams. Extensive experiments on the MM-Fi dataset demonstrate that SymMotion achieves a SOTA accuracy of 98.44%, outperforming existing asymmetric baselines. In particular, the system exhibits exceptional robustness, maintaining 87.26% accuracy even in complete darkness by adaptively exploiting the wireless stream.**

Index Terms—**Multi-modal Human Action Recognition, Symmetric Motion Fusion, WiFi-Vision Fusion, Physical Homology, Robust Sensing**

I. Introduction

Human Action Recognition (HAR) is critical for ubiquitous computing, underlying applications ranging from smart surveillance [1] and human computer interaction [2] to healthcare care [3]. Although single-modal approaches using RGB [4] or wireless signals [5] have progressed, they suffer from inherent limitations: vision is sensitive to illumination and occlusion, while wireless sensing lacks fine-grained spatial semantics. Consequently, multi-modal fusion has emerged as a key solution to leverage domain-specific strengths.

However, existing fusion paradigms often lack physical consistency. We identify two primary limitations: (1) Static-Dynamic Mismatch: Combining CSI amplitude with RGB images [6] inefficiently mixes static textures with dynamic motion features; (2) Asymmetric Alignment: Combining Doppler spectrograms with raw RGB [7] forces the network to reconcile wireless multipath effects with visual background clutter. These asymmetries hinder effective feature alignment and fail to take advantage of the true physical complementarity of the modalities.

To address these challenges, we propose a novel symmetric motion fusion paradigm. Our core insight is that human action induces homologous perturbations across physical domains: pixel displacement (Optical Flow) in vision and channel fluctuations (Doppler) in wireless signals. Both are pure motion derivatives that naturally filter out static environmental noise, ensuring physical consistency and enhancing privacy compared to raw RGB.

Building on this insight, we present SymMotion, a framework designed to maximize synergy between these homologous modalities. It incorporates a Feature-level Symmetric Gating (FSG) module, which applies spatial attention to filter visual background noise and channel attention to suppress static wireless components. Furthermore, we introduce a Conceptual Fusion Decoder (CFD) that utilizes learnable queries to actively aggregate high-level motion semantics, bridging the gap between heterogeneous streams more effectively than simple concatenation.

We rigorously evaluated SymMotion in the MM-Fi dataset [8]. Robustness tests demonstrate clear complementarity: the Doppler stream ensures stability in low-light conditions where vision fails, while Optical Flow compensates for the semantic sparsity of wireless signals.

The main contributions of this paper are:

1) We propose the first symmetric fusion paradigm that theoretically aligns WiFi Doppler and Visual Optical Flow based on physical homology, establishing a robust foundation for multi-modal HAR.
2) We design the FSG module for modality-specific feature purification and the CFD module to resolve heterogeneous misalignment through query-based decoding.
3) Extensive experiments demonstrate that our method outperforms existing single-modal and asymmetric fusion baselines, exhibiting superior robustness against environmental variations.

II. Related Work

A. Visual HAR

HAR has evolved from handcrafted descriptors to deep learning paradigms, where optical flow has emerged as a critical modality for capturing temporal dynamics [9], [10]. Although early two-stream networks [11], [12] established the importance of combining RGB appearance with motion signals, recent studies [13] confirm that optical flow provides superior generalization by being invariant to appearance factors such as clothing and background textures. This appearance invariance makes optical flow a more robust physical descriptor of action than raw RGB frames.

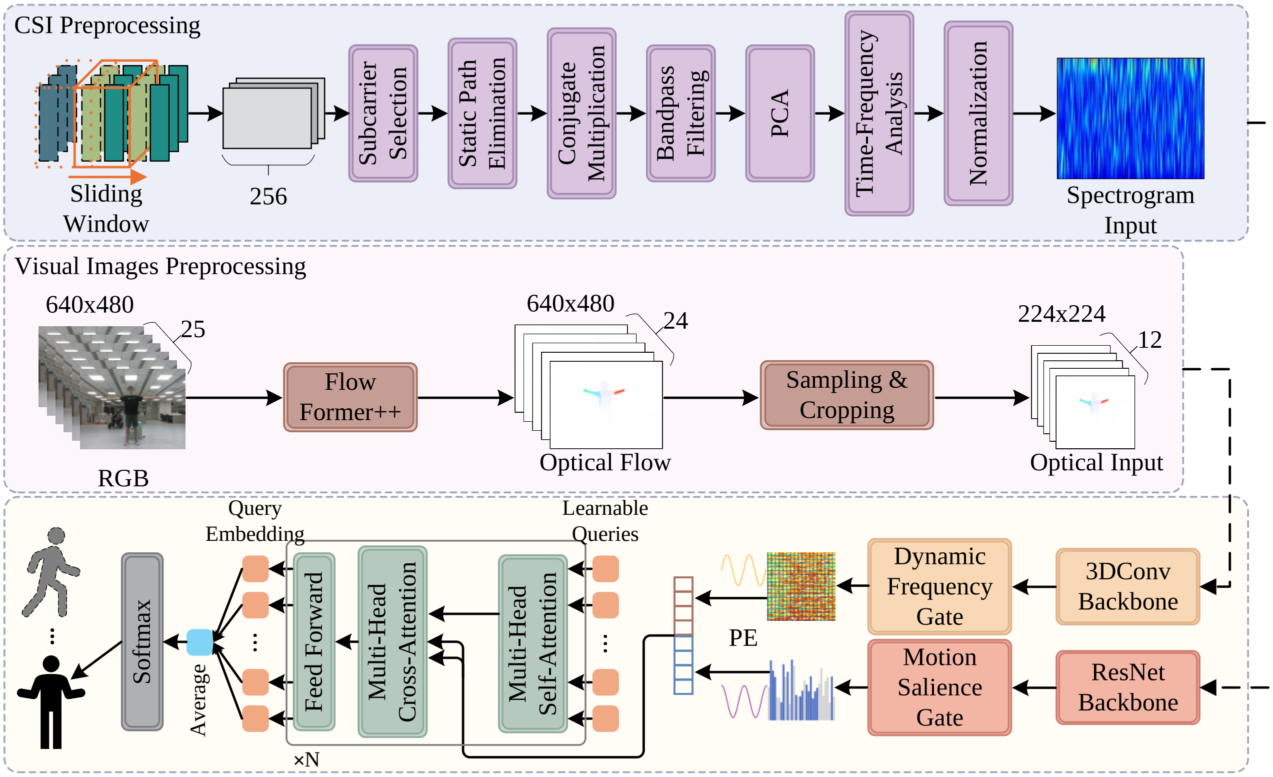

Fig. 1. Overview framework of SymMotion.

However, reliance on optical imaging introduces two fundamental vulnerabilities.

First, standard optical flow estimation relies on the brightness constancy assumption, which collapses in low-light or pitch-black environments. Although recent work like CED-Flow++ [14] performs well in dark environments, DarkLight Networks [15] and specific low-light flow estimators [16] attempt to mitigate this through image enhancement or synthetic training, they effectively hallucinate motion from low-SNR signals rather than measuring it directly.

Second, high-resolution RGB feeds inherently leak sensitive biometric identity information such as face and gait. Although privacy-preserving techniques such as lens-less coded apertures [17] or adversarial anonymization [18] have been proposed, they often degrade recognition performance, creating a difficult trade-off between privacy and utility.

B. Wireless HAR

In parallel to vision, WiFi-based sensing has transformed from the analysis of coarse received signal strength to fine-grained CSI. A paradigm shift occurred with the introduction of doppler-based methods [5], [19], [20], which transform raw CSI variations into body velocity profiles or doppler spectrograms. Unlike CSI in the time-domain that is environment-dependent, doppler signatures capture the radial velocity of the limbs, offering a physical representation of motion that is theoretically analogous to visual optical flow, but immune to lighting conditions [21].

Despite these advances, WiFi sensing faces the problem of semantic gap, while it excels at capturing kinematic dynamics, it lacks the spatial resolution to capture context, such as holding a cup or holding a phone, or multi-user spatial semantics effectively. This limitation necessitates the integration of visual modalities to provide semantic grounding, yet current fusion approaches remain rudimentary.

C. Multimodal Method

Existing RGB-WiFi fusion frameworks largely ignore the physical properties of the signals, leading to what we define as the physical inconsistency problem. We categorize current research into two types of asymmetry.

Static-Dynamic mismatch. A common practice is to fuse static RGB frames (static) with CSI time-series (motion). For example, some approaches convert CSI into image-like heat maps to feed into 2D-CNNs alongside RGB images [22], [23]. We argue that this fusion is physically misaligned, as RGB frames capture position of pixels while CSI captures velocity or change. Forcing a network to correlate static texture with dynamic fluctuations creates a difficult optimization landscape and fails to exploit the temporal continuity of human action.

Asymmetric Representation. More recent work fuse video streams with CSI [24]–[26]. However, these methods often employ asymmetric backbones where the visual stream dominates the inference, causing the model to treat WiFi as a redundant auxiliary rather than a complementary sensor. Consequently, in adverse visual conditions, the dominant visual branch collapses and the uncalibrated WiFi branch fails to compensate effectively. Furthermore, direct fusion of high-dimensional video features with low-dimensional CSI often leads to modality imbalance, where the stronger modality suppresses the weaker one during gradient updates.

III. Model and Method

We propose SymMotion, a multi-modal framework designed to capture the physical homology of human motion through WiFi and visual signals. As illustrated in Fig. 1, our system comprises three main stages.

Data preprocessing and representation stage. We transform raw CSI signals into doppler spectrograms and RGB frames into optical flow sequences to isolate pure motion information from both domains.

Dual-stream feature extraction with symmetric gating and feature extraction stage. A dual-stream network extracts high-level features. Crucially, an FSG Module is inserted to filter modality-specific noise—spatial clutter in vision and static frequency components in wireless signals.

Fusion and classification stage. Instead of passive concatenation, we employ a CFD module with learnable queries to actively decode motion semantics from the purified multi-modal features for final classification.

A. Data Preprocessing & Representation

We utilize the MM-Fi [8] dataset, a large-scale multi-modal dataset. However, the raw data structure (segmented into fragments of $3 \times 114 \times 10$) is insufficient for continuous doppler analysis. We introduce a data reassemble and sliding window strategy to construct robust motion representations.

To capture continuous motion patterns, we first concatenate fragmented raw files within each action instance to form a continuous CSI stream of shape $3 \times 114 \times 2970$. We then apply a 256-sliding window with a specific stride to generate samples $X_{csi} \in \mathbb{C}^{3 \times 114 \times 256}$. Following the pipeline established in Widar3.0 [5], the doppler spectrogram is generated through the following steps.

First, we select a reference subcarrier for each antenna by identifying the one with the highest mean-to-variance ratio, which ensures optimal signal stability for subsequent processing. Second, we remove the static multipath components by subtracting the minimum amplitude from each subcarrier while preserving the phase information, effectively isolating the dynamic components caused by human motion. Third, we perform conjugate multiplication between the adjusted signal and the reference signal to extract phase differences, which simultaneously cancel the carrier frequency offset. Fourth, we apply a butterworth bandpass filter to eliminate DC components associated with stationary objects and suppress high-frequency noise. Fifth, we employ Principal Component Analysis (PCA) to reduce the multi-subcarrier signal into a single representative time series by projecting onto the first principal component. Finally, we applied STFT with a gaussian window to convert the processed signal into a time-frequency representation, yielding the Doppler spectrum that captures the velocity profile of human movements over time.

To ensure temporal alignment, the sliding window of 256 CSI samples corresponds to approximately 25 video frames. We employ the state-of-the-art FlowFormer++ model [27] to extract dense optical flow from these RGB frames. Directly using all 24 flow maps creates excessive computational redundancy. Therefore, we adopt a uniform sampling strategy, selecting $T_v = 12$ frames from the generated flow sequence. These frames are downsampling and cropping to 224×224.

B. Dual-Stream Feature Extraction

We design a two-stream architecture to extract features from heterogeneous inputs. For the wireless stream, we employ a modified ResNet-18 model. The first convolutional layer is adapted to accept single-channel spectrogram input. This backbone extracts the pattern of motion in the frequency-domain, producing a feature map $F_w \in \mathbb{R}^{C_w \times H_w \times W_w}$. For the visual Backbone. Given that the input is a sequence of optical flows, 2D convolution is insufficient to capture temporal evolution. We utilized a 5-layer 3D-CNN. The 3D convolutions $(C \times T \times H \times W)$ effectively model the spatiotemporal continuity of the pixel displacements. This produces the visual feature map $F_v \in \mathbb{R}^{C_v \times T_v' \times H_v \times W_v}$.

C. Feature-level Symmetric Gating Module

To ensure that only high-quality motion semantics enter the fusion stage, we design a gating module that respects the physical characteristics of each modality. More importantly, it ensures stable and high-quality feature selection when issues arise in the feature stream of a particular stream.

Dynamic frequency gating for the visual stream. In the optical flow domain, motion is spatially sparse; meaningful information is confined to the human body, while the majority of the background flow is zero or noise. To suppress this background noise, we apply a spatial attention mechanism for dynamic frequency gating:

$$M_s(F_v) = \sigma(Conv_{2D}([AvgPool(F_v); MaxPool(F_v)])) \tag{1}$$

$$F_v' = F_v \odot M_s(F_v) \tag{2}$$

Here, σ denotes the sigmoid function. This step not only enhances the key parts of the motion features, but also provides a switch for feature quality control.

Motion salience gating for the wireless stream. In the doppler domain, motion information is distributed across specific frequency bands corresponding to the limb speeds. The zero-frequency or static components often dominate the energy, but do not carry motion information. We employ a channel attention mechanism inspired by SE-Block to adaptively recalibrate channel importance:

$$M_c(F_w) = \sigma(MLP(AvgPool(F_w))) \tag{3}$$

$$F'_w = F_w \odot M_c(F_w) \tag{4}$$

This mechanism allows the network to focus on the frequency components that are actively modulated by human motion, effectively filtering out static environmental noise.

D. Conceptual Fusion Decoder

Traditional fusion methods, such as concatenation or summation, passively merge features. We find that effective fusion should be an active process of searching for motion patterns. We propose the CFD, a Transformer-based decoder architecture.

We define a set of learnable parameters $Q \in \mathbb{R}^{K \times d}$, where $K = 32$. These queries represent latent motion concepts such as the abstract notion of "raising an arm" that the model learns to identify during training.

The purified features F'_v and F'_w are flattened, projected, and concatenated to form a unified Key-Value memory bank M_{kv}. The queries Q interact with M_{kv} via multi-head cross-attention:

$$Attention(Q, M_{kv}) = Softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{5}$$

where K and V are projections of M_{kv}. This process allows queries to dynamically attend to the most relevant spatial regions (from Vision) and frequency components (from WiFi) simultaneously, reconstructing a comprehensive representation of the action, which is then fed into a generic MLP head for classification.

E. Training Objective

To ensure robust convergence and enforce the learning of discriminative motion representations across both modalities, we employ a multi-task joint training strategy consisting of three components.

The primary objective is to minimize the standard cross-entropy loss between the final prediction of the fusion network and the ground truth labels:

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(p_i) \tag{6}$$

where N is the batch size, y_i is the ground truth vector, and p_i is the predicted probability distribution of the fusion decoder.

To alleviate the modality laziness problem—where the model might over-rely on the dominant modality (e.g., Optical Flow) and ignore the other—and to ensure the effectiveness of our FSG Module, we introduce deep supervision. We attach two auxiliary classifiers h_v and h_w directly after the gated features F'_v and F'_w. This requires each stream to learn semantically meaningful features independently before fusion:

$$\mathcal{L}_{aux} = \mathcal{L}_{CE}(h_v(F'v), y) + \mathcal{L}_{CE}(h_w(F'_w), y) \tag{7}$$

This explicitly constrains the gating modules to filter noise and retain class-discriminative motion cues.

To prevent queries from collapsing into similar representations in CFD, we apply an orthogonality constraint to maximize their diversity:

$$\mathcal{L}_{div} = ||\frac{Q}{|Q|_2}(\frac{Q}{|Q|_2})^T - I||_F \tag{8}$$

where I is the identity matrix and $\| \cdot \|_F$ denotes the Frobenius norm. This encourages each query to attend to distinct spatiotemporal or frequency patterns.

The final objective function is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_{aux}\mathcal{L}_{aux} + \lambda_{div}\mathcal{L}_{div} \tag{9}$$

where λ_{aux} and λ_{div} are balance hyperparameters, set at 0.5 and 0.1, respectively.

IV. Experiments and results

A. Experimental Setup

We evaluate SymMotion on the MM-Fi dataset. This dataset contains 27 action categories from 40 subjects. We utilize Wi-Fi CSI collected via a TP-Link 750 wireless router and RGB images collected by the Intel D435 camera. All experiments are conducted on a GPU server with NVIDIA A100-SXM4-80GB using PyTorch 2.3.1. The batch size is set to 64.

B. Overall Performance

We compare SymMotion with several state-of-the-art baselines in the MM-Fi dataset: MetaFi++ [8], GaitFi with RGB [28], GaitFi with optical flow [28] and Wi-Flow [28].

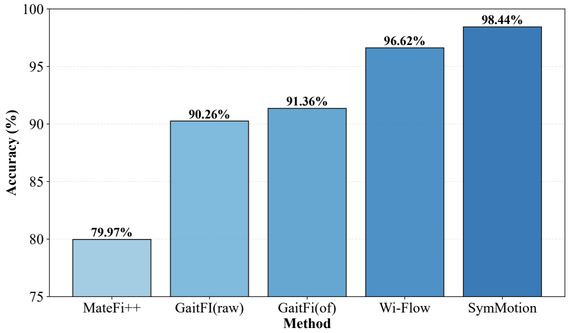

Fig. 2. Accuracy Comparison of Different Methods

Our method outperforms the strongest multi-modal baseline, Wi-Flow, by a margin of 1.82%. In particular, compared to GaitFi (of), which is based on CSI and visual motion, SymMotion improves accuracy by more than 7%. This significant boost confirms that the doppler spectrogram provides critical motion information that CSI misses. Furthermore, SymMotion surpasses MetaFi++ (79.97%) by a substantial margin, demonstrating the superiority of our continuous doppler and flow fusion strategy over traditional transformer-based approaches that may not explicitly align physical motion features.

C. Symmetry Ablation Study

To empirically validate our core about symmetric motion fusion as the optimal pairing, we conducted a systematic study comparing different modality combinations. We kept the backbone and fusion architecture constant and only swapped the input modalities. The results as shown in Table I.

The results strongly corroborate our motion homology hypothesis. The proposed Symmetric (Ours) configuration (Doppler + Optical Flow) achieves the highest accuracy of 98.44%. This indicates that combining two pure motion derivatives—one from the visual domain and one from the wireless domain—minimizes the semantic gap and maximizes feature compatibility.

In particular, the Asymmetric B combination (Doppler + RGB Image) yields the lowest performance (89.25%). This sharp drop reveals the difficulty of aligning high-frequency wireless motion features with static visual textures. The network struggles to ignore the background noise in RGB images to match the pure motion in Doppler spectrograms.

Although Asymmetric A (CSI Amplitude + Optical Flow) performs relatively well (96.76%) due to the strong guidance of optical flow, it still lags behind our method. This suggests that the raw CSI amplitude contains environmental noise that the Doppler transformation successfully filters out, proving that purified wireless features are essential for optimal fusion.

D. Component Effectiveness Analysis

We further analyze the contribution of our proposed architectural modules: the FSG module and the CFD module.

As shown in Table II, removing the CFD results in the most significant performance drop (-6.29%, from 98.44% to 92.15%). This demonstrates that simple feature concatenation is insufficient for heterogeneous modalities. Our learnable queries effectively bridge the domain gap by actively searching for correlated semantic concepts.

The removal of the FSG module leads to a drop of 2.75% (95.69%). This confirms that modality-specific noise (visual background and static wireless paths) harms the fusion process. The FSG successfully purifies the input features before they enter the interaction stage, ensuring that the decoder focuses on valid motion semantics.

E. Robustness Analysis

A key advantage of our system is its robustness against environmental constraints. We designed two stress tests:

We simulate varying lighting conditions by adjusting the brightness of the input images, ranging from normal lighting (100%) to complete darkness (0%).

As shown in Fig. 3, SymMotion exhibits remarkable resilience to changes in brightness.

When brightness drops to 0%, the optical flow quality degrades to None, rendering visual information useless. A vision-only model would theoretically drop to random guessing ($\sim$3.7%). However, SymMotion maintains a high accuracy of 87.26%.

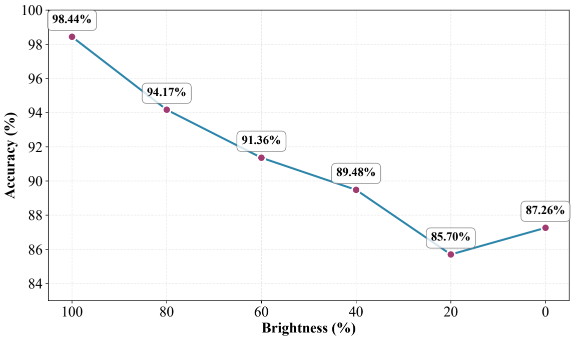

Fig. 3. Impact of Brightness on Accuracy

This performance floor is entirely sustained by the Doppler stream. It proves that our Conceptual Fusion Decoder effectively shifts its attention weight to the wireless modality when the visual stream becomes unreliable.

We observe a slight fluctuation between 20% and 0% brightness (85.7% vs. 87.26%). This is likely due to the network completely detaching from the noisy, low-quality optical flow features at 0% and relying purely on the clean Doppler signal, whereas at 20%, the very bad optical flow might introduce misleading noise.

V. Conclusion

In this paper, we addressed the fundamental challenge of physical inconsistency in multi-modal HAR by introducing SymMotion, a unified framework rooted in the principle of Symmetric Motion Fusion. By theoretically aligning WiFi Doppler spectrograms and Visual Optical Flow as homologous motion derivatives, we bridged the semantic gap that has long hindered effective fusion. We proposed the FSG module to ensure feature purity and the CFD module to achieve active, query-driven semantic alignment. Our empirical results on the MM-Fi dataset not only establish a new SOTA but also validate the hypothesis that fusing symmetric motion-to-motion representations is superior to traditional static-to-dynamic approaches. Furthermore, the demonstrated robustness in zero-light scenarios confirms the practical value of our system for privacy-preserving, all-weather sensing applications. We hope this work inspires future research to prioritize physical consistency in multi-modal learning architectures.

References

[1] M Zaigham Zaheer, Arif Mahmood, M Haris Khan, Mattia Segu, Fisher Yu, and Seung-Ik Lee, "Generative cooperative learning for unsupervised video anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 14744–14754.

[2] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai, "Revisiting skeleton-based action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 2969–2978.

[3] Kaishun Wu, Yandao Huang, Minghui Qiu, Zhenkan Peng, and Lu Wang, "Toward device-free and user-independent fall detection using floor vibration," *ACM Transactions on Sensor Networks*, vol. 19, no. 1, pp. 1–20, 2023.

TABLE I
SYMMETRY ABLATION STUDY

Combination Type	Wireless Modality	Visual Modality	Physical Consistency	Accuracy (%)
Static Symmetric	CSI Amplitude	RGB Image	Low (Static+Static/Mixed)	94.64
Asymmetric A	CSI Amplitude	Optical Flow	Low (Mixed+Motion)	96.76
Asymmetric B	Doppler Spectrogram	RGB Image	Low (Motion+Static)	89.25
Symmetric (Ours)	Doppler Spectrogram	Optical Flow	High (Motion+Motion)	98.44

TABLE II
COMPONENT EFFECTIVENESS ANALYSIS

FSG	CFD	Accuracy (%)
✗	✓	95.69
✓	✗	92.15
✓	✓	98.44

[4] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao, "Videomae v2: Scaling video masked autoencoders with dual masking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 14549–14560.

[5] Yi Zhang, Yue Zheng, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang, "Widar3. 0: Zero-effort cross-domain gesture recognition with wi-fi," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 8671–8688, 2021.

[6] Han Zou, Jianfei Yang, Hari Prasanna Das, Huihan Liu, Yuxun Zhou, and Costas J Spanos, "Wifi and vision multimodal learning for accurate and robust device-free human activity recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 0–0.

[7] Mohammud Junaid Bocus, Xiaoyang Wang, and Robert J Piechocki, "Streamlining multimodal data fusion in wireless communication and sensor networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 10, no. 1, pp. 252–262, 2023.

[8] Jianfei Yang, He Huang, Yunjiao Zhou, Xinyan Chen, Yuecong Xu, Shenghai Yuan, Han Zou, Chris Xiaoxuan Lu, and Lihua Xie, "Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing," *Advances in Neural Information Processing Systems*, vol. 36, pp. 18756–18768, 2023.

[9] Ammar Ladjailia, Imed Bouchrika, Hayet Farida Merouani, Nouzha Harrati, and Zohra Mahfouf, "Human activity recognition via optical flow: decomposing activities into basic actions," *Neural Computing and Applications*, vol. 32, no. 21, pp. 16387–16400, 2020.

[10] Anurag Ranjan, David T Hoffmann, Dimitrios Tzionas, Siyu Tang, Javier Romero, and Michael J Black, "Learning multi-human optical flow," *International Journal of Computer Vision*, vol. 128, no. 4, pp. 873–890, 2020.

[11] Lei Wang, Piotr Koniusz, and Du Q Huynh, "Hallucinating idt descriptors and i3d optical flow features for action recognition with cnns," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8698–8708.

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

[13] Laura Sevilla-Lara, Yiyi Liao, Fatma Güney, Varun Jampani, Andreas Geiger, and Michael J Black, "On the integration of optical flow and action recognition," in *German conference on pattern recognition*. 2018, pp. 281–297, Springer.

[14] Fengyuan Zuo, Haiyan Jin, Zhaolin Xiao, Haonan Su, and Meng Zhang, "Cedflow++: Latent contour enhancement for dark optical flow estimation," *International Journal of Computer Vision*, vol. 133, no. 10, pp. 7222–7241, 2025.

[15] Rui Chen, Jiajun Chen, Zixi Liang, Huaien Gao, and Shan Lin, "Dark-light networks for action recognition in the dark," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 846–852.

[16] Yinqiang Zheng, Mingfang Zhang, and Feng Lu, "Optical flow in the dark," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6749–6757.

[17] Zihao W Wang, Vibhav Vineet, Francesco Pittaluga, Sudipta N Sinha, Oliver Cossairt, and Sing Bing Kang, "Privacy-preserving action recognition using coded aperture videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[18] Lixin Chen, Chaomeng Chen, Jiale Zhou, Zhijian Wu, and Xun Lin, "Stegavar: Privacy-preserving video action recognition via steganographic domain analysis," *arXiv preprint arXiv:2512.12586*, 2025.

[19] Zheng Yang, Yi Zhang, Kun Qian, and Chenshu Wu, "SLNet: A spectrogram learning neural network for deep wireless sensing," in *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*, 2023, pp. 1221–1236.

[20] Guoxuan Chi, Zheng Yang, Chenshu Wu, Jingao Xu, Yuchong Gao, Yunhao Liu, and Tony Xiao Han, "Rf-diffusion: Radio signal generation via time-frequency diffusion," in *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, 2024, pp. 77–92.

[21] Guanghang Liao, Jieming Ma, and Fei Luo, "Human activity recognition by using enhanced radar point cloud 2d histograms and doppler feature fusion," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. 2025, pp. 11234–11241, IEEE.

[22] Changsheng Zhang and Wanguo Jiao, "Imgfi: A high accuracy and lightweight human activity recognition framework using csi image," *IEEE Sensors Journal*, vol. 23, no. 18, pp. 21966–21977, 2023.

[23] Anjun Chen, Xiangyu Wang, Kun Shi, Shaohao Zhu, Bin Fang, Yingfeng Chen, Jiming Chen, Yuchi Huo, and Qi Ye, "Immfusion: Robust mmwave-rgb fusion for 3d human body reconstruction in all weather conditions," *arXiv preprint arXiv:2210.01346*, 2022.

[24] Kian Behzad, Rojin Zandi, Elaheh Motamedi, Hojjat Salehinejad, and Milad Siami, "Robomnist: A multimodal dataset for multi-robot activity recognition using wifi sensing, video, and audio," *Scientific Data*, vol. 12, no. 1, pp. 326, 2025.

[25] Cheng Chen, Shoki Ohta, Takayuki Nishio, Mehdi Bennis, Jihong Park, and Mohamed Wahib, "Enabling visual scene recovery from wi-fi csi for occlusion-free surveillance," *IEEE Internet of Things Journal*, 2025.

[26] Yangyang Gu, Jing Chen, Congrui Chen, Kun He, Jia Ju, Yebo Feng, Ruiying Du, and Cong Wu, "Csipose: Unveiling human poses using commodity wifi devices through the wall," *IEEE Transactions on Mobile Computing*, 2025.

[27] Xiaoyu Shi, Zhaoyang Huang, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li, "Flowformer++: Masked cost volume autoencoding for pretraining optical flow estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 1599–1610.

[28] Zheng Zhang and Junxing Zhang, "Wi-flow: Multimodal human action recognition based on dynamic features in wifi csi and optical flow," in *2025 IEEE Smartworld, Ubiquitous Intelligence and Computing, Scalable Computing and Communications, Digital Twin, Privacy Computing, Metaverse, Autonomous and Trusted Vehicles (SmartWorld/UIC/ScalCom/DigitalTwin/PriComp/Meta)*. 2025, IEEE.