VR-PCT: Enhanced VR Semantic Performance via Edge-Client Collaborative Multi-modal Point Cloud Transformers

Luoyu Mei, Shuai Wang, Ruofeng Liu, Yun Cheng, Shuai Wang*, *Senior Member, IEEE*
Wenchao Jiang, Zhimeng Yin, Tian He, *Fellow, IEEE*

***Abstract*—Real-time semantic recognition is crucial for virtual reality (VR) applications, but the efficient fusion of multi-modal data poses significant challenges under resource-constrained VR scenarios. While integrating millimeter-wave (mmWave) radar point clouds with vision data offers a promising solution, existing methods often suffer from excessive data overhead and degraded accuracy due to redundant and noisy information. To address this limitation, this paper presents VR-PCT, a multi-modal transformer for edge-client collaborative VR semantic recognition that fuses mmWave radar point cloud and vision data for VR applications. VR-PCT introduces a novel collaborative design where VR clients perform lightweight semantic region detection while VR edge processes multi-modal VR semantic recognition. Through efficient edge-client collaboration, VR-PCT optimizes the transmission of mmWave point cloud and vision data by transmitting only the VR semantic region of vision data instead of the entire video. Additionally, it incorporates adaptive cross-modal data selection and fusion strategies to achieve real-time semantic recognition while significantly reducing data redundancy. Across 22 participants engaged in four experimental scenes utilizing VR devices from three different manufacturers, our evaluation demonstrates that VR-PCT achieves 97.6% recognition accuracy while reducing transmission overhead by 81.5% compared to existing approaches. These results highlight the effectiveness of VR-PCT in enabling efficient and accurate multi-modal VR semantic recognition for VR applications. The code and data of VR-PCT are released on https://github.com/luoyumei1-a/VR-PCT.**

***Index Terms*—mmWave Radar, Enhanced VR Semantic, Multi-modal Transformer, Edge-client Collaboration.**

I. Introduction

The rapid evolution of Virtual Reality (VR) technology has created increasing demands for robust and privacy-preserving semantic content recognition capabilities, such as recognizing hand gestures and head movements for intuitive user activity types and keystroke input [1]–[3]. Current VR systems rely on two primary sensing modalities with complementary strengths and limitations: (i) Camera-based visual sensing [4] provides rich semantic information and high-resolution imagery but becomes ineffective under occlusion scenarios and raises significant privacy concerns during data collection and transmission [5]; (ii) Millimeter-wave (mmWave) radar sensing [6], [7] offers unique advantages in penetrating non-metallic obstacles and preserving privacy through point cloud generation, but suffers from sparse point distribution and limited semantic information [8], [9]. This complementary nature presents a compelling opportunity for multi-modal fusion where vision data provides detailed semantic information under line-of-sight conditions, while mmWave radar ensures continuous tracking under non-line-of-sight scenarios. However, processing and fusing these diverse data streams presents significant challenges due to VR devices' limited computational resources, and balancing recognition accuracy with privacy protection.

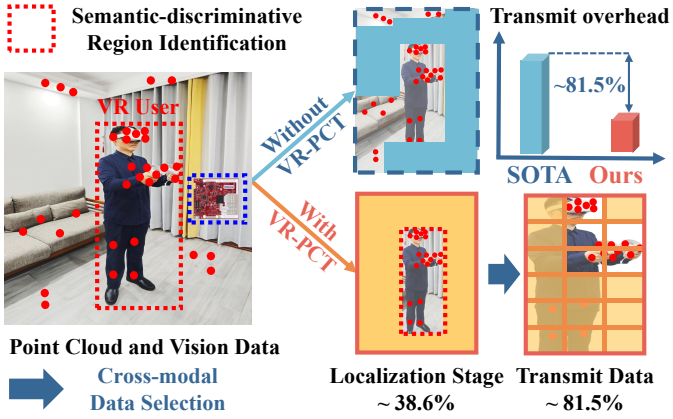


Fig. 1: VR-PCT enhances model accuracy while reducing the VR data transmission overhead by 81.5%.

Existing state-of-the-art (SOTA) approaches in this domain face fundamental limitations in effectively leveraging these complementary modalities: (i) Vision transformer (ViTs) based methods [10], [11] achieve high accuracy through processing rich visual features but fail under occlusion scenarios and raise privacy concerns when streaming raw visual data [12], [13]; (ii) Point transformer-based methods [14], [15] enable privacy-preserving sensing through point clouds but struggle with semantic content recognition due to sparse point distribution and limited feature representation [16]. Neither approach effectively combines the complementary strengths of both modalities while meeting VR's critical requirements [17], where reducing data transmission overhead is essential to minimize latency while improving semantic content recognition accuracy for a more immersive experience [18].

The impact of these limitations hampers VR applications' multi-modal processing and increases latency. Current VR

Corresponding author: Shuai Wang
L. Mei, S. Wang, S. Wang, and T. He are with the Southeast University, China. (email: lymei-@seu.edu.cn; shuaiwang_iot@seu.edu.cn; shuaiwang@seu.edu.cn; tianhe@seu.edu.cn)
L. Mei and Z. Yin are with the City University of Hong Kong, Hong Kong. (email: zhimeyin@cityu.edu.hk)
R. Liu is with the Michigan State University, USA. (email: liuruofe@msu.edu)
Y. Cheng is with the ETH Zurich, Switzerland. (email: yun.cheng@sdsc.ethz.ch)
W. Jiang is with the Singapore University of Technology and Design, Singapore. (email: wenchao_jiang@sutd.edu.sg)

designs inefficiently transmit complete vision and point cloud data without considering semantic relevance, causing network congestion and privacy risks. Moreover, the lack of efficient cross-modal selection mechanisms wastes bandwidth and increases delay in dynamic VR environments. Therefore, an edge-client collaborative framework is needed to selectively transmit only semantically relevant multi-modal data fragments instead of complete streams, thereby preserving privacy while maintaining recognition accuracy.

To overcome these limitations, we propose VR-PCT, a multi-modal edge-client collaboration transformer for VR semantic content recognition. VR-PCT addresses three key challenges: (i) How to achieve efficient semantic region localization utilizing mmWave point clouds on resource-constrained VR devices, (ii) How to optimize cross-modal data filtering while preserving semantic information, and (iii) How to effectively fuse reduced point cloud and visual data at VR edge to maximize recognition performance. The VR-PCT model addresses these challenges with an edge-client collaboration framework, based on our key observation that *not all multi-modal data contribute positively to recognition accuracy, certain segments even degrade model performance by introducing noise and redundancy*. Therefore, VR-PCT leverages mmWave point clouds on clients to detect semantic regions, enabling targeted filtering of both vision and point cloud data. By transmitting semantic data segments to the edge, VR-PCT reduces noise, preserves privacy, and enhances semantic content recognition through a multi-modal transformer.

Specifically, VR-PCT employs a lightweight cross-modal semantic localization mechanism on clients that utilizes mmWave point clouds to identify semantic regions, as illustrated in Fig. 1, where the system selectively transmits only semantically relevant regions. The inherent spatial correlation between mmWave and vision data enables efficient semantic detection through privacy-preserving mmWave sensing. This correlation enables effective filtering of both modalities, preserving semantic information while reducing transmission overhead by 81.5%. At the edge, VR-PCT implements a multi-modal transformer architecture that integrates spatial-temporal features from both filtered data streams, where mmWave point clouds provide robust spatial information through occlusions while vision data contributes detailed semantic features. Through cross-modal attention mechanisms, VR-PCT captures correlations between modalities, enabling accurate VR semantic content recognition in challenging scenarios.

Our comprehensive evaluation demonstrates VR-PCT's superior performance in both robustness and efficiency. The framework exhibits robust recognition capabilities under various challenging scenarios, including through-wall occlusions and privacy-sensitive situations. These results validate VR-PCT's effectiveness in real-world VR applications, where our adaptive multi-modal approach ensures consistent performance across different environmental conditions while preserving user privacy. Our key contributions are:

- We are the first to propose an edge-client collaboration multi-modal VR semantic content recognition framework that leverages the complementary strengths of vision and mmWave point cloud data for robust VR semantic content recognition while preserving privacy.
- We develop a novel two-stage processing pipeline that combines lightweight semantic region localization on VR clients with multi-modal semantic content recognition at the VR edge, enabling efficient data filtering and transmission while maintaining recognition accuracy.
- We demonstrate VR-PCT's effectiveness through extensive experiments with 22 participants across four scenes with VR devices from three manufacturers, achieving 97.6% recognition accuracy while reducing transmission overhead by 81.5% compared to SOTA approaches.

The remainder of the paper is organized as follows. Section II reviews existing state-of-the-art approaches, highlighting their limitations for VR semantic content recognition. Section III presents the system overview of VR-PCT. Section IV elaborates on the VR-PCT design, including VR semantic region localization, cross-modal VR semantic data selection, and multi-modal VR semantic recognition. Section V presents our comprehensive evaluation methodology and experimental results across various VR platforms, occlusion scenarios, and network conditions. Finally, Section VI concludes the paper.

II. Related Works

Recent advances in VR semantic content recognition are categorized into three main approaches: Vision Transformer, Point Transformer, and Multi-modal Transformer. However, these approaches do not effectively leverage client-edge collaboration to optimize the trade-off between wireless transmission overhead and semantic content recognition performance.

Vision Transformer. Vision transformers [19] have transformed computer vision by adapting transformer architectures to process image patches as sequential tokens. These architectures excel in tasks including object detection [20], [21], semantic segmentation [22], and video comprehension [23]–[25]. However, Vision Transformers require transmission of comprehensive vision data, imposing substantial bandwidth and computational demands [26]. They also raise privacy vulnerabilities [27], [28], making them inadequate for privacy-sensitive VR applications [29]. Additionally, these models have limitations in processing 3D point cloud data due to irregularity and sparsity characteristics [30], [31] and perform poorly in non-line-of-sight conditions [12], [32], diminishing their efficacy for VR semantic content recognition across diverse operating conditions [10], [33].

TABLE I: Comparison of VR-PCT and SOTA methods.

	Motion Tracking	**VR Semantic**
Vision or Point Cloud	[14], [19]	[22], [29], [34]
Multi-modal	[35]–[37]	**VR-PCT (Our Work)**

Point Transformer. Point transformers [14], [38] adopt transformer mechanisms to point cloud data without requiring preprocessing. Recent advances include Self-supervised 4D [39] for representation learning, Interpretable3D [16] for architectural interpretability, and Point Transformer V3 [14] for computational scalability. The 3D Object Tracking [15],

[38] employs hierarchical operations with relation-aware sampling and point-BEV fusion. However, these architectures face fundamental challenges in simultaneously optimizing accuracy and efficiency, demonstrating an inherent trade-off between performance and system overhead while failing to achieve reduced data transmission costs with high recognition accuracy.

Multi-modal Transformer. Multi-modal transformers demonstrate promising fusion capabilities but have limitations in bandwidth-constrained VR environments. CMDTrack [35] employs a teacher-student knowledge distillation framework to bridge the performance gap between multimodalities. MM-STC [36] utilizes an encoder-decoder Transformer architecture for video and time-series data, UFAFormer [40] combines frequency domain with spatial features, and IS-Fusion [37] implements instance-scene fusion for 3D detection. However, these architectures exhibit limitations in bandwidth-constrained scenarios due to their requirement for comprehensive multi-modal data transmission. Processing across different modalities imposes substantial bandwidth demands and introduces significant latency through both transmission and the synchronized processing of heterogeneous data types requirements. These limitations are pronounced in real-time applications where bandwidth and latency are crucial.

In contrast to existing approaches, VR-PCT introduces a client-edge collaborative framework for multi-modal fusion in VR semantic content recognition. Table I reveals vision and point transformers [14], [19], [22] focus on motion tracking but suffer from high computational demands and privacy vulnerabilities when applied to VR semantics, while single-modal approaches [29], [34] address VR semantic content recognition but struggle with limited accuracy and efficiency. Existing multi-modal methods [35]–[37] are predominantly utilized for motion tracking. To address these limitations, VR-PCT proposes a multi-modal transformer for VR semantic content recognition, performing lightweight semantic region localization on point cloud data at the VR client to identify the VR semantic region. This approach reduces overhead while maintaining accuracy, leveraging complementary strengths of both modalities through mmWave point cloud localization on the client for efficient data filtering before edge transmission, thereby optimizing computational efficiency and recognition performance through edge-client collaboration.

III. Overview

VR-PCT leverages lightweight mmWave point cloud and vision data for client-edge collaborative VR semantic content recognition. It significantly reduces the data volume transmitted from VR clients (e.g., VR headsets) to VR edge devices (e.g., VR PCs, switches) while maintaining high VR semantic content recognition accuracy. By strategically selecting and processing only semantically relevant data, VR-PCT efficiently addresses bandwidth constraints and computational limitations inherent in resource-constrained VR devices, enabling robust semantic content recognition across diverse usage scenarios.

Fig. 2 illustrates the design architecture of VR-PCT, which processes original mmWave point cloud and vision data as input and generates VR semantic information (e.g., activity types and keystroke inputs) as output. The system comprises two primary components: the VR client and the VR edge. The VR client performs lightweight semantic region localization and semantic-discriminative regions selection, while the VR edge conducts comprehensive multi-modal VR semantic content recognition. The following sections detail the technical components of these two essential modules. The system processes multi-modal edge-client collaboration as follows:

- At the VR client, VR-PCT employs VR semantic region localization utilizing mmWave point clouds to rapidly identify VR semantic positions. Not all vision data positively contributes to model accuracy, with certain portions potentially misleading the model. VR-PCT subsequently utilizes these identified semantic regions to perform semantic-discriminative region selection. This process efficiently filters semantically rich vision data while eliminating interference data. By selectively transmitting only semantically enriched vision and mmWave data, VR-PCT substantially reduces data transmission overhead while preserving critical VR semantic information.
- At the VR edge, upon receiving the filtered vision data and semantically annotated mmWave point clouds from the VR client, VR-PCT performs multi-modal VR semantic content recognition through sophisticated transformer-based architectures. The edge processing integrates modality-specific feature enhancement with environment-aware adaptive fusion, enabling robust recognition of VR user activity types and keystroke inputs. This multi-modal approach leverages the complementary characteristics of each modality, where high-resolution vision data provides detailed semantic features within direct line-of-sight, while mmWave point clouds maintain spatial tracking through obstacles.

This comprehensive approach enables VR-PCT to achieve efficient VR semantic content recognition through client-edge collaboration. By intelligently localizing semantic regions at the client and performing VR semantic content recognition through multi-modal fusion at the edge, the system significantly reduces transmission bandwidth while maintaining high recognition accuracy. This synergistic design effectively addresses resource constraints in VR environments while enabling robust activity and keystroke recognition across diverse usage scenarios and environmental conditions.

IV. VR-PCT Design

A. VR Semantic Region Localization at VR Client

VR-PCT employs a lightweight semantic region localization module at the VR client, leveraging mmWave point cloud sparsity for computational efficiency on resource-limited devices. This approach preserves privacy by capturing semantic representations without sensitive vision information. These low computational overhead and privacy-preserving characteristics form the foundation for VR semantic region localization.

Fig. 3 illustrates the VR semantic region localization module. The module processes multi-frame mmWave point cloud data of dimensionality $s \times N \times d$, containing spatiotemporal features. A point transformer computes attention scores for

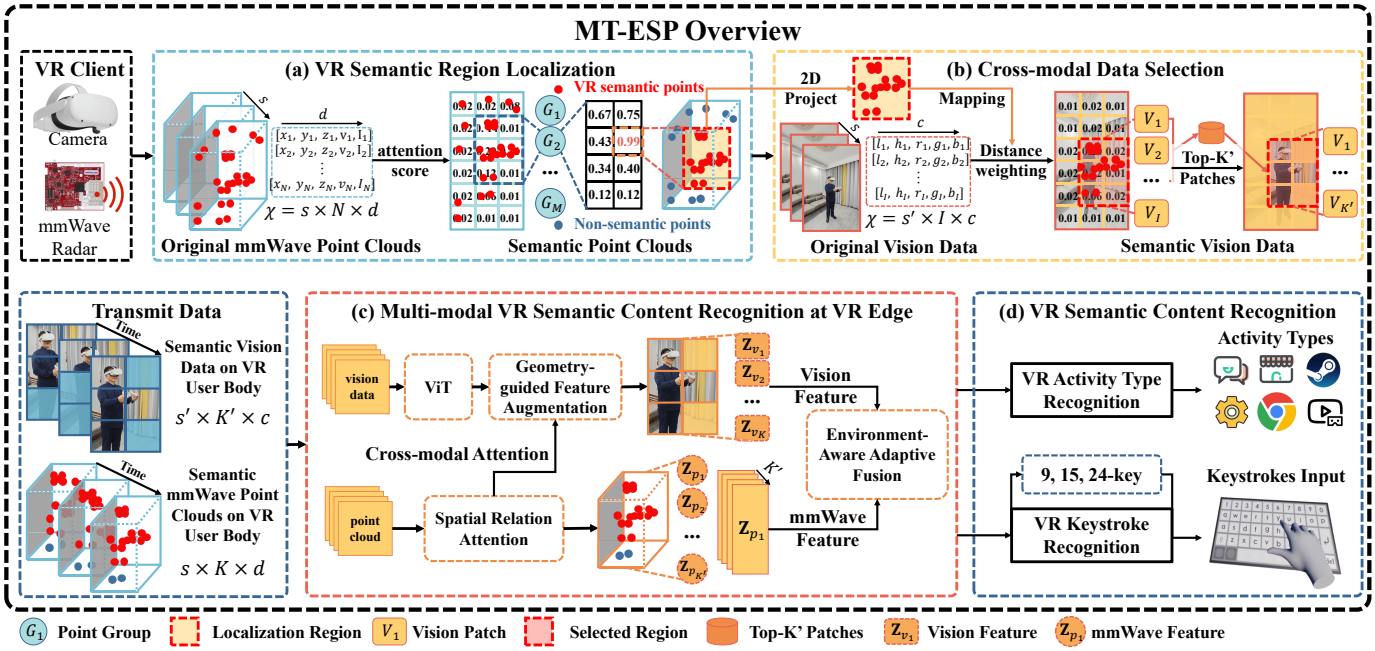


Fig. 2: Overview of VR-PCT. (a) VR semantic region localization identifies semantically significant areas by processing mmWave point clouds through attention mechanisms. (b) Cross-modal data selection projects 3D semantic points onto 2D vision data and selects and transmits the semantically relevant segments to the VR edge. (c) Multi-modal VR Semantic Recognition at VR edge processes the transmitted semantic data through modality-specific feature enhancement and environment-aware adaptive fusion. (d) VR semantic content recognition modules analyze the fused features to determine activity types and recognize VR keystroke inputs across various keyboard layouts.

individual points, quantifying semantic significance through spatial and feature relationships. The system clusters attention coefficients based on spatial proximity to identify semantically coherent regions. For each spatial cluster, cumulative attention scores measure semantic relevance, with the maximum-scoring cluster located as the primary semantic region.

The semantic localization framework employs an iterative refinement process through Top-K point selection within each frame. Starting from regions with maximum cumulative attention scores, the system implements a central expansion algorithm to identify points with the highest semantic information. This optimization reduces point cloud dimensionality while preserving critical semantic information. Applied across the temporal dimension, the process produces a refined point cloud representation of dimensionality $s \times K \times d$, where K represents selected points per frame. The system facilitates interpretation of user interactions and intentionality through precise localization of semantically salient regions such as the upper extremities, head, and torso.

Specifically, to achieve efficient VR semantic content recognition localization on resource-constrained client devices, we implement a point transformer [14] for processing mmWave point cloud data $\mathcal{P}$. The point cloud data is formalized as a tensor $\mathcal{X} \in \mathbb{R}^{s \times N \times d}$, where s represents the combined batch and sequence dimension, N denotes the cardinality of points per frame, and d corresponds to the feature dimensionality encompassing spatiotemporal characteristics (x, y, z, v, I), incorporating spatial coordinates, velocity, and intensity measurements. This representation encodes the temporal evolution of user motion patterns within the VR environment.

The point transformer architecture employs a vector attention mechanism to analyze point-wise spatial and feature domain relationships, facilitating semantic point selection for VR recognition. The vector attention is expressed as:

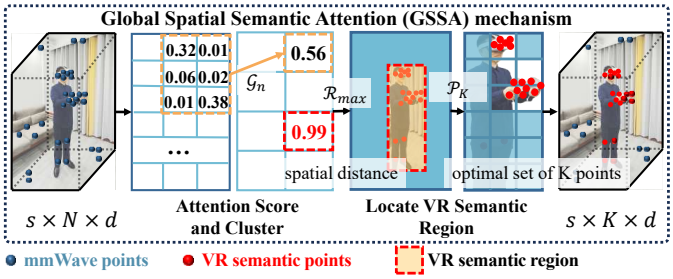


Fig. 3: VR Semantic Region Localization.

$$\vec{y}_i = \sum_{\vec{x}_j \in \mathcal{X}} \vec{a}_{ij} \odot h(\vec{x}_j), \tag{1}$$

where $\vec{y}_i$ denotes the output feature vector for the i-th point, $\vec{a}_{ij}$ represents the attention weight between points i and j, h is a learnable feature transformation function. The attention weight $\vec{a}_{ij}$ is computed through a composition of feature transformations ϕ and ψ, relation function β, mapping function γ, learned bias δ, and non-linear activation function σ:

$$\vec{a}_{ij} = \sigma\left(\gamma\left(\beta\left(\phi(\vec{x}_i), \psi(\vec{x}_j)\right) + \delta\right)\right). \tag{2}$$

After obtaining the attention scores, VR-PCT spatial clusters semantically coherent regions through the Global Spatial Semantic Attention (GSSA) mechanism, which operates by analyzing the spatial distribution of attention weights across

the multi-modal feature space to identify regions with high semantic relevance for VR interactions:

$$\mathcal{G}_n = \sum_{\vec{a}_{ij} \in \mathcal{S}_k} \vec{a}_{ij}, \tag{3}$$

where $\mathcal{G}_n$ represents the n-th spatial cluster, and $\mathcal{S}_k$ denotes the k-th attention weights set. The primary semantic region is identified through maximum attention aggregation:

$$\mathcal{R}_{max} = \underset{n}{\operatorname{argmax}}\, \mathcal{G}_n, \tag{4}$$

where $\mathcal{R}_{max}$ represents the region with the highest VR semantic. As illustrated in the GSSA mechanism workflow, this process systematically evaluates all spatial clusters to locate the region containing the most semantically significant VR interaction points. Utilizing this region as the center, VR-PCT employs a diffusion-based Top-K point selection strategy:

$$\mathcal{P}_K = \{p_i \mid d(p_i, \mathcal{R}_{max}) \leq r_k, |\mathcal{P}_K| = K\}, \tag{5}$$

where $d(p_i, \mathcal{R}_{max})$ measures the spatial distance between point p_i and $\mathcal{R}_{max}$, r_k is an adaptive radius threshold ensuring selection of exactly K points, and $\mathcal{P}_K$ represents the optimal set of K points with maximal semantic information per frame.

This mechanism identifies semantically relevant regions on VR users' hands and headsets, which is critical for understanding VR interactions. The GSSA mechanism ensures selected regions maintain both high semantic relevance and spatial coherence, essential for accurate VR semantic recognition under occlusion. When the maximal GSSA score in $\mathcal{R}_{max}$ falls below the threshold η, the algorithm performs iterative refinement through recursive GSSA application until achieving adequate semantic significance, enabling precise localization of key regions for enhanced VR interaction interpretation.

B. Cross-modal VR Semantic Data Selection

Building upon the semantic regions identified in the previous stage, VR-PCT employs a cross-modal VR semantic-discriminative region selection mechanism to efficiently filter and transmit both mmWave point cloud and vision data to the VR edge. The fundamental principle exploits the inherent spatial correspondence between mmWave point cloud and vision data captured from the same VR user, enabling identification of VR semantic locations through their correlated distribution patterns and filtering out irrelevant data outside these semantic regions. As illustrated in Fig. 4, the upper part demonstrates semantic point selection from 3D point clouds. To enable cross-modal selection, VR-PCT projects these identified 3D semantic points onto a 2D plane while preserving their spatial relationship. The mmWave point cloud and vision data exhibit inherent spatial correspondence with a specific spatial transformation as shown in Equation 6. By projecting the identified semantic point clouds onto vision frames, illustrated as red points in the lower part, VR-PCT locates corresponding semantic regions within the vision data, enabling selection of vision patches with VR semantic information.

This cross-modal selection approach effectively reduces data transmission requirements by locating and only transmitting

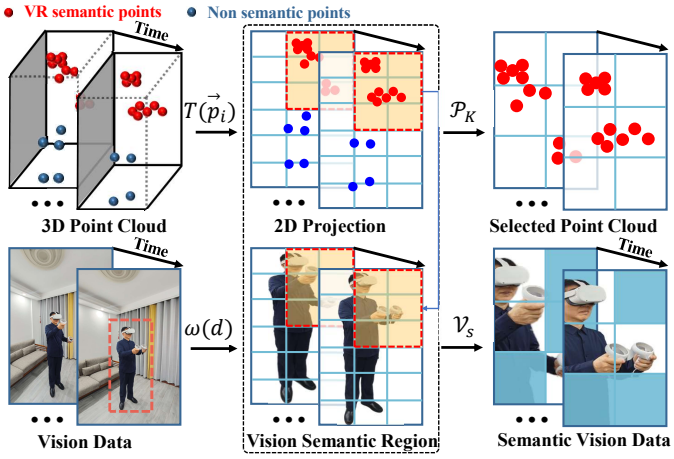


Fig. 4: Cross-modal VR Semantic Data Selection.

data in the VR semantic regions while filtering out irrelevant noise. Thus, it enables efficient and privacy-aware data transmission to the VR edge for VR semantic content recognition. Specifically, VR-PCT projects the 3D mmWave point cloud semantic regions onto the two-dimensional vision coordinate space. The spatial transformation $T(\cdot)$ from 3D point cloud coordinates to 2D vision coordinates is formulated as:

$$T(\vec{p}_i) : \begin{bmatrix} u & v & 1 \end{bmatrix} = \frac{1}{y} M_1 M_2 \begin{bmatrix} x & y & z & 1 \end{bmatrix}, \tag{6}$$

where (u, v) is the corresponding pixel location in the vision data's pixel coordinate system, M_1 is the 3×3 intrinsic camera parameter matrix of the vision modality, and M_2 is the 3×4 extrinsic matrix encoding the relative position and orientation between the mmWave radar and vision sensors. Because the mmWave radar and vision sensors are fixed at known positions on the VR device, the transformation parameters M_1 and M_2 are pre-calibrated and remain constant.

Utilizing this projection, the cross-modal selection process identifies relevant vision image patches corresponding to the semantic regions. For the transformed point cloud regions $T(\vec{p}_i)$ from the identified semantic points in the previous stage, the corresponding Top-K' vision image patches around these projected locations are selected:

$$\mathcal{V}_s = \{ \underset{|I_{selected}|=K'}{I(u,v) \in \mathcal{V}} \mid \sum_{p_i} \vec{a}_{ij} \cdot \omega(I(u,v), T(\vec{p}_i))) \}, \tag{7}$$

where $I(u, v)$ represents the image patch centered at pixel coordinates (u, v), $T(\cdot)$ is the spatial transformation from 3D point cloud to 2D vision coordinates as defined in Equ. 6, $\vec{a}_{ij}$ denotes the attention weight of point p_i obtained from Equ. 2, K' is the number of selected image patches, and $\omega(\cdot)$ is a distance-based weighting function defined as:

$$\omega(d) = \begin{cases} \exp(-\frac{d^2}{2\sigma^2}) & \text{if } d \leq r_v \\ 0 & \text{otherwise} \end{cases}, \tag{8}$$

where d represents the Euclidean distance in the 2D image plane between the projected point and patch center, σ is a

scaling parameter controlling the influence range of each point, and r_v is the maximum radius threshold for patch selection.

The reformulated selection mechanism incorporates both spatial proximity and semantic importance through the attention weights $\vec{a}_{ij}$. The exponential weighting function $\omega(d)$ provides smooth distance-based importance transitions, while the cutoff radius r_v maintains spatial locality. This cross-modal selection approach leverages complementary characteristics where 3D point clouds preserve spatial positioning and user poses through obstacle penetration, while 2D vision captures fine-grained VR semantics like controller and headset movements. By selecting Top-K' image patches $\mathcal{V}_s$ with the highest weighted scores corresponding to identified VR semantic regions, our approach achieves efficient semantic preservation with minimal data redundancy at the VR edge.

To quantify the impact of cross-modal VR semantic-discriminative region selection on VR semantic content recognition performance, we conduct an analysis of both latency reduction and computational overhead. The total end-to-end latency comprises three principal components: semantic region localization time T_l, data transmission time T_t, and edge processing time T_p. Through theoretical analysis, we establish that both transmission and processing latencies exhibit proportional relationships with the data volume ratio η:

$$T_t = T_0 \cdot \frac{V_s + P_k}{V_t + P_t}, \quad T_p = T_b \cdot \frac{V_s + P_k}{V_t + P_t}, \tag{9}$$

where T_0 and T_b represent the original transmission latency and baseline processing latency, respectively. The parameters V_t and P_t denote the total volume of vision and point cloud data, while V_s and P_k represent the selected vision and top-K point cloud data volume, with values K and $K^{'}$ respectively.

The semantic region localization mechanism introduces computational overhead comprising feature transformation, attention computation, and point selection operations. To evaluate the efficiency gains, we define the data reduction ratio:

$$\eta = 1 - \frac{V_s + P_k}{V_t + P_t}. \tag{10}$$

This formulation enables precise measurement of data volume reduction while maintaining semantic integrity through selective data transmission. This overhead is illustrated as:

$$T_l = f(\mathbf{x}) + g(\mathbf{a}) + h(K), \tag{11}$$

where $f(\cdot)$ denotes feature transformation operations, $g(\cdot)$ represents attention computation and grouping mechanisms, and $h(\cdot)$ encapsulates the top-K selection process. Subsequently, the end-to-end latency reduction is formulated as:

$$\Delta T = T_0 \cdot (1 - \eta) - T_l, \tag{12}$$

where T_0 represents the original processing latency and η is the data reduction ratio. The empirical analysis demonstrates that the localization overhead is substantially lower than network transmission latency. Furthermore, given the inherent sparsity of point cloud data compared to high-resolution vision data, our approach achieves significant latency reduction while maintaining semantic content recognition accuracy through precise selection of semantically relevant regions.

C. Multi-modal VR Semantic Recognition at VR Edge

At the VR edge, VR-PCT implements a sophisticated multi-modal transformer architecture to effectively fuse semantic-enriched mmWave point cloud $\mathcal{P}_K$ and vision data $\mathcal{V}_s$ received from VR clients. As illustrated in Fig. 5, VR-PCT first enhances the semantic features of both mmWave and vision data through modality-specific feature enhancement. A spatial relation attention mechanism extracts VR users' spatial semantic information from mmWave data to obtain point features. Meanwhile, cross-modal attention integrates visual semantics extracted by vision transformers with spatial information from mmWave point clouds to generate vision features with a three-dimensional spatial context. These semantic features are then fed into environment-aware adaptive fusion for multi-modal integration, where environmental-modulated attention gating dynamically adjusts feature weights based on the VR user's environmental conditions. For instance, it prioritizes mmWave semantics under heavy occlusion and favors visual features when line-of-sight is clear. Following the attention gating, adaptive feature integration combines mmWave and visual modalities to produce fused VR semantic features.

1) Modality-Specific Feature Enhancement: To facilitate effective multi-modal fusion, VR-PCT first processes each modality through specialized attention networks designed to enhance modality-specific characteristics.

For semantic mmWave point cloud data $\mathcal{P}_K \in \mathbb{R}^{s\times K\times d}$, we introduce a hierarchical spatial relation enhanced attention mechanism that integrates local geometric structures and global spatial correlations:

$$\mathcal{F}_p(p_i) = \sum_{p_j \in \mathcal{N}_k(p_i)} \omega(p_i, p_j) \cdot (\Phi(p_j) + \mathcal{E}(p_i, p_j)), \tag{13}$$

where $\mathcal{F}_p(p_i)$ denotes the enhanced feature representation for point p_i, $\mathcal{N}_k(p_i)$ represents the k-nearest neighbor set in the metric space defined by Euclidean distance, $\Phi(\cdot)$ is a learnable transformation function implemented as a multi-layer perceptron (MLP), and $\mathcal{E}(p_i, p_j)$ encodes relative spatial positions through sinusoidal positional encoding. The attention weight $\omega(p_i, p_j)$ between points p_i and p_j is computed through a normalized attention mechanism:

$$\omega(p_i, p_j) = \frac{\exp(\sigma(\Theta_q(p_i))^T \cdot \sigma(\Theta_k(p_j)))}{\sum_{x_m \in \mathcal{N}_k(p_i)} \exp(\sigma(\Theta_q(p_i))^T \cdot \sigma(\Theta_k(x_m)))}, \tag{14}$$

where Θ_q, Θ_k denote learnable query and key transformation functions, respectively implemented as MLPs with layer normalization, and σ represents the ReLU activation function ensuring non-negative attention weights.

This attention mechanism enables the aggregation of information from its spatial neighbors by computing normalized weights. To capture multi-scale geometric relationships, we introduce a hierarchical feature aggregation mechanism:

$$\mathcal{H}_l(p_i) = \Lambda(\{\mathcal{F}_p(p_j) | p_j \in \mathcal{N}_{r_l}(p_i)\}), \tag{15}$$

where $\mathcal{H}_l(p_i)$ represents the hierarchical feature at level l, $\mathcal{N}_{r_l}(p_i)$ denotes the neighborhood with radius r_l, and increases

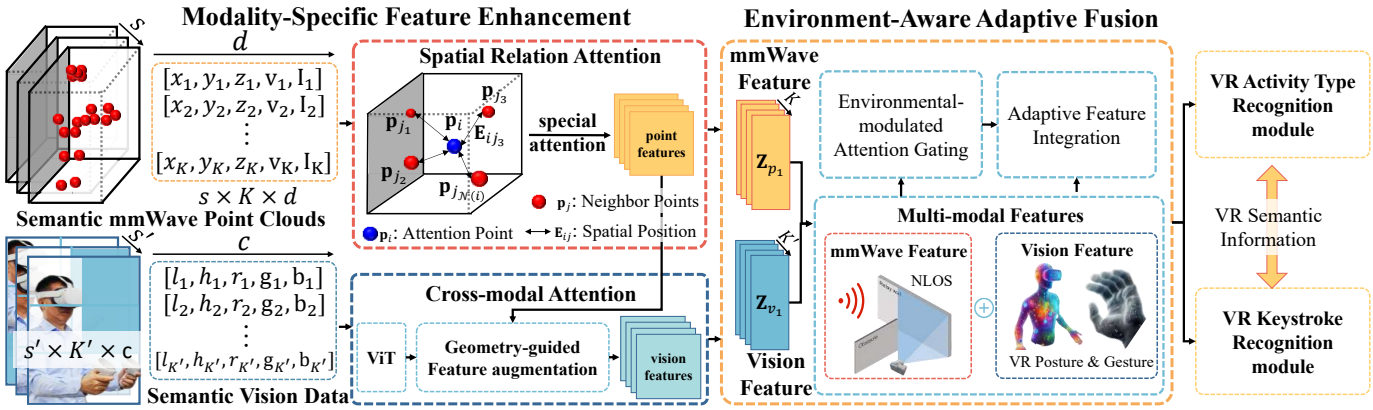


Fig. 5: Multi-modal VR Semantic Recognition.

with level l to capture increasingly spatial contexts. $\Lambda(\cdot)$ represents a learnable aggregation function that adaptively combines features from neighboring points.

For vision data $\mathcal{V}_s \in \mathbb{R}^{s\times K'\times c}$, we propose a geometry-guided cross-modal attention mechanism that enhances vision features by incorporating special features from point clouds:

$$\mathcal{F}_v(v_i) = \Psi_2(\sigma(\Psi_1(\alpha(v_i)\cdot v_i + (1-\alpha(v_i))\cdot \mathcal{C}(v_i)))), \quad (16)$$

where $\Psi_1 \in \mathbb{R}^{d_h\times c}$ and $\Psi_2 \in \mathbb{R}^{c\times d_h}$ represent learnable projection functions, $\alpha(v_i)$ is a dynamic balancing function that adapts to local feature characteristics:

$$\alpha(v_i) = \sigma(W_\alpha[v_i; \Lambda(\mathcal{C}(v_i))] + b_\alpha), \quad (17)$$

The geometric context feature $\mathcal{C}(v_i)$ is computed through a cross-modal attention mechanism:

$$\mathcal{C}(v_i) = \sum_{j=1}^{K} \gamma(v_i, p_j)\cdot(\Upsilon(p_j) + \Delta(v_i, p_j)), \quad (18)$$

where Υ denotes a learnable point cloud feature transformation, and $\Delta(v_i, p_j)$ encodes geometric relationships through a learned mapping function. The cross-modal attention weight $\gamma(v_i, p_j)$ incorporates both content and spatial attention:

$$\gamma(v_i, p_j) = \text{softmax}(\frac{\mathcal{Q}(v_i)\mathcal{K}(p_j)^T}{\sqrt{d_k}})\cdot \mathcal{S}(v_i, p_j), \quad (19)$$

where $\mathcal{Q}$ and $\mathcal{K}$ represent query and key transformation functions, and $\mathcal{S}(v_i, p_j)$ is a learned spatial attention mask that enforces geometric consistency between modalities.

2) Environment-Aware Adaptive Fusion: To leverage the complementary characteristics of enhanced mmWave point cloud features and vision features, VR-PCT implements an environment-aware fusion strategy consisting of two key components: environmental-modulated attention gating and adaptive feature integration. This dual-stage fusion architecture explicitly addresses the challenges of varying occlusion conditions in VR environments, where visual information may become temporarily unavailable due to physical obstacles while mmWave signals maintain their penetrative sensing capabilities for continuous VR semantic information capture.

In the environmental-modulated attention gating stage, VR-PCT dynamically integrates the enhanced features from both modalities through an environment-sensitive attention module:

$$\Phi_{gate} = \sum_{m\in\{p,v\}} \Lambda_m \odot \mathcal{F}_m + \lambda\cdot\sum_{h=1}^{H}\mathcal{W}_h\Gamma(\mathcal{F}_p, \mathcal{F}_v), \quad (20)$$

where the modality-specific adaptive weights Λ_m are computed through a sophisticated attention mechanism that incorporates comprehensive environmental context awareness:

$$\Lambda_m = \phi(\frac{\mathcal{Q}_m(\Omega[\mathcal{F}_p; \mathcal{F}_v; \mathcal{O}])^T}{\sqrt{d}})\cdot\sigma(\mathcal{V}_m\mathcal{H} + \beta_m), \quad (21)$$

where $\phi(\cdot)$ denotes a normalized exponential function that maps inputs to a probability distribution, $\mathcal{O}$ represents a rich set of occlusion context features derived from real-time environmental analysis, $\mathcal{H}$ encodes hierarchical environmental information capturing multi-scale spatial relationships and temporal dynamics, and Ω, $\mathcal{Q}_m$, $\mathcal{V}_m$, and β_m represent learnable transformation parameters optimized during the training process to capture complex cross-modal interactions.

Following the environmental gating, we introduce an adaptive feature modulation mechanism that employs a context-aware attention framework for cross-modal feature integration:

$$\mathcal{A} = \phi(\frac{\mathcal{Q}\mathcal{K}^T}{\sqrt{d_k}})\mathcal{V}, \quad (22)$$

where $\phi(\cdot)$ represents a normalized exponential operator that captures the intrinsic correlations between modalities. The integrated features undergo a non-linear transformation through a learnable mapping function:

$$\mathcal{F} = W_2(\rho(W_1\mathcal{A} + b_1)) + b_2. \quad (23)$$

This sophisticated fusion architecture enables VR-PCT to adaptively calibrate and synthesize complementary modal characteristics via environment-aware attention mechanisms. The proposed framework exhibits robust modal importance modulation capabilities, automatically emphasizing mmWave features during occlusion scenarios while leveraging high-fidelity visual information under clear line-of-sight conditions.

The resultant integrated representation $\mathcal{F}$, in conjunction with point cloud and visual features, serves as input to specialized modules for activity recognition and VR interaction analysis, facilitating a comprehensive understanding of VR user activity types and keystroke input.

3) VR Activity Recognition: Leveraging the environment-aware adaptive multi-modal fusion features $\mathcal{F}$, mmWave point cloud $\mathcal{P}_{body}$, and vision data $\mathcal{V}_{body}$ of the VR user, we develop a temporal modeling framework for VR activity recognition. This framework distinguishes various VR activities by analyzing behavioral patterns through both integrated and modality-specific features. VR-PCT identifies characteristic activities, including chatting, shopping, gaming, system configuration, browsing, and video watching, by capturing distinct motion patterns and postures. Each activity exhibits unique signatures. For instance, chatting involves frequent controller movements for text input, gaming shows large-scale body motions, shopping features repetitive sliding and clicking, system settings combine brief sliding with keyboard input, and video watching maintains stable sitting or standing postures.

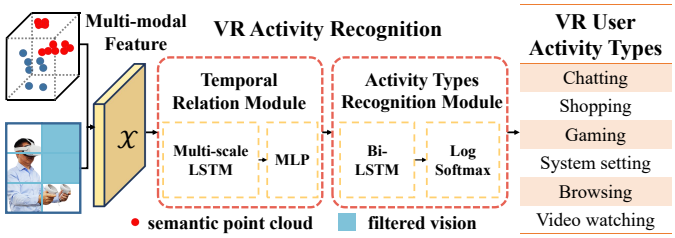


Fig. 6: VR Activity Type Recognition.

The temporal model extracts temporal relationships from the multi-modal input features:

$$\mathcal{X} = [\mathcal{F} \| \mathcal{P}_K \| \mathcal{V}_s], \tag{24}$$

where $\mathcal{X}$ represents the concatenated feature vector that combines environment-aware fusion features and modality-specific features, and $\|$ denotes feature concatenation. To capture temporal dependencies across time, temporal semantic features are computed through a multi-scale attention mechanism:

$$\mathcal{H}_t = \sum_{\tau \in \mathcal{T}} w_\tau \cdot \Psi(\mathcal{X}_t, \mathcal{X}_{t+\tau}), \tag{25}$$

where $\mathcal{H}_t$ represents the temporal semantic features at time step t, $\mathcal{T}$ represents the set of temporal scales, w_τ are learnable scale weights, and $\Psi(\cdot)$ computes temporal relationships between features at different time steps:

$$\Psi(\mathcal{X}_1, \mathcal{X}_2) = \phi(\varphi(\mathcal{X}_1), \psi(\mathcal{X}_2)) \cdot \gamma([\mathcal{X}_1 \| \mathcal{X}_2]), \tag{26}$$

where $\phi(\cdot)$ measures feature relevance, $\varphi(\cdot)$ and $\psi(\cdot)$ are feature transformation functions, $\gamma(\cdot)$ combines temporal information, and $\|$ denotes feature concatenation.

To enhance recognition robustness across different environments and occlusion conditions, we incorporate an adaptive feature aggregation mechanism:

$$\mathcal{R} = \sum_{t=1}^{T} \alpha_t \cdot \Omega(\mathcal{H}_t, \mathcal{C}_t), \tag{27}$$

where $\mathcal{R}$ represents the aggregated features, $\Omega(\cdot)$ is a non-linear transformation function, and α_t represents temporal importance weights computed through:

$$\alpha_t = \zeta(\delta(\mathcal{H}_t) + \beta(\mathcal{C}_t)), \tag{28}$$

where $\zeta(\cdot)$ normalizes importance weights, $\delta(\cdot)$ and $\beta(\cdot)$ transform temporal and contextual features respectively, and $\mathcal{C}_t$ captures environmental conditions. The activity type prediction is obtained through:

$$\mathcal{Y} = \rho(\theta(\mathcal{R})), \tag{29}$$

where $\mathcal{Y}$ represents the predicted probabilities over different activity types, $\theta(\cdot)$ transforms aggregated features into prediction scores, and $\rho(\cdot)$ normalizes the predictions.

This framework effectively combines environment-aware fusion features with modality-specific characteristics to identify distinct VR activity patterns. Semantic vision features capture detailed controller movements and user postures when line-of-sight is available, while semantic point cloud features maintain tracking during occlusions. This multi-modal approach enables robust activity recognition across various usage scenarios and environmental conditions.

4) VR Keystroke Recognition: VR-PCT introduces a collaborative keystroke recognition framework that leverages the semantic correlations between VR headset and controller dynamics through multi-modal fusion. This framework incorporates mmWave point cloud and vision data on the VR controller and headset to achieve precise keystroke identification.

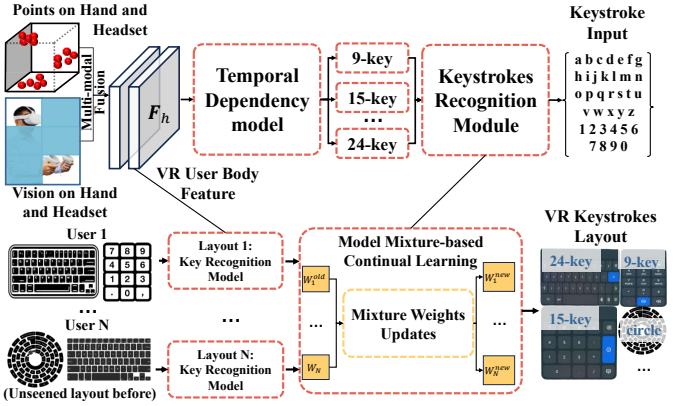


Fig. 7: VR Keystroke Recognition.

To construct a comprehensive representation encapsulating spatial-temporal keystroke dynamics, the proposed framework integrates multi-modal input features formulated as follows:

$$\mathcal{Z} = [\mathcal{F} \| \Delta\mathcal{P} \| \Delta\mathcal{V}], \tag{30}$$

where $\mathcal{Z}$ denotes the concatenated feature vector, comprising environment-aware fusion features $\mathcal{F}$, as well as relative motion features $\Delta\mathcal{P}$ and $\Delta\mathcal{V}$, which are extracted from point cloud and vision modalities, respectively. This fusion

mechanism ensures the preservation of both global and local keystroke-related motion cues.

To model the temporal dependencies in keystroke sequences, we propose a multi-scale temporal relation network:

$$\mathcal{G}_t = \sum_{s \in \mathcal{S}} \alpha_s \cdot \Theta(\mathcal{Z}_t, \mathcal{Z}_{t+s}), \tag{31}$$

where $\mathcal{G}_t$ represents the temporally aggregated semantic features at time step t, $\mathcal{S}$ denotes the set of temporal scales, α_s are learnable scale-dependent weights, and $\Theta(\cdot)$ quantifies temporal relationships between feature representations.

To enhance recognition robustness across various VR keyboard configurations, we introduce an adaptive feature integration mechanism that accounts for the distinct key distributions associated with different keyboard layouts. VR keyboards, e.g., 9-key, 15-key, and 24-key layouts, exhibit unique spatial arrangements, which in turn influence the corresponding VR semantic features. VR-PCT infers the specific keyboard type by analysing the keyboard distribution.

$$\mathcal{K} = \sum_{t=1}^{T} \beta_t \cdot \Psi(\mathcal{G}_t, \mathcal{L}_t), \tag{32}$$

where $\mathcal{K}$ denotes the integrated feature representation, $\Psi(\cdot)$ represents a non-linear transformation function, and β_t corresponds to the temporal importance weight, defined as:

$$\beta_t = \phi(\omega(\mathcal{G}_t) + \eta(\mathcal{L}_t)), \tag{33}$$

where $\phi(\cdot)$ normalizes the computed importance weights, while $\omega(\cdot)$ and $\eta(\cdot)$ extract temporal motion features and keyboard layout characteristics, respectively. The inclusion of $\mathcal{L}_t$, which encodes both the statistical key distribution and keystroke positions, ensures adaptability across different keyboard designs and spatial configurations, thereby improving the accuracy and robustness of VR keystroke recognition.

The keystroke identification process in VR-PCT leverages a multi-task learning framework that integrates both the inferred keyboard type and the detected keystroke positions to determine the final keyboard input:

$$[\mathcal{Y}_{key}, \mathcal{Y}_{conf}] = \rho(\theta(\mathcal{K}, \mathcal{L}_{type}, \mathcal{P}_{click})), \tag{34}$$

where $\mathcal{Y}_{key}$ represents the predicted keystroke probabilities, $\mathcal{Y}_{conf}$ denotes confidence estimations, $\mathcal{L}_{type}$ encodes the identified keyboard layout, and $\mathcal{P}_{click}$ is the detected keystroke location. The function $\theta(\cdot)$ transforms the integrated feature representation $\mathcal{K}$, the keyboard layout information $\mathcal{L}_{type}$, and the spatial keystroke position $\mathcal{P}_{click}$ into prediction scores.

To address previously unseen keyboard layouts and maintain model robustness across diverse VR input interfaces, we implement a model mixture-based continual learning approach for keyboard adaptation, as shown in the bottom part of Fig. 7. This method allows the system to continuously learn and adapt to new keyboard configurations while preserving performance on known layouts. The mixture model for keyboard recognition is represented as:

$$P(\mathcal{Y}_{key}|\mathcal{K}) = \sum_{i=1}^{M} \pi_i P_i(\mathcal{Y}_{key}|\mathcal{K}), \tag{35}$$

where $P(\mathcal{Y}_{key}|\mathcal{K})$ is the probability of keystroke prediction given the integrated feature representation $\mathcal{K}$, M is the number of keyboard layout mixture components, π_i are mixture weights corresponding to different keyboard models, and $P_i(\mathcal{Y}_{key}|\mathcal{K})$ are individual keyboard model probabilities. When a new keyboard layout is encountered, the system updates the mixture weights and individual models:

$$\pi_i^{new} = (1-\gamma)\pi_i^{old} + \gamma P_i(\mathcal{Y}_{key}|\mathcal{K}), \tag{36}$$

$$\theta^{new} = \theta^{old} + \xi \nabla_\theta \log P_i(\mathcal{Y}_{key}|\mathcal{K}), \tag{37}$$

where γ and ξ are the learning rates for mixture weights and model parameters respectively, and θ represents the keyboard recognition model parameters. This continual learning approach enables the system to adapt to new keyboard layouts based on newly collected keystroke data. The updated keystroke recognition module then incorporates the new keyboard configuration:

$$[\mathcal{Y}_{key}^{new}, \mathcal{Y}_{conf}^{new}] = \rho(\theta^{new}(\mathcal{K}, \mathcal{L}_{type}^{new}, \mathcal{P}_{click})). \tag{38}$$

This adaptive mechanism ensures that the model can recognize keystrokes from both known and previously unseen keyboard layouts, maintaining robust performance across a diverse range of VR input interfaces.

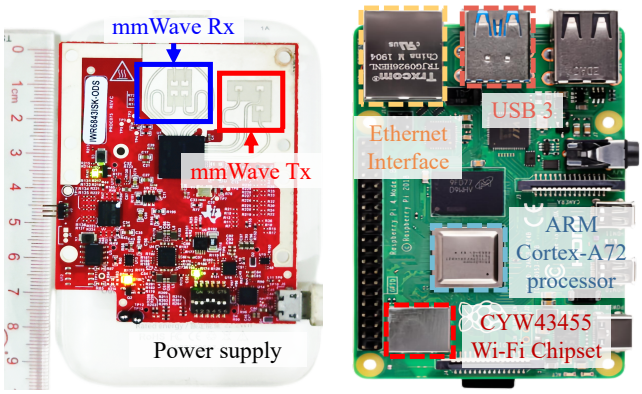
Integrating keyboard layout classification with positional data enables precise keystroke recognition across various VR keyboards. Our framework uses hierarchical temporal attention and adaptive feature fusion for robust keystroke pattern capture. Vision features provide trajectories in clear conditions, while point cloud features maintain tracking during occlusions. Joint optimization of classification and confidence estimation improves reliability in real-world VR environments.

V. Evaluation

A. Experimental Setup

In this section, we discuss the hardware implementation and experimental configuration of VR-PCT, emphasizing its performance in various real-world VR scenarios and alignment with the design principles.

Data Collection. VR-PCT employs a multi-modal sensing infrastructure to capture comprehensive data for VR semantic content recognition. As shown in Fig. 8, our platform integrates two key components. For mmWave point cloud collection, as illustrated in Fig. 8a, we utilize the TI IWR6843-ODS radar operating at 60 GHz with 4000.14 MHz bandwidth. Its 3TX and 4RX antenna array provides 0.75 cm distance resolution and 0.24 m/s velocity resolution, enabling precise spatial capture. For vision data, we utilize an Azure Kinect capturing RGB frames at 30 FPS with 1920×1080 resolution, featuring a 1-MP depth sensor, 7-microphone array, and IMU sensors for spatial tracking. Both sensors connect to our VR client device, as demonstrated in Fig. 8b, a Raspberry Pi with



(a) Detailed components of the mmWave radar device.

(b) Detailed components of the VR client for data collection.

Fig. 8: VR Semantic Data Collection Devices.

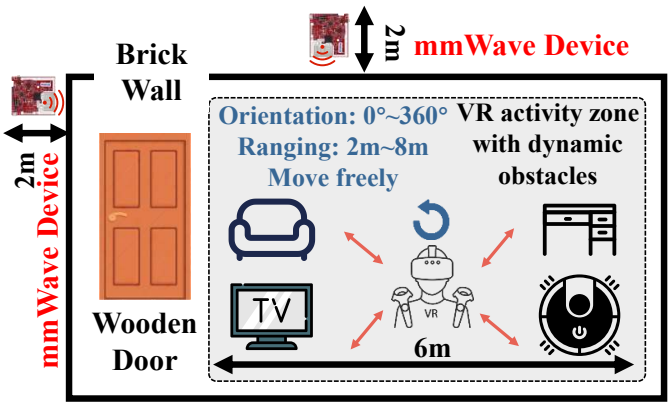


Fig. 9: Implementation Scenario.

ARM Cortex-A72 processor, CYW43455 Wi-Fi chipset, and Ethernet interface—which serves as our resource-constrained edge computing platform. This complementary setup facilitates the synchronized acquisition of point cloud and vision information crucial for our semantic-discriminative region selection and multi-modal recognition approach.

Our VR-PCT prototype's data collection components are portable, fitting into a 96-gram power bank measuring 8×8 cm. We utilized commercial VR devices as target platforms and involved 32 participants, including 16 males and 16 females aged 21-58, from whom we obtained informed consent before participation in IRB-approved studies (IRB 2021ZDSYLL089-P01). We accumulated 4,200 datasets totaling 14TB of mmWave point cloud and vision data, with 300 sets for each keystroke type. Each dataset contains 30 seconds of raw signal and point cloud data sampled at 15 FPS.

Implementation. Our experiments span multiple commercial VR platforms, including Meta Quest, Sony PlayStation VR, and HTC VIVE XR, evaluating VR-PCT across diverse hardware configurations in dynamic VR environments. As shown in Fig. 9, following standard VR safety protocols, users define a safe boundary area within the activity zone 2-8m from the mmWave device to prevent collisions with surrounding walls or furniture. In such scenarios, each VR-PCT deployment handles semantic recognition for its respective user. This activity zone consists of an open space where users move freely from multiple orientations (0° 360°), while the external environment contains dynamic obstacles such as robotic vacuum cleaners and other moving household devices. The setup incorporates brick walls and wooden doors to assess system penetration capabilities, creating realistic usage conditions that reflect typical VR deployment scenarios.

We utilize the Raspberry Pi as our VR client platform due to limited sensor access in commercial VR headsets, demonstrating VR-PCT's efficiency on computational capabilities more constrained than contemporary headsets like the Meta Quest 3 with its Snapdragon XR2 Gen 2. For the VR edge device, we utilize an Asus ROG laptop with Intel® Core™ i9-13980HX and NVIDIA GeForce RTX 4060 GPU for multi-modal fusion and semantic content recognition.

We evaluate VR-PCT on activity and keystroke recognition for model implementation utilizing our 14TB multi-modal dataset. The system processes mmWave point clouds with a 16×16 patch size for efficient partitioning. Input consists of point cloud sequences of 25 frames per keystroke recognition instance. We adhere to original training methodologies, including data augmentation, regularization strategies, and optimization parameters. Training spans 700 epochs with early stopping at 200 epochs to prevent overfitting, initializing validation loss at infinity, and updating when lower values are achieved to ensure optimal model convergence.

Evaluation Matrix. In our comprehensive evaluation, we compare VR-PCT against state-of-the-art vision transformers HIRI-ViT [25] and SpikingResformer [22] and point transformers Point Transformer V3 [14] and OneFormer3D [15]. Vision transformers require transmitting complete vision data, causing excessive bandwidth consumption and privacy concerns. Point transformers process entire point clouds with noise without regard to VR semantics. In contrast, VR-PCT locates semantic regions with the lightweight point cloud, enabling selective transmission of semantic data to the edge, reducing overhead while improving recognition performance.

Our evaluation examines VR-PCT's performance across multiple dimensions. We first present its effectiveness on three VR platforms under four occlusion configurations in Subsection V-B, comparing against state-of-the-art methods. In Subsection V-C, we evaluate our semantic-discriminative region selection approach versus existing transformer techniques. Subsection V-E analyzes VR-PCT's accuracy in categorizing diverse VR activities from chatting to gaming, with further processing for keystroke recognition when applicable. Subsection V-F evaluates keystroke recognition precision across common VR keyboard layouts, including 9-key, 15-key, and 24-key configurations for all alphanumeric characters, while VR-PCT's mixture-based continual learning approach enables extension to additional keyboard types. Subsection V-D tests our edge-client collaborative approach under various network conditions, including wired and WiFi connections with congestion. Finally, Subsection V-G examines VR-PCT's privacy protection capabilities through selective semantic transmission compared to conventional full data transmission methods.

TABLE II: Performance comparison of VR-PCT against SOTA methods across four occlusion scenarios. Metrics include transmission requirements, recognition accuracy, computational complexity, and parameter count. Gray cells show the highest accuracy configurations; underlined values indicate the most efficient models. ***Note:** Point transformer models report artificially low bandwidth by excluding essential visual data transmission required in practical VR systems [1], [41]–[43].

Occlusion	Model	Transmission Data (Mbps)	VR Application Type Recognition			VR Keystrokes Recognition		
			Acc. (%)	FLOPs (G)	Params (K)	Acc. (%)	FLOPs(G)	Params (K)
No Occlusion	Baseline (Attention mechanism)	17.8	82.1	4.2	2530	69.5	1.4	1372
	Point Transformer V3 [14]	0.6*	87.2	1.8	1239	74.7	0.6	568
	OneFormer3D [15]	0.7*	88.5	1.6	1106	77.2	0.6	524
	SpikingResformer [22]	11.4	92.6	2.3	1428	82.6	0.7	758
	HIRI-ViT [25]	12.9	93.7	2.6	1680	85.4	0.8	812
	VR-PCT (K'=32, K=450)	**6.5**	**93.2**	**0.6**	**367**	**82.1**	**0.2**	**186**
	VR-PCT (K'=64, K=680)	**7.8**	**95.8**	**0.9**	**693**	**85.3**	**0.3**	**215**
	VR-PCT (K'=96, K=720)	**8.7**	**97.6**	**1.5**	**986**	**92.8**	**0.5**	**297**
Wood Occlusion	Baseline (Attention mechanism)	16.7	78.4	3.8	2398	66.1	1.2	1263
	Point Transformer V3 [14]	0.3*	84.1	1.6	1123	71.4	0.5	525
	OneFormer3D [15]	0.4*	85.7	1.4	1031	73.3	0.5	481
	SpikingResformer [22]	10.4	89.1	2.1	1381	77.9	0.7	654
	HIRI-ViT [25]	11.7	90.4	2.4	1531	80.8	0.7	710
	VR-PCT (K'=32, K=450)	**5.7**	**90.9**	**0.4**	**361**	**79.6**	**0.2**	**183**
	VR-PCT (K'=64, K=680)	**6.9**	**93.6**	**0.8**	**679**	**82.9**	**0.3**	**208**
	VR-PCT (K'=96, K=720)	**7.3**	**94.8**	**1.3**	**875**	**88.7**	**0.4**	**273**
Brick Occlusion	Baseline (Attention mechanism)	17.3	75.8	4.0	2478	63.5	1.3	1283
	Point Transformer V3 [14]	0.5*	81.6	1.7	1178	69.2	0.6	554
	OneFormer3D [15]	0.6*	83.4	1.5	1112	71.5	0.6	509
	SpikingResformer [22]	12.4	87.9	2.2	1454	75.2	0.7	683
	HIRI-ViT [25]	13.7	86.2	2.5	1582	74.1	0.8	748
	VR-PCT (K'=32, K=450)	**6.7**	**87.7**	**0.5**	**378**	**76.4**	**0.2**	**189**
	VR-PCT (K'=64, K=680)	**7.9**	**90.5**	**0.9**	**703**	**79.6**	**0.3**	**221**
	VR-PCT (K'=96, K=720)	**8.3**	**92.3**	**1.4**	**923**	**84.9**	**0.5**	**281**
Combined Occlusion	Baseline (Attention mechanism)	16.5	72.3	3.9	2318	60.7	1.1	1223
	Point Transformer V3 [14]	0.2*	78.9	1.5	1083	66.8	0.5	513
	OneFormer3D [15]	0.3*	80.8	1.3	1074	68.4	0.5	468
	SpikingResformer [22]	9.4	85.3	2.0	1375	71.1	0.6	628
	HIRI-ViT [25]	10.7	83.6	2.3	1487	70.3	0.7	656
	VR-PCT (K'=32, K=450)	**4.7**	**84.1**	**0.4**	**339**	**72.8**	**0.1**	**172**
	VR-PCT (K'=64, K=680)	**5.9**	**87.2**	**0.7**	**645**	**75.7**	**0.2**	**198**
	VR-PCT (K'=96, K=720)	**6.3**	**89.4**	**1.2**	**857**	**79.2**	**0.3**	**256**

B. Overall Performance

This section presents comprehensive experiments evaluating VR-PCT across various real-world VR environments, ranging from unobstructed scenarios to those with multiple occlusion barriers and different point cloud densities.

Table II demonstrates VR-PCT's superior performance compared to state-of-the-art visual and point cloud transformer architectures across various occlusion scenarios. In unobstructed environments, our approach with configuration K'=96, K=720 achieves 97.6% accuracy in application type recognition while requiring 8.7 Mbps transmission, less than HIRI-ViT at 12.9 Mbps and SpikingResformer at 11.4 Mbps. For keystroke recognition, VR-PCT achieves 92.8% accuracy, surpassing HIRI-ViT at 85.4% and SpikingResformer at 82.6% with lower computational requirements at 1.5G FLOPs. Reducing Top-K values to K'=32, K=450 decreases transmission to 6.5 Mbps and computational overhead to 0.6G FLOPs while maintaining competitive 93.2% application recognition accuracy.

VR-PCT's robustness becomes evident in challenging environments with strategic parameter adaptation. Under wood occlusion, VR-PCT maintains 94.8% application recognition accuracy with only 7.3 Mbps transmission, compared to HIRI-ViT's 90.4% accuracy requiring 11.7 Mbps. In combined occlusion scenarios, VR-PCT delivers 89.4% application recognition accuracy and 79.2% keystroke recognition accuracy, outperforming SOTA methods while requiring 6.3 Mbps bandwidth, 41% less than vision transformers. Lower Top-K configurations maintain remarkable efficiency under severe occlusion with only 4.7 Mbps transmission and 0.4G FLOPs while achieving 84.1% accuracy.

The ablation study results in Table III highlight the critical contribution of each VR-PCT component to computational and transmission overhead. Without attention score calculation, the model incurs the highest computational burden with 6.2G FLOPs and 3720K parameters alongside the heaviest transmission load at 11.2 Mbps, achieving only 68.7% application recognition and 48.1% keystroke recognition accuracy. VR semantic region localization reduces computational costs to 4.6G FLOPs and 2819K parameters while improving accuracy to 86.3% and 77.6% and decreasing bandwidth to 10.4 Mbps. Semantic-discriminative region selection enhances computational efficiency with 2.7G FLOPs and 1803K parameters, achieving 93.1% and 82.5% accuracy while reducing transmission to 6.8 Mbps. Geometry-guided feature augmentation demonstrates cost reduction with 1.8G FLOPs and 1248K parameters, improving accuracy to 93.9% and 85.3% with

TABLE III: The ablation study shows each VR-PCT component's contribution to system performance and corresponding overhead. Results quantify how attention score calculation, semantic region localization, discriminative region selection, and geometry-guided augmentation affect transmission efficiency, recognition accuracy, and computational requirements.

Model	Transmission Data (Mbps)	VR Application Type Recognition			VR Keystrokes Recognition		
		Acc. (%)	FLOPs (G)	Params (K)	Acc. (%)	FLOPs (G)	Params (K)
VR-PCT (w/o calculate the attention score)	11.2	68.7	6.2	3720	48.1	1.8	2105
VR-PCT (w/o VR semantic region localization)	10.4	86.3	4.6	2819	77.6	1.6	1874
VR-PCT (w/o semantic-discriminative regions selection)	6.8	93.1	2.7	1803	82.5	1.2	1063
VR-PCT (w/o geometry-guided feature augmentation)	2.1	93.9	1.8	1248	85.3	0.7	487
VR-PCT (K'=96, K=720)	**3.3**	**97.6**	**1.5**	**986**	**95.7**	**0.5**	**297**

bandwidth reduction to 2.1 Mbps. The complete VR-PCT framework achieves performance with 97.6% and 95.7% accuracy, utilizing only 1.5G FLOPs and 986K parameters at 3.3 Mbps, demonstrating our edge-client architecture eliminates redundancy while enhancing recognition.

C. Performance of Semantic-discriminative Regions Selection.

Fig. 10 illustrates the semantic-discriminative regions selected by VR-PCT compared to state-of-the-art Vision and Point Transformer approaches. Vision Transformer models like SpikingResformer [22] and HIRI-ViT [25] capture visual semantics effectively but require complete vision data transmission, resulting in substantial computational overhead and privacy vulnerabilities. Point Transformer models like Point Transformer V3 [14] and OneFormer3D [15] partially address transmission limitations by operating exclusively on geometric data but still process entire point clouds, degrading recognition accuracy due to noise interference while failing to differentiate the semantic importance of individual points in VR contexts. VR-PCT addresses these limitations through multi-modal fusion, leveraging complementary strengths of both visual and point cloud data while eliminating their respective weaknesses.

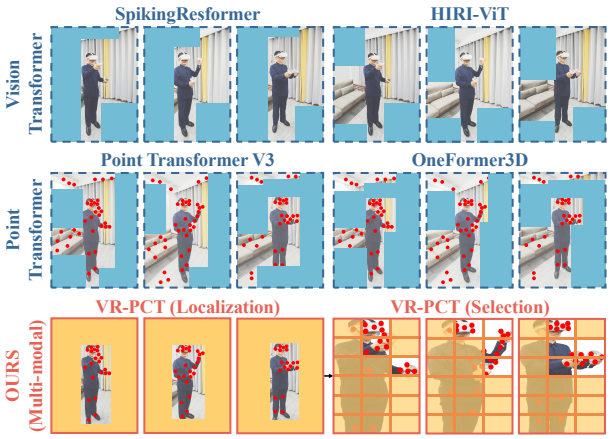


Fig. 10: Semantic-discriminative region selected by VR-PCT, Vision and Point Transformer SOTAs.

VR-PCT implements a two-stage approach. During the localization stage, it leverages mmWave point clouds to identify semantically significant regions containing user interactions. The selection stage then performs semantic-discriminative region selection by retaining the semantic vision data based on the initially localized areas. This hierarchical processing substantially reduces mmWave point cloud and visual data transmission requirements while enhancing semantic content recognition accuracy by focusing computational resources exclusively on the most informative regions. Eliminating redundant information across both modalities enables VR-PCT to achieve superior efficiency in computational demands and data transmission overhead while improving recognition accuracy.

D. Performance of Edge-client Collaboration

VR-PCT demonstrates significant advantages in edge-client collaboration through adaptive parameter adjustment and efficient semantic processing across diverse network conditions. Fig. 11 illustrates data transmission and semantic processing latency across various network connections between the Raspberry Pi client and VR edge. The horizontal axis represents different connections from congested Wi-Fi 5 to USB 3.2.VR-PCT adapts to varying network conditions by dynamically adjusting the K' and K parameters to balance recognition accuracy with transmission efficiency. Under optimal USB 3.2 connection, VR-PCT achieves 2.1ms data transmission versus 3.6ms for standard VR, representing a 41.7% reduction while maintaining 30.5ms semantic processing time. Under congested Wi-Fi 5 conditions, VR-PCT demonstrates remarkable adaptability by reducing parameters to K'=32, K=450, achieving 88.5% data transmission reduction compared to standard VR while maintaining responsive semantic processing at 31ms. This adaptive parameter selection enables VR-PCT to maintain real-time performance under adverse network conditions by reducing transmitted semantic data volume.

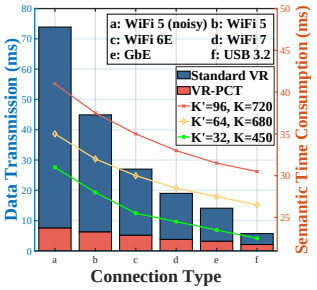


Fig. 11: Latency of Various Settings vs. Standard VR.

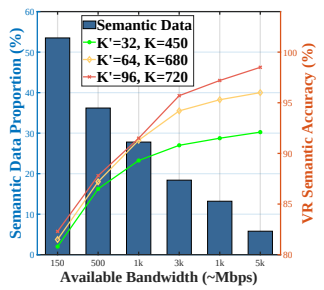


Fig. 12: Semantic Recognition Performance via Bandwidth.

Fig. 12 demonstrates VR-PCT's robust performance across varying bandwidth conditions, illustrating how the system maintains high semantic recognition accuracy even when employing reduced K' and K parameters to accommodate

network constraints. At a constrained bandwidth of 150 Mbps, the lightweight configuration, e.g., K'=32, K=450, achieves 80.8% accuracy with 53.5% semantic data proportion. As bandwidth increases, VR-PCT progressively increases parameters for enhanced accuracy, with semantic data requirements reducing to 5.8% at 5Gbps while achieving 92.1% accuracy with lightweight settings and 98.5% accuracy with the comprehensive model (K'=96, K=720). VR-PCT sustains high semantic recognition accuracy consistently above 80% across network conditions. Beyond computational FLOPs, we evaluate VR-PCT's impact on VR device battery consumption through data transmission analysis. The blue bars represent the proportion of semantic data that VR-PCT transmits relative to total VR information, reflecting its impact on battery consumption. Compared to traditional approaches requiring complete data transmission, VR-PCT reduces wireless communication power consumption by up to 94.2% under optimal conditions and 46.5% under constrained conditions.

E. VR Activity Type Recognition

We conduct comprehensive experiments comparing VR-PCT with state-of-the-art methods for VR activity recognition. As Fig. 13 shows, VR-PCT consistently outperforms all baseline models across varying computational complexities. With FLOPs below 0.5G per frame, VR-PCT achieves superior performance by effectively leveraging mmWave point cloud information without extensive visual data processing. Point transformers deliver reasonable accuracy in this low-computation regime, while vision transformers struggle with limited resources. As computational capacity increases beyond 0.5G FLOPs, VR-PCT maintains its advantage by incorporating visual semantic information while preserving point cloud benefits. At 1.5G FLOPs per frame, VR-PCT achieves 97.6% accuracy while reducing computational overhead by 76.9% compared to Point Transformer V3, enabling real-time processing with a 35ms inference time.

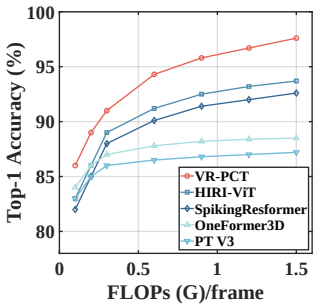


Fig. 13: Activity Type Recognition Performance vs. SOTA.

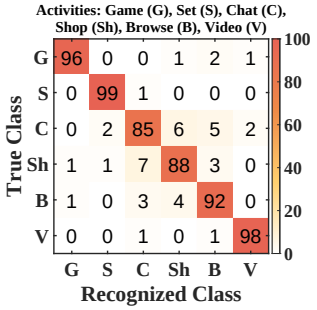


Fig. 14: Confusion Matrix of Activity Type Recognition.

Fig. 14 presents the confusion matrix of VR-PCT's activity recognition performance. System Setting and Game activities achieve the highest accuracies at 99% and 96%, respectively, due to their distinctive interaction patterns. Video Watching demonstrates 98% accuracy, characterized by minimal interaction followed by physical stillness. More challenging to distinguish are Browsing with 92% accuracy and Shopping with 88% accuracy, which share similar controller gestures but differ in Shopping's extended keyboard input sequences. Chatting exhibits the lowest accuracy at 85%, as its mixture of interactions creates confusion with both Shopping and Browsing activities. Despite these challenges, VR-PCT identifies fine-grained semantic differences in VR interactions while maintaining recognition accuracy across various activities.

F. VR Reystrokes Recognition

We conduct comprehensive experiments comparing VR-PCT with state-of-the-art methods for VR activity recognition. As Fig. 13 shows, VR-PCT consistently outperforms all baseline models across varying computational complexities. With FLOPs below 0.5G per frame, VR-PCT achieves superior performance by effectively leveraging mmWave point cloud information without extensive visual data processing. Point transformers deliver reasonable accuracy in this low-computation regime, while vision transformers struggle with limited resources. As computational capacity increases beyond 0.5G FLOPs, VR-PCT maintains its advantage by incorporating visual semantic information while preserving point cloud benefits. At 1.5G FLOPs per frame, VR-PCT achieves 97.6% accuracy while reducing computational overhead by 76.9% compared to Point Transformer V3, enabling real-time processing with a 35ms inference time.

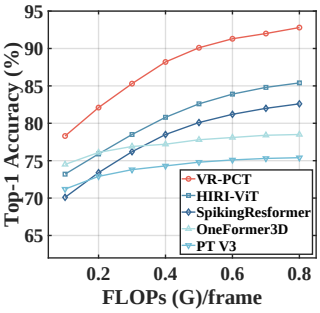


Fig. 15: Keystrokes Recognition Performance vs. SOTA.

Fig. 16: Illustration of VR Keystroke Layouts.

VR environments employ diverse keyboard layouts that maintain conventional input paradigms while accommodating specific interaction requirements. Fig. 16 illustrates three predominant keyboard configurations found in VR applications. The 24-key layout adopts a standard QWERTY arrangement primarily utilized for alphanumeric text entry in messaging. The 15-key layout features a numerical pad with mathematical operators arranged in a grid-like structure, commonly implemented for calculation tasks. The 9-key layout presents a compact 3×3 grid with an additional "0" key at the bottom, similar to telephone keypads, utilized for numeric input. The spatial distribution of keystroke positions creates distinctive interaction patterns that remain consistent throughout user sessions, enabling accurate keyboard layout recognition through pattern analysis rather than template matching. VR-PCT identifies keyboard types through user keystroke distribution patterns, enabling adaptability to diverse input interfaces.

Fig. 17 presents a detailed analysis of VR-PCT's keystroke recognition performance across all alphanumeric keys from a to z and 0 to 9. We evaluated VR-PCT on multiple keyboard

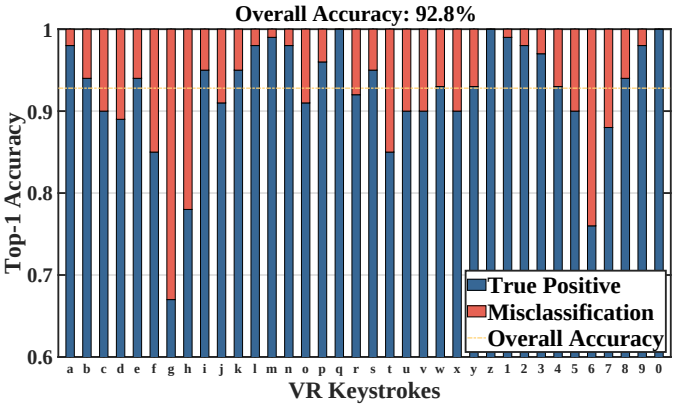


Fig. 17: Top-1 Accuracy of Keystrokes Recognition.

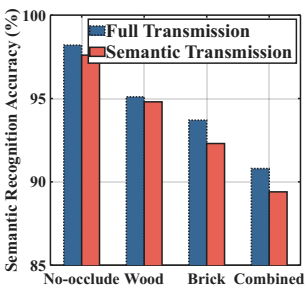


Fig. 18: Privacy Protection vs. Full Data Transmission.

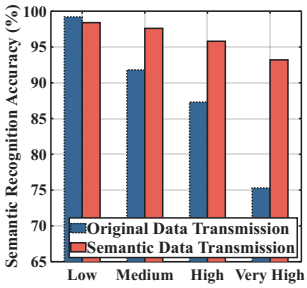


Fig. 19: Performance via Privacy Protection Level.

layouts commonly utilized in VR environments, including 9-key, 15-key, and 24-key configurations. The results demonstrate that VR-PCT achieves an overall top-1 accuracy of 92.8% across all keystrokes. Individual key recognition performance varies based on spatial positioning, with peripheral keys like q, z, and 0 exhibiting near-perfect recognition rates at 100% due to their distinct spatial characteristics. Keys in more densely populated regions of the keyboard, such as g at 67% and 6 at 76%, present significant recognition challenges due to potential confusion with adjacent keys. Despite these variations, VR-PCT maintains robust performance across the entire keyboard layout, demonstrating its effectiveness in capturing the semantics of VR interactions via multi-modal fusion.

G. VR Privacy Protection

Fig. 18 presents VR-PCT's privacy protection capabilities compared to conventional full data transmission across four occlusion scenarios. VR-PCT's semantic transmission maintains competitive performance while providing substantial privacy benefits through selective data transmission. In unobstructed environments, VR-PCT achieves 97.6% accuracy compared to full transmission's 98.2% (0.6% trade-off). Under wood occlusion, VR-PCT maintains 94.8% accuracy against 95.1%, while brick occlusion shows 92.3% versus 93.7%. In combined occlusion scenarios, VR-PCT sustains 89.4% compared to 90.8%. Across all scenarios, performance degradation remains below 1.5%, demonstrating VR-PCT's ability to provide robust privacy protection without significantly compromising recognition accuracy. These results underscore the effectiveness of selective semantic region transmission in maintaining high-quality VR performance while substantially reducing sensitive information exposure.

Fig. 19 presents VR-PCT's performance across varying privacy protection levels, demonstrating semantic data transmission's effectiveness compared to conventional original data transmission methods. At low privacy protection levels, original data transmission achieves 99.2% accuracy compared to VR-PCT's 98.4%. As privacy requirements intensify, VR-PCT maintains 97.6% accuracy at medium levels while original transmission drops to 91.8%. Under high privacy protection, VR-PCT achieves 95.8% accuracy versus the original transmission's 87.3%. Under very high privacy protection scenarios, VR-PCT sustains 93.2% accuracy while original data transmission degrades significantly to 75.3%, demonstrating a 17.9% superiority. VR-PCT's semantic region identification enables effective recognition with limited privacy-constrained data transmission, whereas conventional methods suffer when privacy-sensitive data cannot be transmitted. These results underscore VR-PCT's capability to maintain robust semantic recognition performance while providing enhanced privacy protection across varying security requirements.

VI. Conclusion

This paper presents VR-PCT, the first multi-modal transformer framework for VR semantic content recognition that enhances performance while reducing redundancy through edge-client collaboration. The key insight leverages semantic region detection on VR clients utilizing mmWave point clouds to enable selective transmission of semantically relevant data to the edge. By fusing reduced point cloud and visual modalities, VR-PCT demonstrates superior recognition accuracy while significantly reducing computational requirements. Our evaluation shows VR-PCT achieving 97.6% accuracy in VR application type recognition while maintaining robust 89.4% accuracy under challenging combined occlusions and reducing transmission data by up to 81.5% compared to vision and point transformer SOTA approaches. These results validate VR-PCT's effectiveness in optimizing cross-modal data filtering while preserving critical semantic information across diverse VR environments. Beyond VR applications, VR-PCT's client-edge collaboration architecture and cross-modal semantic principles are able to extended to autonomous systems, augmented reality environments, and smart environments for privacy-aware multi-modal processing. We release a 14TB dataset of mmWave and vision data under various VR scenarios to facilitate further research in this promising direction.

References

[1] D. Martin, S. Malpica, D. Gutierrez, B. Masia, and A. Serrano, "Multimodality in vr: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, sep 2022.

[2] Y. Wang, Z. Su, N. Zhang, R. Xing, D. Liu, T. H. Luan, and X. Shen, "A survey on metaverse: Fundamentals, security, and privacy," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 319–352, 2023.

[3] X. Qiao, P. Ren, S. Dustdar, L. Liu, H. Ma, and J. Chen, "Web ar: A promising future for mobile augmented reality—state of the art, challenges, and insights," *Proceedings of the IEEE*, vol. 107, no. 4, pp. 651–666, 2019.

[4] J. Suo, W. Zhang, J. Gong, X. Yuan, D. J. Brady, and Q. Dai, "Computational imaging and artificial intelligence: The next revolution of mobile vision," *Proceedings of the IEEE*, vol. 111, no. 12, pp. 1607–1639, 2023.

[5] L. Mei, R. Liu, Z. Yin, Q. Zhao, W. Jiang, S. Wang, S. Wang, K. Lu, and T. He, "mmspyvr: Exploiting mmwave radar for penetrating obstacles to uncover privacy vulnerability of virtual reality," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 8, no. 4, Nov. 2024.

[6] J. Zhang, R. Xi, Y. He, Y. Sun, X. Guo, W. Wang, X. Na, Y. Liu, Z. Shi, and T. Gu, "A survey of mmwave-based human sensing: Technology, platforms and applications," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 4, pp. 2052–2087, 2023.

[7] W. Li, R. Liu, S. Wang, D. Cao, and W. Jiang, "Egocentric human pose estimation using head-mounted mmwave radar," in *Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems*, 2023, pp. 431–444.

[8] Y. Qiu, J. Zhang, Y. Chen, J. Zhang, and B. Ji, "Radar2: Passive spy radar detection and localization using cots mmwave radar," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 2810–2825, 2023.

[9] S. Wang, D. Cao, R. Liu, W. Jiang, T. Yao, and C. X. Lu, "Human parsing with joint learning for dynamic mmwave radar point cloud," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 1, pp. 1–22, 2023.

[10] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, "A survey on vision transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 87–110, 2023.

[11] Y. Hu, Y. Cheng, A. Lu, Z. Cao, D. Wei, J. Liu, and Z. Li, "Lf-vit: Reducing spatial redundancy in vision transformer for efficient image recognition," *arXiv preprint arXiv:2402.00033*, 2024.

[12] Y. Li, J. Peng, J. Ye, Y. Zhang, F. Xu, and Z. Xiong, "Nlost: Non-line-of-sight imaging with transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 13 313–13 322.

[13] L. Zhong, X. Chen, C. Xu, Y. Ma, M. Wang, Y. Zhao, and G.-M. Muntean, "A multi-user cost-efficient crowd-assisted vr content delivery solution in 5g-and-beyond heterogeneous networks," *IEEE Transactions on Mobile Computing*, vol. 22, no. 8, pp. 4405–4421, 2023.

[14] X. Wu, L. Jiang, P.-S. Wang, Z. Liu, X. Liu, Y. Qiao, W. Ouyang, T. He, and H. Zhao, "Point transformer v3: Simpler faster stronger," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 4840–4851.

[15] M. Kolodiazhnyi, A. Vorontsova, A. Konushin, and D. Rukhovich, "Oneformer3d: One transformer for unified point cloud segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2024, pp. 20 943–20 953.

[16] T. Feng, R. Quan, X. Wang, W. Wang, and Y. Yang, "Interpretable3d: An ad-hoc interpretable classifier for 3d point clouds," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 2, pp. 1761–1769, Mar. 2024.

[17] N. Sabri, B. Chen, A. Teoh, S. P. Dow, K. Vaccaro, and M. Elsherief, "Challenges of moderating social virtual reality," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2023.

[18] Y. Liu, S. Jia, C. K. Yiu, W. Park, Z. Chen, J. Nan, X. Huang, H. Chen, W. Li, Y. Gao *et al.*, "Intelligent wearable olfactory interface for latency-free mixed reality and fast olfactory enhancement," *Nature Communications*, vol. 15, no. 1, p. 4474, 2024.

[19] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, "Cswin transformer: A general vision transformer backbone with cross-shaped windows," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2022, pp. 12 124–12 134.

[20] S. Chen, P. Sun, Y. Song, and P. Luo, "Diffusiondet: Diffusion model for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2023, pp. 19 830–19 843.

[21] R. Herzig, E. Ben-Avraham, K. Mangalam, A. Bar, G. Chechik, A. Rohrbach, T. Darrell, and A. Globerson, "Object-region video transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2022, pp. 3148–3159.

[22] X. Shi, Z. Hao, and Z. Yu, "Spikingresformer: Bridging resnet and vision transformer in spiking neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 5610–5619.

[23] C.-Y. Wu, Y. Li, K. Mangalam, H. Fan, B. Xiong, J. Malik, and C. Feichtenhofer, "Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 13 587–13 597.

[24] J. Yang, X. Dong, L. Liu, C. Zhang, J. Shen, and D. Yu, "Recurring the transformer for video action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 14 063–14 073.

[25] T. Yao, Y. Li, Y. Pan, and T. Mei, "Hiri-vit: Scaling vision transformer with high resolution inputs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 9, pp. 6431–6442, 2024.

[26] E.-C. Chen, P.-Y. Chen, I.-H. Chung, and C.-R. Lee, "Overload: Latency attacks on object detection for edge devices," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2024, pp. 24 716–24 725.

[27] Z. Yang, Z. Sarwar, I. Hwang, R. Bhaskar, B. Y. Zhao, and H. Zheng, "Can virtual reality protect users from keystroke inference attacks?" in *33rd USENIX Security Symposium (USENIX Security 24)*. Philadelphia, PA: USENIX Association, Aug. 2024, pp. 2725–2742.

[28] Y. Meng, Y. Zhan, J. Li, S. Du, H. Zhu, and X. Shen, "De-anonymizing avatars in virtual reality: Attacks and countermeasures," *IEEE Transactions on Mobile Computing*, vol. 23, no. 12, pp. 13 342–13 357, 2024.

[29] E. Wilson, A. Ibragimov, M. J. Proulx, S. D. Tetali, K. Butler, and E. Jain, "Privacy-preserving gaze data streaming in immersive interactive virtual reality: Robustness and user experience," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 5, pp. 2257–2268, 2024.

[30] J. Selva, A. S. Johansen, S. Escalera, K. Nasrollahi, T. B. Moeslund, and A. Clapés, "Video transformers: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 12 922–12 943, 2023.

[31] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2021, pp. 6836–6846.

[32] H. Yu, Z. Qin, J. Hou, M. Saleh, D. Li, B. Busam, and S. Ilic, "Rotation-invariant transformer for point cloud matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2023, pp. 5384–5393.

[33] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12 113–12 132, 2023.

[34] C. Slocum, Y. Zhang, N. Abu-Ghazaleh, and J. Chen, "Going through the motions: AR/VR keylogging from user head motions," in *32nd USENIX Security Symposium (USENIX Security 23)*. Anaheim, CA: USENIX Association, Aug. 2023, pp. 159–174.

[35] T. Zhang, Q. Zhang, K. Debattista, and J. Han, "Cross-modality distillation for multi-modal tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 7, pp. 5847–5865, 2025.

[36] T. Zhang, Q. Jiao, Q. Zhang, and J. Han, "Exploring multi-modal spatial–temporal contexts for high-performance rgb-t tracking," *IEEE Transactions on Image Processing*, vol. 33, pp. 4303–4318, 2024.

[37] J. Yin, J. Shen, R. Chen, W. Li, R. Yang, P. Frossard, and W. Wang, "Is-fusion: Instance-scene collaborative fusion for multimodal 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 14 905–14 915.

[38] Z. Luo, C. Zhou, L. Pan, G. Zhang, T. Liu, Y. Luo, H. Zhao, Z. Liu, and S. Lu, "Exploring point-bev fusion for 3d point cloud object tracking with transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 9, pp. 5921–5935, 2024.

[39] Z. Zhang, Y. Dong, Y. Liu, and L. Yi, "Complete-to-partial 4d distillation for self-supervised point cloud sequence representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2023, pp. 17 661–17 670.

[40] H. Liu, Z. Tan, Q. Chen, Y. Wei, Y. Zhao, and J. Wang, "Unified frequency-assisted transformer framework for detecting and grounding multi-modal manipulation," *International Journal of Computer Vision*, pp. 1–18, 2024.

[41] D. Yu, T. Dingler, E. Velloso, and J. Goncalves, "Object selection and manipulation in vr headsets: Research challenges, solutions, and success measurements," *ACM Comput. Surv.*, vol. 57, no. 4, Dec. 2024.

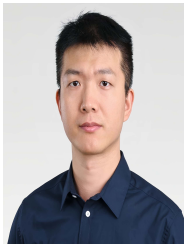
[42] A. Bentaleb, M. Lim, M. N. Akcay, A. C. Begen, S. Hammoudi, and R. Zimmermann, "Toward one-second latency: Evolution of live media streaming," *IEEE Communications Surveys & Tutorials*, pp. 1–1, 2025.

[43] J. van der Hooft, H. Amirpour, M. T. Vega, Y. Sanchez, R. Schatz, T. Schierl, and C. Timmerer, "A tutorial on immersive video delivery: From omnidirectional video to holography," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 2, pp. 1336–1375, 2023.

Luoyu Mei received the B.S. degrees from the Special Class for the Gifted Young in Southeast University, China. He is currently pursuing the Ph.D. degree of the joint program with the School of Computer Science and Engineering, Southeast University, and the Department of Computer Science, City University of Hong Kong. His research interests include the Wireless Sensing, Internet of Things, Wireless Network and mobile systems.

Shuai Wang is a Lecturer and Master's Supervisor at the School of Computer Science and Engineering, Southeast University. He received his Ph.D. from the Department of Computer Science at George Mason University in October 2023. His research focuses on mobile computing and the Internet of Things (IoT). He has published over thirty papers in prestigious journals and conferences, including MobiCom, SenSys, IJCAI, KDD, IPSN, and received the Best Paper Award at ICDCS 2018.

Ruofeng Liu is an assistant professor at Michigan State University. Before joining MSU, he was a research scientist at Robert Bosch. He received the Ph.D. degree from the University of Minnesota. Dr. Liu's research broadly lies in designing intelligent sensing and wireless communication systems, with a wide range of applications on the Internet of Things, human-computer interaction, and mobile networks. He published papers in top conferences including NSDI, MobiCom, IJCAI, CIKM, Sensys, SIGMETRICS, Ubicomp, and journals such as TON, TOSN, and COMST. He was granted several US patents. His research innovations were adopted by several companies including Nokia Bell Lab, Samsung Research, and Robert Bosch to solve real-world problems in logistics, automotive, and telecommunication.



Yun Cheng received the BSc and MSc degrees in computer science and technology from the Harbin Institute of Technology, Harbin, China, in 2012 and 2015, respectively, and the PhD degree from ETH Zürich, in 2022. His research interests include machine intelligence and efficient, applied machine learning.

Shuai Wang is a young chief professor in the School of Computer Science and Engineering at Southeast University. He has been serving as the head of the computer engineering department since Nov. 2022. He received his Ph.D. degree in the Department of Computer Science and Engineering at the University of Minnesota, Twin Cities in March 2017, under the guidance of Prof. Tian He. His research interests include artificial intelligence, data analytics, Internet of Things (IoT), cyber-physical systems, wireless networks and sensors and related areas.

Wenchao Jiang has been an Assistant Professor in ISTD, SUTD (Singapore) since September 2019. He received his PhD from the University of Minnesota, Twin Cities in 2019 under the supervision of Prof. Tian He. His research focuses on building direct communication between COTS heterogeneous wireless technologies in the PHY layer and its applications across the network stack. His research interests include Internet of Things (IoT), wireless communication, embedded systems, low-power sensor networks, mobile computing, and localization. He has published in top conferences and journals including MobiCom, SenSys, INFOCOM, SigComm, NSDI, TON, TMC, and TPDS.

Zhimeng Yin is an Assistant Professor at the Department of Computer Science at the City University of Hong Kong. He obtained his Ph.D. degree from the University of Minnesota (supervised by Prof. Tian He) in 2020. Before that, he received his bachelor's degree and master's degree (advised by Prof. Hongbo Jiang) from Huazhong University of Science and Technology in 2011 and 2014.

Tian He (Fellow, IEEE) received the Ph.D. degree in engineering from the Department of Computer Science, Hong Kong University of Science and Technology, Hong Kong, in 2008. He is currently a Full Professor with the Department of Computer Science and Engineering, University of Minnesota-Twin Cities, Minneapolis, MN, USA. He is the author and coauthor of over 280 papers in premier network journals and conferences with over 33,000 citations (H-Index 84). His research interests include wireless networks, networked sensing systems, cyber-physical systems, the Internet of Things, and distributed systems in general. Dr. He is the recipient of the NSF CAREER Award in 2009, the McKnight Land-Grant Chaired Professorship in 2011, the George W. Taylor Distinguished Research Award in 2015, the China NSF Outstanding Overseas Young Researcher I and II in 2012 and 2016, and eight best paper awards in international conferences, including MobiCom, SenSys, and ICDCS. He has served as a few general/program chair positions in international conferences and on many program committees and also has been an Editorial Board Member for six international journals, including ACM Transactions on Sensor Networks, IEEE Transactions on Computers, and IEEE/ACM Transactions on Networking. He is a Fellow of ACM.