Recognizing Actions from Robotic View for Natural Human-Robot Interaction

Ziyi Wang[1] Peiming Li[1] Hong Liu[1] Zhichao Deng[2] Can Wang[3]
Jun Liu[4] Junsong Yuan[5] Mengyuan Liu[1†]

[1]State Key Laboratory of General Artificial Intelligence, Peking University, Shenzhen Graduate School
[2]Sun Yat-sen University [3]Kiel University [4]Lancaster University [5]State University of New York at Buffalo

{ziyiwang,lipeiming1001}@stu.pku.edu.cn hongliu@pku.edu.cn dengzhch3@mail2.sysu.edu.cn
canwang@pku.edu.cn j.liu81@lancaster.ac.uk jsyuan@buffalo.edu liumengyuan@pku.edu.cn

2025 IEEE/CVF International Conference on Computer Vision (ICCV) | 979-8-3315-8775-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICCV51701.2025.01319

Abstract

Natural Human-Robot Interaction (N-HRI) requires robots to recognize human actions at varying distances and states, regardless of whether the robot itself is in motion or stationary. This setup is more flexible and practical than conventional human action recognition tasks. However, existing benchmarks designed for traditional action recognition fail to address the unique complexities in N-HRI due to limited data, modalities, task categories, and diversity of subjects and environments. To address these challenges, we introduce ACTIVE (Action from Robotic View), a large-scale dataset tailored specifically for perception-centric robotic views prevalent in mobile service robots. ACTIVE comprises 30 composite action categories, 80 participants, and 46,868 annotated video instances, covering both RGB and point cloud modalities. Participants performed various human actions in diverse environments at distances ranging from 3m to 50m, while the camera platform was also mobile, simulating real-world scenarios of robot perception with varying camera heights due to uneven ground. This comprehensive and challenging benchmark aims to advance action and attribute recognition research in N-HRI. Furthermore, we propose ACTIVE-PC, a method that accurately perceives human actions at long distances using Multilevel Neighborhood Sampling, Layered Recognizers, Elastic Ellipse Query, and precise decoupling of kinematic interference from human actions. Experimental results demonstrate the effectiveness of ACTIVE-PC. Our code is available at: https://github.com/wangzy01/ACTIVE-Action-from-Robotic-View.

1. Introduction

A robust Natural Human-Robot Interaction (N-HRI) system can provide a more personalized and intelligent interaction experience by adapting to changes in user needs and evolving action patterns. It has a wide range of applications, such as in service robots [19], public safety surveillance [10], and medical diagnostics [3, 32]. N-HRI requires robots to have the ability to adapt in real-time, not only understanding human actions but also considering changes in human position (*e.g.*, walking, standing) and their own motion. Therefore, compared to traditional human action recognition, human action recognition in N-HRI involves handling more complex human-robot interaction scenarios.

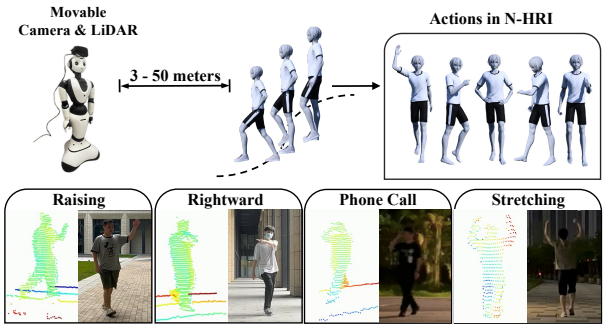


Figure 1. **Data Collection Process and Samples of the ACTIVE Dataset:** The movable camera and LiDAR capture human motion at distances ranging from 3 to 50 meters, with scene shaking introduced during the recording process. For visualization purposes, the samples have been scaled and cropped.

Compared to traditional human action recognition videos, videos captured from perception-centric robotic views in N-HRI scenarios exhibit diverse viewpoints, greater kinematic disturbances, and varied subject resolutions due to changing human-robot positions, robot movements, and uneven ground. While different from manipulator-mounted or end-effector perspectives common in industrial robots, this perception-centric viewpoint aligns closely with real-world applications, such as mobile assistants and patrol robots sensing humans at distances of 3–50 meters using forward-facing sensors. These challenges necessitate specialized methods tailored specifically to the

† Corresponding author: Mengyuan Liu (liumengyuan@pku.edu.cn)

2380-7504/25/$31.00 ©2025 IEEE
DOI 10.1109/ICCV51701.2025.01319

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 06:15:37 UTC from IEEE Xplore. Restrictions apply.

Table 1. A comparison of the ACTIVE dataset, designed for N-HRI, with other widely used general action and micro-action datasets. ACTIVE is the first large-scale action recognition dataset specifically focused on N-HRI.

		General Action					Fine-Grained Action			Actions from Robotic View
Dataset Attribute		MSR Action3D	Multiview 3D Event	NTU RGB+D 120	UAV Human	PA HMDB51	HOMAGE	iMiGUE	SMG	**ACTIVE (Ours)**
Action Recog.	# Videos	567	3,815	114,480	67,428	515	1,752	18,499	3,712	**46,868**
	# Subjects	20	10	106	119	-	27	72	40	**80**
	# Classes	10	10	120	155	51	75	32	16	**30**
Attribute Recog.	# Samples	×	×	×	22,476	515	×	×	×	**46,868**
Data Modality		D	RGB+D	RGB+D+IR	RGB+D+IR	RGB	RGB+IR	RGB	RGB+D	**RGB+Point Cloud**
Sensors		Kinect	Kinect	Kinect V2	Azure DK	NA	RGB Cam.	RGB Cam.	Kinect V2	**LiDAR, RGB Cam.**
Capturing Scenarios	# Sites	1	NA	3	45	NA	NA	NA	1	**6**
	Outdoor	×	×	✓	✓	×	×	×	×	✓
Shooting Time	Night	×	×	×	✓	×	×	×	×	✓
Composite Action		×	×	×	×	×	✓	✓	✓	✓
Dynamic Distance		×	×	×	×	×	×	×	×	✓
Platform Motion		×	×	×	✓	×	×	×	×	✓
Source		CVPR	ICCV	TPAMI	CVPR	TPAMI	CVPR	CVPR	IJCV	-
Year		2010	2013	2019	2021	2020	2021	2021	2023	-

unique characteristics of robotic views in N-HRI.

Existing work has already demonstrated the importance of using large, comprehensive, and challenging datasets to develop and evaluate state-of-the-art deep learning methods. However, in the field of human action understanding for N-HRI scenarios, current datasets [13, 21] have limitations in several aspects, including: 1) Limited Sample Size: Large-scale datasets are often critical for mitigating overfitting and enhancing the generalization ability of models developed on them. 2) Fixed Capture Modes: Most existing datasets rely on fixed relative positions and backgrounds, with subjects typically standing still and performing predefined actions or postures. In contrast, humans in N-HRI are usually mobile. 3) Constrained Capture Scenes: In real-world applications, N-HRI needs to function in diverse environments (*e.g.*, indoor, outdoor) and across various time periods (*e.g.*, day, night). However, samples in current datasets are often collected under similar conditions, which oversimplifies the challenges in real-world N-HRI scenarios. 4) Limited Capture Distances: N-HRI often involves frequent changes in the human-robot position, leading not only to significant motion blur but also to considerable resolution changes. Yet, in most existing datasets, humans only exhibit slow and subtle movements, with fixed capture distances. 5) Limited Data Modalities: In real-world applications, different types of sensors are often deployed to collect data under varying conditions. For instance, LiDAR sensors can be used for long-distance N-HRI understanding, while RGB cameras are typically used for close-range scenes that require rich detail. This highlights the importance of collecting multimodal data to analyze human action under different conditions. However, most existing human action datasets only provide traditional RGB video samples or depth modalities. Devices like Kinect suffer from limited capture range and poor performance under daylight conditions. These limitations emphasize the need for a more robust and diverse dataset that better reflects the dynamic and real-world complexities of N-HRI scenarios.

The aforementioned limitations of existing datasets clearly highlight the need for a larger, more challenging, and comprehensive dataset for human action analysis for N-HRI. To this end, we introduce ACTIVE, a large-scale multimodal dataset in this field. Specifically, ACTIVE includes 30 action categories with disturbing motion (*e.g.*, Heading + Nodding), involving 80 participants and attribute information such as gender and clothing. It consists of a total of 46,868 video instances, captured using sensors including LiDAR and RGB cameras. During data capture, participants perform various human actions in different environments (indoor and outdoor, day and night) at varying capture distances (ranging from 3m to a long-range 50m). Additionally, we simulate the robot's motion by introducing movement to the camera platform. The use of different sensors allows our dataset to provide rich data modalities, including both RGB and point cloud data. Overall, ACTIVE serves as a more challenging dataset, encompassing action recognition video instances in N-HRI scenarios and introducing diverse kinematic disturbances, which can promote research on human action understanding in N-HRI settings. Tab. 1 presents a statistical comparison between ACTIVE and existing datasets. More details about ACTIVE will be discussed in Sec. 3.

As shown in Fig. 1, unlike general actions, ACTIVE incorporates variations in human-robot states (*e.g.*, changes in relative distance and position through disturbing motion like walking and camera platform movement). Recognizing actions from robotic view aims to understand human actions in N-HRI scenarios, with the goal of filtering out the influence of kinematic disturbances from the robot's perspective and recognizing human actions to guide N-HRI. The main challenges of ACTIVE are as follows: 1) At long distances (*e.g.*, 50 meters), human actions are subtle and rapid, making it even more difficult to capture the fine details of these actions. 2) In dynamic interaction scenarios, large-scale kinematic disturbances (such as the motion of the robot platform, environmental changes like lighting and obstacles, and the relative motion between the human and the robot) further complicate action recognition.

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 06:15:37 UTC from IEEE Xplore. Restrictions apply.

Recognizing human actions from a robotic view presents several challenges, particularly when dealing with long distances and the interference of robotic motion. Traditional methods often fail to capture local details across varying scales and cannot effectively differentiate subtle human actions from kinematic disturbances caused by robot platform movements. To address these issues, we propose ACTIVE-PC, a comprehensive framework that utilizes Multilevel Neighborhood Sampling (MNS), Layered Recognizers (LR), and Elastic Ellipse Query (EEQ). MNS ensures both global coverage and retention of critical local details, while LR decouple features at different layers to preserve both fine-grained human actions and global kinematic variations. EEQ introduces axis-specific adaptive scaling to handle varying interaction distances, enhancing action differentiation by adjusting to the spatial distribution of motion. Experimental results demonstrate that this method improves action recognition performance in N-HRI scenarios.

2. Related Work

2.1. Action Recognition Datasets

Early 3D action recognition datasets, such as MSR-Action3D [13], Multiview 3D Event [31], and NTU RGB+D [21], have played a crucial role in advancing action recognition technology. The MSR-Action3D dataset [13] primarily consists of sequences of gaming actions and was one of the first to incorporate depth information for action analysis. Many methods have been evaluated on this dataset, but due to the limited dataset size, recent approaches have achieved near saturation in accuracy. To address the limited sample size, the Multiview 3D Event dataset [31] employs multiple depth sensors to capture multi-view representations of the same action. The NTU RGB+D dataset [21], the first large-scale RGB+D action recognition dataset, features a broader range of samples, more intra-class variations, and more camera views, capturing RGB video, depth sequences, and skeleton data. In recent years, research has shifted towards fine-grained action recognition tasks. The iMiGUE dataset [15] aims to interpret athletes' emotional states by analyzing their micro-gestures, recording spontaneous micro-actions after a match. Similarly, the SMG dataset [4] focuses on both micro-gesture and emotion recognition, documenting participants' micro-actions while narrating stories. However, action recognition for N-HRI involves more complex and diverse interaction states. Existing datasets typically rely on fixed relative positions and backgrounds, which are not ideal for action recognition in real-world human-machine interactions. In contrast, ACTIVE is the first large-scale action recognition dataset focused on N-HRI, capturing a variety of actions with disturbing motions in complex interaction scenarios.

2.2. Action Recognition Methods

RGB-based methods, such as TSN [27], VideoMAE [24], UniFormer [11], and InternVideo [29], have achieved high accuracy in general action recognition. Similarly, various methods have been developed for general action recognition from point cloud videos [16, 22, 34]. To enhance the receptive field, P4Transformer [6] employs transformers to avoid explicit point tracking, although it does not fully encode the spatiotemporal structure. PST-Transformer [9] improves upon this by encoding features based on the spatiotemporal displacement between reference points and all points in the video. However, existing point cloud-based methods are typically limited to simple, single-action recognition and perform poorly on actions with disturbing motions. Our approach focuses on recognizing actions with disturbing motions in N-HRI, while remaining compatible with general action recognition.

Currently, there is limited research specifically addressing micro-action recognition from point cloud videos. In RGB-based methods, Yonetani et al. [33] extracted handcrafted features from both first-person and third-person videos for micro-action recognition. However, this approach requires both first-person and third-person videos, along with manual synchronization. Mi et al. [20] proposed a segment-level temporal pyramid for micro-action videos. Nonetheless, current micro-action recognition methods typically rely on fixed relative positions and backgrounds. In contrast, our approach adapts to different complex interaction scenarios, making it more suitable for action recognition in real-world human-machine interactions.

3. Dataset

3.1. The Action from Robotic View Dataset

ACTIVE is the first large-scale human behavior understanding dataset specifically designed for N-HRI scenarios (shown in Fig. 2), featuring data from 80 participants along with annotated attribute information. It covers six distinct environments (including both indoor and outdoor settings, under varying daylight and nighttime conditions) and includes 30 action (*e.g.*, Walking forward + Waving) categories. The dataset comprises 46,868 video instances, with an average duration of 3.2 seconds, totaling 41.7 hours. It includes two modalities: RGB and point cloud data. The point cloud data is captured using the HESAI AT128P LiDAR, with a horizontal field of view of 120°, a horizontal angular resolution of 0.1°, a vertical field of view of 25.4°, and a vertical resolution of 0.2°. The scanning frame rate is set at 10 Hz. The RGB video has a resolution of 1280×720 and a frame rate of 30 fps. Tab. 1 compares ACTIVE dataset with existing human action recognition datasets. Below are some key features of ACTIVE.

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 06:15:37 UTC from IEEE Xplore. Restrictions apply.

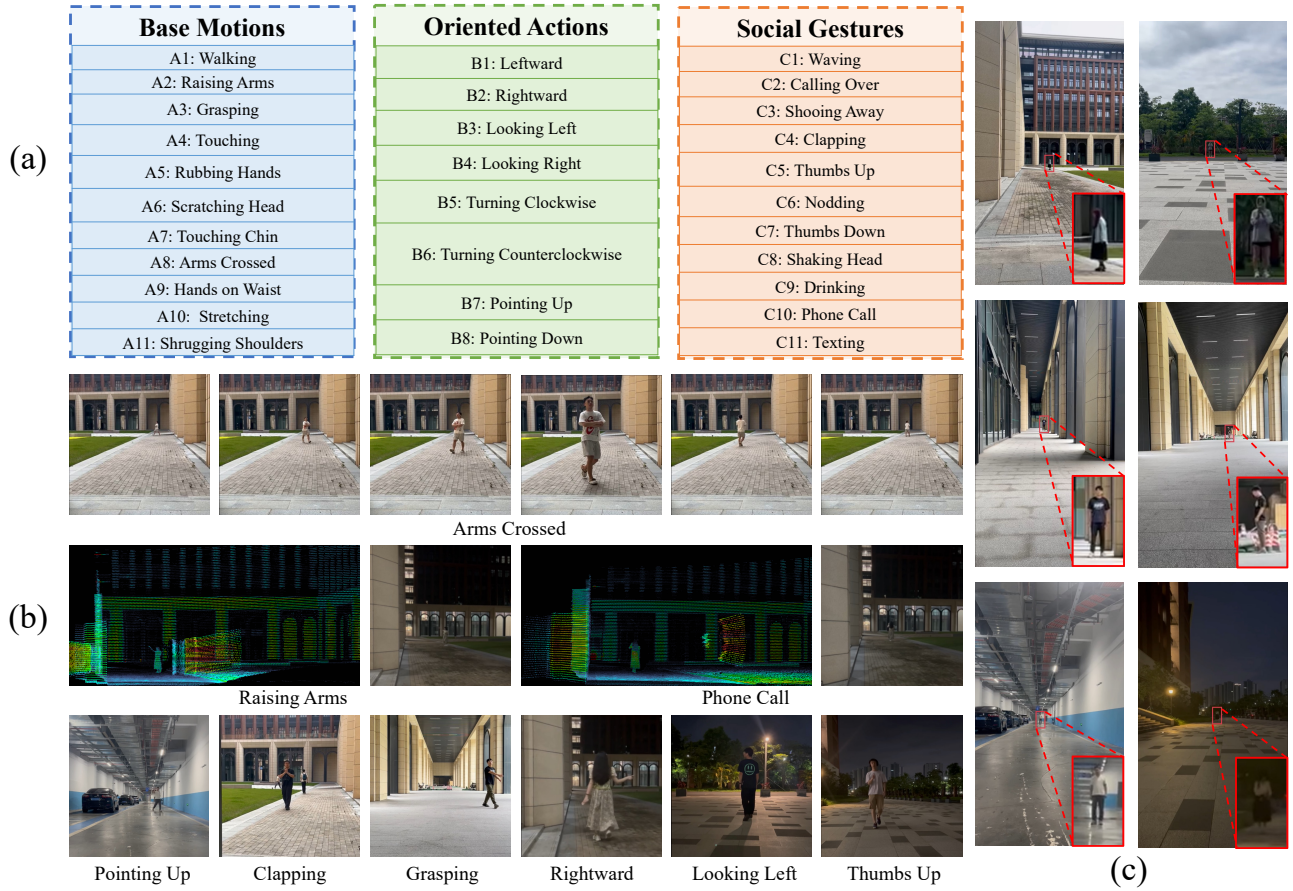


Figure 2. **Details of ACTIVE:** (a) Action Labels: Three categories are included: Base Motions, Oriented Actions, and Social Gestures. (b) The first row illustrates the significant variations in the captured results of the same action at different human-robot distances. The second row presents the raw point cloud captured by LiDAR and its corresponding RGB image. The point cloud modality retains more detailed geometric information, enabling long-range N-HRI understanding with strong robustness to interference. In contrast, the RGB modality complements this by providing rich texture and semantic information at close distances. The final row showcases additional action samples from our dataset. (c) Shooting Scenes: Six shooting scenes used in ACTIVE, covering both indoor and outdoor settings, as well as day and night conditions.

Diversity of Capture States and Behavior Patterns. Traditional human behavior datasets typically focus on fixed capture modes and distances, often emphasizing a single behavior. In contrast, human behavior understanding in N-HRI requires handling more complex human-robot interaction scenarios. Therefore, ACTIVE collects data from subjects performing various actions under different human-robot states. In different settings, participants are instructed to perform multiple natural actions at varying distances. The linear distance between participants and the data capture platform ranges from 3 meters to 50 meters, with additional movement and vibrations to simulate changes in the robot's state. Minor disturbances from vehicles and pedestrians are also allowed during recording.

More Relevant Data Modalities. Most existing human behavior datasets only provide traditional RGB video samples or depth modalities. While depth data can be converted into point clouds, it often lacks detailed geometric structure. Additionally, devices like Kinect suffer from short capture distances and poor performance under daylight conditions. In contrast, point cloud data captured by LiDAR retains richer geometric details, supports long-range N-HRI understanding, and is more robust to interference. The RGB modality complements this by offering rich texture and semantic information at close range. This underscores the importance of collecting diverse data modalities. Consequently, our varied data sources are designed to foster the development of robust models for human behavior analysis in complex N-HRI scenarios.

Multiple Human Behavior Understanding Tasks. To comprehensively analyze human behavior and actions in N-HRI, ACTIVE currently provides annotations for two primary tasks: human action recognition and human attribute recognition. Detailed annotations for the pose estimation task will be released at a later stage.

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 06:15:37 UTC from IEEE Xplore. Restrictions apply.

3.2. Benchmark Evaluations

Action Recognition. To ensure consistency in benchmarking, we define the Cross-Subject Evaluation protocol and use classification accuracy (percentage) as the evaluation metric. A total of 80 participants were divided into a training group and a testing group, consisting of 53 and 27 subjects, respectively.

Attribute Recognition. For attribute recognition, we use the same dataset partitioning strategy as for action recognition. The performance of attribute recognition is evaluated by measuring the classification accuracy for each attribute.

4. Methodologies

The ACTIVE dataset captures natural human-robot interactions from robotic view, encompassing both large-scale global motion disturbances and subtle local postural changes in the human body. In this work, we primarily aim to address two key challenges posed by the ACTIVE dataset: (1) At long distances, human actions are subtle and rapid, making it a complex task to recognize these fine movements. (2) In dynamic interaction scenarios, large-scale global motion disturbances further increase the difficulty of action recognition. To address these challenges in ACTIVE, we introduce ACTIVE-PC. In Sec. 4.1, we define the RARV task. In Sec. 4.2, we present the pipeline of ACTIVE-PC. In Sec. 4.3 and Sec. 4.4, we propose Multilevel Neighborhood Sampling (MNS) and Layered Recognizers (LR) to address challenge (1) and (2), and Elastic Ellipse Query to tackle challenge (2).

4.1. Task Definition of RARV

Consider an input video, $\mathbf{X} = \{I_1, I_2, \ldots, I_T\}$, where T is the length of the video. Each frame contains N points, and the video includes both human actions (*e.g.*, raising arms) and kinematic disturbances (*e.g.*, human-robot proximity and robot movement). Our goal is to filter out the influence of kinematic disturbances from the robot's perspective, in order to obtain the human action label Y.

4.2. Pipeline Overview

The pipeline of ACTIVE-PC is shown in Fig. 3. The input video $\mathbf{X} \in \mathbb{R}^{T\times N\times 3}$ first undergoes MNS, producing point clouds with different densities: $\mathbf{X_1} \in \mathbb{R}^{T/2\times N/4\times 3}$, $\mathbf{X_2} \in \mathbb{R}^{T/2\times N/16\times 3}$ and $\mathbf{X_3} \in \mathbb{R}^{T/2\times N/32\times 3}$. These point clouds, $\mathbf{X_1}$, $\mathbf{X_2}$ and $\mathbf{X_3}$ are then processed through EEQ for adaptive neighbor querying, enabling spatiotemporal tracking and differentiation of human actions and kinematic disturbances, resulting in feature maps $\mathbf{F}_1$, $\mathbf{F}_2$ and $\mathbf{F}_3$.

Low-density point clouds, which capture deeper features $\mathbf{F}_2$ and $\mathbf{F}_3$, are passed through the Kinematic Interpreter for human-robot state understanding. High-density point clouds, containing finer local details in $\mathbf{F}_1$, are input to the Action Recognizer for human-action recognition. $\mathbf{Y}_{human}$ and $\mathbf{Y}_{kinematic}$ are fused to obtain the fine-grained action score $\mathbf{Y}$. The process is formulated as follows:

$$\begin{aligned} \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3 &= \mathbf{MNS}(\mathbf{X}), \\ \mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3 &= \mathbf{PointTube}_{\mathbf{EEQ}}(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3), \\ \mathbf{Y}_{kinematic} &= \mathbf{Interpreter}_{\mathbf{kinematic}}(\mathbf{F}_2, \mathbf{F}_3), \\ \mathbf{Y}_{human} &= \mathbf{Recognizer}_{\mathbf{human}}(\mathbf{F}_1), \\ \mathbf{Y} &= \frac{(\mathbf{Y}_{human} + \mathbf{Y}_{kinematic})}{2}. \end{aligned} \tag{1}$$

4.3. Multilevel Neighborhood Sampling

To address the core challenges of the ACTIVE dataset, which include the precise perception of subtle human motions at long distances and the accurate decoupling of kinematic interference from human actions in dynamic environments, Multilevel Neighborhood Sampling (MNS, Sec. 4.3) and Layered Recognizers (LR, Sec. 4.4) work in tandem.

Traditional methods typically apply multi-scale or different sampling rates of farthest point sampling (FPS) directly to the entire point cloud. While this approach ensures globally uniform sampling, it often neglects the correlation between local details across different scales, which is critical in the ACTIVE dataset. In situations where the human-robot distance is large, high-density local sampling can better capture small yet distinguishable motion features.

To address this, we propose Multilevel Neighborhood Sampling (MNS), which builds hierarchical dependencies using local neighborhood constraints, ensuring both global coverage and retention of local details. The implementation of MNS is shown in Fig. 3. The black points represent the anchor points obtained through downsampling, while the red points represent the neighboring points. The process is as follows: First, for the input point cloud video data $\mathbf{X} \in \mathbb{R}^{T\times N\times 3}$, we apply FPS to sample N_1 points globally, obtaining the first layer of anchor points:

$$\mathbf{X}_1 = \mathrm{FPS}(\mathbf{X}; N_1). \tag{2}$$

For the l-th layer, instead of directly sampling globally, we construct a candidate set within the local neighborhood of the previous layer's sampled points $\mathbf{x}_{l-1,j}$ (where $j = 1, \ldots, N_{l-1}$). Using a radius r and a candidate number k, we define the spherical neighborhood as:

$$\mathcal{N}(\mathbf{x}_{l-1,j}) = \mathbf{x}_i : |\mathbf{x}_i - \mathbf{x}_{l-1,j}|_2 \leq r^{(l)}, |\mathcal{N}| \leq k^{(l)}. \tag{3}$$

Then, the union of the neighborhoods across all j is taken and duplicates are removed, forming the local candidate set:

$$\widehat{\mathbf{X}}_l = \bigcup \mathcal{N}(\mathbf{x}_{l-1,j}). \tag{4}$$

Through MNS, we can capture global structure in the sparse layers (counteracting the kinematic interference

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 06:15:37 UTC from IEEE Xplore. Restrictions apply.

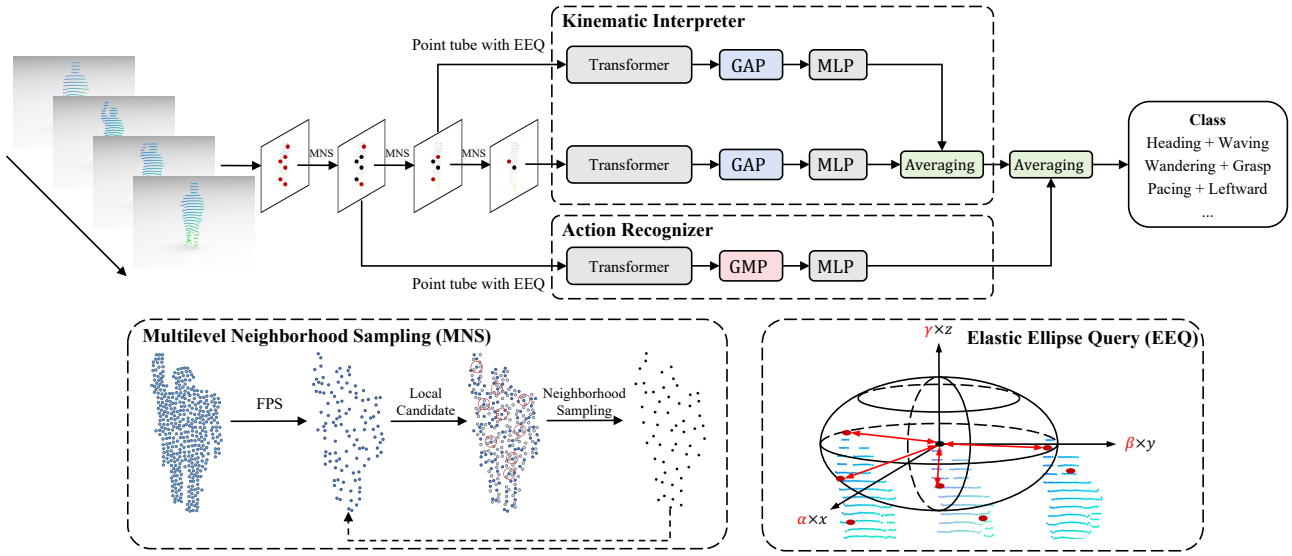


Figure 3. **ACTIVE-PC architecture.** First, the point cloud sequence undergoes Multilevel Neighborhood Sampling, followed by feature extraction using a point tube with Elastic Ellipse Query. The dense features are fed into Action Recognizer, while the sparse features are input into Kinematic Interpreter. The resulting human-action and kinematic scores are then fused to obtain the fine-grained action score.

caused by platform motion from robotic view), while fine-tuning the extraction of subtle human motion changes in the denser layers. This approach precisely aligns with the ACTIVE dataset's high demand for fine-grained local information when capturing complex human-robot dynamics.

4.4. Layered Recognizers

Considering the characteristics of the ACTIVE dataset, human action recognition requires not only capturing fine-grained, small-scale movements but also filtering out global kinematic variations caused by robot platform motion. To address these challenges, we design Layered Recognizers, which decouples features from different layers by sending them through specialized sub-networks. The core of the design includes two branches:

Kinematic Interpreter: For features obtained from larger sampling scales (*i.e.*, sparser layers, such as $\mathbf{F}_2$ and $\mathbf{F}_3$), this branch focuses on global kinematic impacts caused by platform motion and changes in the human-robot spatial relationship from the robot's perspective. Since these global features are relatively dispersed in the temporal dimension, we use Global Average Pooling (GAP) to effectively aggregate information from all spatiotemporal regions, ensuring that contributions from different regions are considered. The process is as follows:

$$\begin{aligned}\mathbf{Y}_{kinematic} &= \mathbf{Interpreter}_{\mathbf{kinematic}}(\mathbf{F}_2, \mathbf{F}_3) \\ &= \frac{1}{2}\left(\mathbf{MLP}(\mathbf{GAP}(\mathbf{Transformer}(\mathbf{F}_2))) \right. \\ &\quad \left. + \mathbf{MLP}(\mathbf{GAP}(\mathbf{Transformer}(\mathbf{F}_3)))\right).\end{aligned} \tag{5}$$

Human Action Recognizer: For features extracted from lower layers with denser point clouds (*e.g.*, $\mathbf{F}_1$), these features more directly reflect subtle human actions. Here, we use Global Max Pooling (GMP) to select the most significant spatiotemporal points, emphasizing key fine-grained action details. The calculation is as follows:

$$\begin{aligned}\mathbf{Y}_{human} &= \mathbf{Recognizer}_{\mathbf{human}}(\mathbf{F}_1) \\ &= \mathbf{MLP}(\mathbf{GMP}(\mathbf{Transformer}(\mathbf{F}_1)).\end{aligned} \tag{6}$$

Finally, we fuse the classification scores from both branches, averaging their contributions. This approach retains the robustness of global kinematic interpretation while capturing the high discriminability of subtle action changes.

4.5. Elastic Ellipse Query

Let the 4D point cloud set be $P = \{p_1, p_2, \ldots, p_N\}$, where each point $p_n = (x_n, y_n, z_n)$ represents the neighboring point n of a query point. Given a query point p_q and a radius r, the objective of the Ball Query is to find all points whose distance from the query point p_q is less than or equal to r. Let $d(p_n, p_q)$ denote the Euclidean distance between point p_n and query point p_q:

$$d(p_n, p_q) = \sqrt{(x_n - x_q)^2 + (y_n - y_q)^2 + (z_n - z_q)^2}, \tag{7}$$

where $p_q = (x_q, y_q, z_q)$ is the query point.

The goal of Ball Query is to identify all points p_n satisfying the following condition:

$$\mathcal{B}(p_q, r) = \{p_n \mid d(p_n, p_q) \leq r\}. \tag{8}$$

That is, the set of points within the spherical neighborhood $\mathcal{B}(p_q, r)$ consists of all points p_n such that $d(p_n, p_q) \leq r$.

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 06:15:37 UTC from IEEE Xplore. Restrictions apply.

In N-HRI, conventional spherical neighborhood queries (*e.g.*, Ball Query) exhibit inherent limitations when confronted with kinematic-induced geometric distortions. The relative motion between mobile robots and humans introduces anisotropic deformations in the spatiotemporal point cloud distribution. Three critical challenges emerge: 1) Scale Discrepancy: Human actions (*e.g.*, hand gestures) manifest as localized deformations (0.1-0.5m) within the robot's egocentric coordinates, while kinematic disturbances induce macroscopic displacements (up to 50m) along the xy-plane. 2) Axis-wise Heterogeneity: Vertical displacements (along the z-axis) primarily reflect genuine human joint movements, while horizontal movements (in the xy-plane) are more susceptible to artifacts caused by the robot's self-motion and changes in the human-robot spatial positioning. 3) Dynamic Occlusion: Transient point cloud density variations caused by viewpoint changes violate the isotropic neighborhood assumption.

For RARV, visual similarities influenced by kinematic disturbances from robotic view complicate the differentiation, causing deformations in human-actions and irregular distributions in the spatiotemporal domain. Unlike kinematic disturbances, which have a large range of movement, human-actions are constrained to smaller areas. For instance, while walking involves significant movement in the xy-plane, human-actions like waving hands are restricted to a smaller range. Due to the influence of kinematic, points with large xy distance at different times are more likely to represent the same point, while points with large differences in the vertical dimension are more likely to be distinct.

To address these challenges, we propose the *Elastic Ellipse Query*. Considering the distribution differences between datasets, we introduce axis-specific adaptive query scale learning. This mechanism establishes an anisotropic metric space through learnable axis-specific scaling, enabling partial tracking of points and differentiation of actions. Given a query point p_q and a neighborhood point p_n, our adaptive distance metric is formulated as follows:

$$d(p_n, p_q) = \sqrt{\alpha(x_n - x_q)^2 + \beta(y_n - y_q)^2 + \gamma(z_n - z_q)^2}, \quad (9)$$

where $\Omega = \{\alpha, \beta, \gamma\}$ are trainable scaling parameters that automatically adjust the neighborhood aspect ratios. For the ACTIVE dataset, the final values of α, β, and γ are $(3.5632, 3.6789, 2.8038)$, consistent with the expected scale ratios. This elliptic metric space provides following distinctive advantages: 1) Kinematic-Aware Adaptation: The horizontal scaling factors (α, β) suppress planar displacements induced by the robot while preserving fine-grained human motions. In our experiments with the ACTIVE dataset, the optimized parameters ($\alpha = 3.56$, $\beta = 3.68$) effectively compress the neighborhood range in the xy-plane relative to the z-axis ($\gamma = 2.80$). The vertical scaling factor γ remains sensitive to subtle posture changes, such as a 0.2-meter arm elevation, while the horizontal scaling factors adaptively filter out irrelevant positional shifts. 2) Dynamic Focus Adjustment: Unlike fixed elliptical kernels, our parameters evolve during training to handle varying interaction distances (3-50m).

Table 2. Action recognition accuracy (%) of ACTIVE-PC on ACTIVE under different dataset settings.

Method	Venue	Setting		
		8 × 512	12 × 512	12 × 768
PSTNet [7]	ICLR'21	18.85	26.49	34.10
P4Transformer [6]	CVPR'21	37.97	40.56	43.88
PSTNet++ [8]	TPAMI'22	34.98	38.81	40.29
PST-Transformer [9]	TPAMI'23	41.61	42.92	46.76
3DinAction [2]	CVPR'24	44.88	46.23	49.19
ACTIVE-PC w/o EEQ	-	52.29	53.32	55.11
ACTIVE-PC (Ours)	-	**58.71**	**59.18**	**60.10**

Table 3. Action recognition accuracy (%) of ACTIVE-PC on NTU RGB+D. CS and CV represent the cross-subject and cross-view evaluation protocols, respectively.

Method	Venue	CS	CV
PSTNet [7]	ICLR'2021	90.5	96.5
P4Transformer [6]	CVPR'2021	90.2	96.4
PointCMP [22]	CVPR'2023	88.5	-
PointCPSC [23]	ICCV'2023	88.0	-
PST-Transformer [9]	TPAMI'2023	91.0	96.4
ACTIVE-PC (Ours)	-	**91.7**	**96.8**

Table 4. Action recognition accuracy (%) of ACTIVE-RGB on ACTIVE under different dataset settings.

Method	Venue	Setting	
		8 × 224	12 × 224
VideoSwin [17]	CVPR'22	41.07	44.20
MViT V2 [14]	CVPR'22	38.15	41.83
VideoMAE [24]	NeurIPS'22	42.46	45.63
UniFormer [11]	TPAMI'23	42.55	45.60
VideoMAE V2 [28]	CVPR'23	45.39	47.98
UniFormerv2 [12]	ICCV'23	48.21	50.74
InternVideo2 [30]	ECCV'24	52.52	54.88
ACTIVE-RGB (Ours)	-	**55.90**	**57.56**

5. Experiment

In this section, we evaluate the performance of state-of-the-art methods and ACTIVE-PC, on the ACTIVE dataset across both point cloud and RGB modalities. We also conduct ablation experiments on the proposed components.

5.1. Metrics and Implementation Details

For evaluation, we use standard action recognition metrics, specifically classification accuracy. In the point cloud modality, we first apply the TransFusion[1] method to detect human subjects within the scene. The point cloud sequence is then segmented into fixed-length clips, with each frame sampled to 768 points. For training, video-level labels are assigned as clip-level labels. The color informa-

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 06:15:37 UTC from IEEE Xplore. Restrictions apply.

Table 5. Impact of Layered Recognizers, MNS and EEQ of ACTIVE-PC on ACTIVE.

Layered Recognizers	MNS	EEQ	Accuracy (%)
✗	✗	✗	46.76
✓	✗	✗	51.98
✓	✓	✗	55.11
✓	✗	✓	55.63
✓	✓	✓	60.10

Table 6. Influence of the spatial radius r. EEQ reduces the sensitivity to the hyperparameter r.

EEQ	Spatial radius r_s						Variance
	0.1	0.2	0.5	1	1.5	2	
✗	54.83	55.11	55.39	54.52	54.21	53.51	0.456
✓	59.98	60.10	60.04	59.73	59.54	59.30	0.100

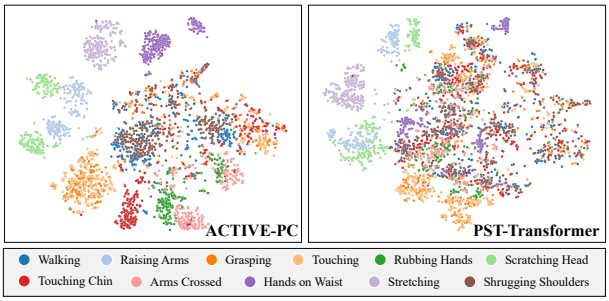


Figure 4. t-SNE distribution of PST-Transformer [6] and ACTIVE-PC on ACTIVE dataset. Different colors represent different action categories, with each point corresponding to a video sample. All 11 categories from the Base Motions are visualized.

tion of the points is not used. For evaluation, the video-level prediction is obtained by averaging the clip-level predicted probabilities. The model is trained using the SGD optimizer for 50 epochs, with a learning rate of 0.01, decaying by a factor of 0.1 at epochs 20 and 30. For the RGB modality, human bodies are detected using the object detection method [26]. Each video is evenly split into 8 or 12 segments, with one frame randomly sampled from each segment. Each frame is resized to 224 × 224. We use the AdamW optimizer [18] for 50 epochs, with a learning rate of 0.0002. All experiments are conducted on the PyTorch platform, utilizing an NVIDIA RTX 4090 GPU.

5.2. Point Cloud Modality Evaluation

We first validate the performance of ACTIVE-PC on the ACTIVE dataset. As shown in Tab. 2, Transformer-based models, such as P4Transformer and PST-Transformer, outperform CNN-based models achieving Acc-Top1 scores of 43.88% and 46.76%, respectively. ACTIVE-PC surpasses PST-Transformer by 13.34%, reaching an Acc-Top1 of 60.10%, demonstrating the effectiveness of our method. To validate the robustness of the model, we conduct experiments on two general action recognition datasets, MSR-Action3D and NTU RGB+D. Tab. 3 presents the results on NTU RGB+D.

5.3. RGB Modality Evaluation

ACTIVE-RGB is built upon the ViT [5], augmented with Layered Recognizers. Tab. 4 presents a performance comparison of ACTIVE-RGB with other state-of-the-art methods in the RGB modality. ACTIVE-RGB outperforms the leading method, the baseline model InternVideo2, by 2.68%, validating the effectiveness of our approach.

5.4. Ablation Study

Evaluation of different Components: We conduct ablation experiments on the key components of our model, as well as the number of sampled frames per point cloud video. Results in Tab. 5 show that each proposed modification contributes positively. Specifically, w/o EEQ and w/o MNS show performance drops of 4.99% and 4.47%, respectively, compared to ACTIVE-PC. The significant performance degradation observed with w/o Layered Recognizers highlights the effectiveness of fusing high-frequency and low-frequency features. Allowing the model to simultaneously focus on both action and kinematic variations enhances its ability to handle complex environments.

Evaluation of the Stability of EEQ: Tab. 6 compares the performance of the full ACTIVE-PC with its variant, which excludes the EEQ module, under different neighborhood radius settings. Traditional neighborhood query methods are constrained by fixed neighborhood radii, whereas EEQ adapts by learning the intrinsic geometric properties of the data distribution, allowing for flexible scaling of the spherical neighborhood, thereby enhancing robustness. The variance of accuracy (acc) for ACTIVE-PC across different neighborhood radii is 0.100, significantly outperforming the variant without EEQ, which exhibits a variance of 0.456.

5.5. Visualization and Analysis

We utilize t-SNE [25] visualization to analyze feature distributions from the point cloud modality, as shown in Fig. 4. ACTIVE-PC distinctly separates similar actions (e.g., "Raising Arms" vs. "Scratching Head" and "Arms Crossed" vs. "Hands on Waist"), whereas PST-Transformer [6] produces more overlapping clusters. This demonstrates ACTIVE-PC's superior ability to differentiate subtle action categories from a robotic viewpoint.

6. Conclusion

To the best of our knowledge, our dataset is the first human action recognition dataset specifically designed for Natural Human-Robot Interaction scenarios, laying the foundation for more accurate and detailed understanding of human actions in N-HRI. We also propose ACTIVE-PC designed for action recognition in N-HRI scenarios. Experimental results demonstrate the effectiveness of this approach.

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 06:15:37 UTC from IEEE Xplore. Restrictions apply.

Acknowledgements This work was supported by National Natural Science Foundation of China (No. 62473007), Natural Science Foundation of Guangdong Province (No. 2024A1515012089), Shenzhen Innovation in Science and Technology Foundation for The Excellent Youth Scholars (No. RCYX20231211090248064).

References

[1] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022. 7

[2] Yizhak Ben-Shabat, Oren Shrout, and Stephen Gould. 3dinaction: Understanding human actions in 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 7

[3] Elif Bozkaya-Aras, Tolga Onel, Levent Eriskin, and Mumtaz Karatas. Intelligent human activity recognition for healthcare digital twin. *Internet of Things*, 2025. 1

[4] Haoyu Chen, Henglin Shi, Xin Liu, Xiaobai Li, and Guoying Zhao. Smg: A micro-gesture dataset towards spontaneous body gestures for emotional stress state analysis. *International Journal of Computer Vision*, 2023. 3

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, 2020. 8

[6] Hehe Fan, Yi Yang, and Mohan Kankanhalli. Point 4d transformer networks for spatio-temporal modeling in point cloud videos. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3, 7, 8

[7] Hehe Fan, Xin Yu, Yuhang Ding, Yi Yang, and MohanS. Kankanhalli. Pstnet: Point spatio-temporal convolution on point cloud sequences. *Cornell University - arXiv,Cornell University - arXiv*, 2022. 7

[8] Hehe Fan, Xin Yu, Yi Yang, and Mohan Kankanhalli. Deep hierarchical representation of point cloud videos via spatio-temporal decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 7

[9] Hehe Fan, Yi Yang, and Mohan Kankanhalli. Point spatio-temporal transformer networks for point cloud video modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3, 7

[10] Shiming Ge, Jia Li, Qiting Ye, and Zhao Luo. Detecting masked faces in the wild with lle-cnns. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[11] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Jiao Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3, 7

[12] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. Uniformerv2: Unlocking the potential of image vits for video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 7

[13] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3d points. In *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*, 2010. 2, 3

[14] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022. 7

[15] Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhao. imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021. 3

[16] Yunze Liu, Junyu Chen, Zekai Zhang, Jingwei Huang, and Li Yi. Leaf: Learning frames for 4d point cloud sequence understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3

[17] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022. 7

[18] I Loshchilov. Decoupled weight decay regularization. 2017. 8

[19] Marcus Mast, Michael Burmester, Birgit Graf, Florian Weisshardt, Georg Arbeiter, Michal Spanel, Zdeněk Materna, Pavel Smrz, and Gernot Kronreif. Design of the human-robot interaction for a semi-autonomous service robot to assist elderly people. 2015. 1

[20] Yang Mi and Song Wang. Recognizing micro actions in videos: learning motion details via segment-level temporal pyramid. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, 2019. 3

[21] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2016. 2, 3

[22] Zhiqiang Shen, Xiaoxiao Sheng, Longguang Wang, Yulan Guo, Qiong Liu, and Xi Zhou. Pointcmp: Contrastive mask prediction for self-supervised learning on point cloud videos. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3, 7

[23] Xiaoxiao Sheng, Zhiqiang Shen, Gang Xiao, Longguang Wang, Yu Kuen Guo, and Hehe Fan. Point contrastive prediction with semantic clustering for self-supervised learning on point cloud videos. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 7

[24] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 2022. 3, 7

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 06:15:37 UTC from IEEE Xplore. Restrictions apply.

[25] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 2008. 8

[26] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yolov10: Real-time end-to-end object detection. *arXiv preprint arXiv:2405.14458*, 2024. 8

[27] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 3

[28] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 7

[29] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning. *ArXiv*, 2022. 3

[30] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, 2024. 7

[31] Ping Wei, Yibiao Zhao, Nanning Zheng, and Song-Chun Zhu. Modeling 4d human-object interactions for event and object recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2013. 3

[32] Sravani Yenduri, Nazil Perveen, Vishnu Chalavadi, and C KrishnaMohan. Fine-grained action recognition using dynamic kernels. *Pattern Recognit.*, 2022. 1

[33] Ryo Yonetani, Kris M Kitani, and Yoichi Sato. Recognizing micro-actions and reactions from paired egocentric videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 3

[34] Jianqi Zhong, Kaichen Zhou, Qingyong Hu, Bing Wang, Niki Trigoni, and A. Markham. No pain, big gain: Classify dynamic point cloud sequences with static models by fitting feature-level space-time surfaces. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 06:15:37 UTC from IEEE Xplore. Restrictions apply.