# CHASE: Learning Convex Hull Adaptive Shift for Skeleton-based Multi-Entity Action Recognition

**Yuhang Wen**
Sun Yat-sen University
wenyh29@mail2.sysu.edu.cn

**Mengyuan Liu***
State Key Laboratory of General Artificial Intelligence
Peking University
nkliuyifang@gmail.com

**Songtao Wu**
Sony R&D Center China
Songtao.Wu@sony.com

**Beichen Ding***
Sun Yat-sen University
dingbch@mail.sysu.edu.cn

## Abstract

Skeleton-based multi-entity action recognition is a challenging task aiming to identify interactive actions or group activities involving multiple diverse entities. Existing models for individuals often fall short in this task due to the inherent distribution discrepancies among entity skeletons, leading to suboptimal backbone optimization. To this end, we introduce a Convex Hull Adaptive Shift based multi-Entity action recognition method (CHASE), which mitigates inter-entity distribution gaps and unbiases subsequent backbones. Specifically, CHASE comprises a learnable parameterized network and an auxiliary objective. The parameterized network achieves plausible, sample-adaptive repositioning of skeleton sequences through two key components. First, the Implicit Convex Hull Constrained Adaptive Shift ensures that the new origin of the coordinate system is within the skeleton convex hull. Second, the Coefficient Learning Block provides a lightweight parameterization of the mapping from skeleton sequences to their specific coefficients in convex combinations. Moreover, to guide the optimization of this network for discrepancy minimization, we propose the Mini-batch Pair-wise Maximum Mean Discrepancy as the additional objective. CHASE operates as a sample-adaptive normalization method to mitigate inter-entity distribution discrepancies, thereby reducing data bias and improving the subsequent classifier's multi-entity action recognition performance. Extensive experiments on six datasets, including NTU Mutual 11/26, H2O, Assembly101, Collective Activity and Volleyball, consistently verify our approach by seamlessly adapting to single-entity backbones and boosting their performance in multi-entity scenarios. Our code is publicly available at https://github.com/Necolizer/CHASE.

## 1 Introduction

Multi-entity action recognition, a challenging task derived from action recognition [1, 2, 3, 4, 5, 6, 7, 8, 9], aims to find the optimal estimator of the mapping from multi-entity motions to semantic labels, where entities involved can range from human bodies [10, 11], hands [12] to various objects [13]. Recent approaches predominantly rely on skeletal data for addressing this challenge [10, 11, 14], given that skeletons serve as a concise representation of spatiotemporal features [15, 16, 17, 18, 19, 20, 21]. This task has broad applications in human-robot interaction [22, 23], scene understanding [24, 25, 26, 27, 28], human motion analysis [29, 30, 31, 32, 33, 34], etc.

*Corresponding Authors.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

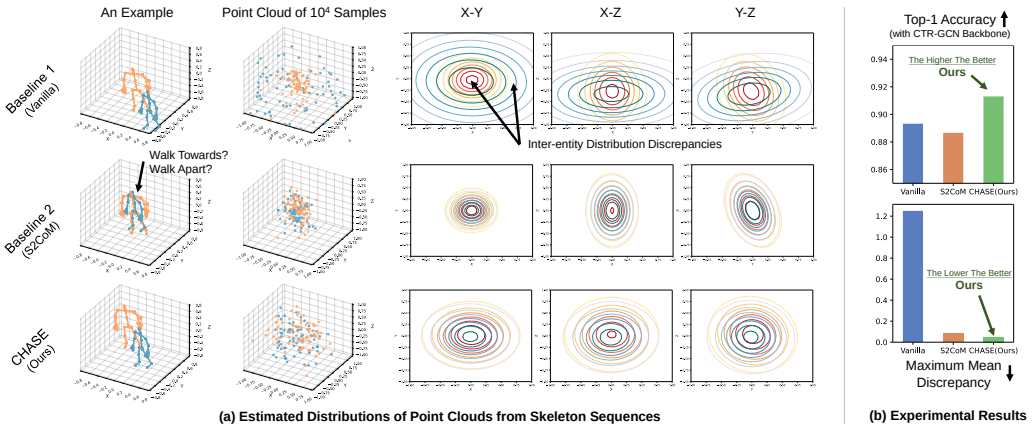



Figure 1: **Inter-entity distribution discrepancies in multi-entity action recognition task.** (a) We delineate three distinct settings: *Vanilla* (a common practice), *S2CoM* (an intuitive baseline approach), and *CHASE* (our proposed method). Column 2 illustrates spatiotemporal point clouds defined by the skeletons over $10^4$ sequences. Column 3-5 depict the projections of estimated distributions of these point clouds onto the x-y, z-x, and y-z planes. These projections reveal significant inter-entity distribution discrepancies when using *Vanilla*. (b) The discrepancies observed in *Vanilla* introduce bias into backbone models, leading to unsatisfactory performance. Although *S2CoM* can reduce these discrepancies, it makes the classifiers produce wrong predictions due to a complete loss of inter-entity information. With the lowest inter-entity discrepancy, our method unbiases the subsequent backbone to get the highest accuracy, underscoring its efficacy.

Experiments have revealed that network architectures tailored for single-entity actions get unsatisfactory performance when confronted with multi-entity actions [10, 35]. This inadequacy can be attributed to a common practice [36, 37, 38, 39, 40] observed in treating interactions: each entity is encoded independently using the same single-entity backbone, and their features are averaged for recognition. This practice is based on an empirical assumption that each entity is independent and identically distributed (i.i.d.). But we demonstrate that different entities depicted by skeletons exhibit evident non-i.i.d. characteristics. Fig. 1 (a) Row 1 reveals significant inter-entity distribution discrepancies using estimated distributions of joints from distinct entities. Such discrepancies can introduce bias into the backbone models, leading to suboptimal optimization and performance. It explains why multi-entity action modeling usually diverges from the single-entity one.

Using local coordinates for each entity holds promise in rendering them i.i.d., achieved by shifting individual origins to the per-entity spatiotemporal centers of mass (S2CoM), as depicted in Fig. 1 (a) Row 2. S2CoM is a straightforward and intuitive baseline to address this problem. However, this approach exacts a significant toll as it entails a complete loss of inter-entity information. Experimental results corroborate this notion, as illustrated in Fig. 1 (b), showcasing the detrimental impact of lacking inter-entity measurements on recognition performance. Nonetheless, this endeavor sparks an insightful realization: the potential for narrowing distribution gaps through origin shifts, thereby improving the performance of single-entity backbones in multi-entity scenarios. A natural question arises: Can we reduce the bias by finding the optimal sample-adaptive shift in $\mathbb{R}^3$ that minimizes the distribution discrepancies among entities?

To address the inter-entity distribution discrepancy problem, we propose a Convex Hull Adaptive Shift based multi-Entity action recognition method (CHASE). Serving as an additional normalization step, CHASE aims to accompany other single-entity backbones for enhanced multi-entity action recognition. Our main insight lies in the adaptive repositioning of skeleton sequences to mitigate inter-entity distribution gaps, thereby unbiasing the subsequent backbone and boosting its performance. Specifically, CHASE consists of a learnable parameterized network and an auxiliary objective. The parameterized network can achieve plausible and sample-adaptive repositioning of skeleton sequences through two crucial components. First, the Implicit Convex Hull Constrained Adaptive Shift (ICHAS) ensures that the new origin of the coordinate system is within the skeleton convex hull. Second, the Coefficient Learning Block (CLB) provides a lightweight parameterization of the mapping from skeleton sequences to their specific coefficients in ICHAS. Moreover, to guide the optimization of this network for discrepancy minimization, we propose the Mini-batch

Pair-wise Maximum Mean Discrepancy (MPMMD) as the additional objective. This loss function quantifies pair-wise entity discrepancies using maximum mean discrepancy and integrates mini-batch sampling strategies to estimate the expectation. In conclusion, CHASE works as a sample-adaptive normalization method to mitigate inter-entity distribution discrepancies, which can reduce bias in the subsequent classifier and enhance its multi-entity action recognition performance.

The contributions of this paper are three-fold:

1. To the best of our knowledge, we are the first to investigate the issue of inter-entity distribution discrepancies in multi-entity action recognition. Our proposed method, Convex Hull Adaptive Shift for Multi-Entity Actions, effectively addresses this challenge. Our main idea is adaptively repositioning skeleton sequences to mitigate inter-entity distribution gaps, thereby unbiasing the subsequent backbones and boosting their performance.
2. Serving as an additional normalization step for backbone models, CHASE consists of a learnable network and an auxiliary objective. Specifically, this network is formulated by the Implicit Convex Hull Constrained Adaptive Shift, together with the parameterization of a lightweight Coefficient Learning Block, which learns sample-adaptive origin shifts within skeleton convex hull. Additionally, the Mini-batch Pair-wise Maximum Mean Discrepancy objective is proposed to guide the discrepancy minimization.
3. Experiments on NTU Mutual 11, NTU Mutual 26, H2O, Assembly101, Collective Activity Dataset and Volleyball Dataset consistently verify our proposed method by improving performance of single-entity backbones in multi-entity action recognition task.

## 2 Related Work

### 2.1 Skeleton-based Action Recognition

**Datasets & Models**. Datasets [41, 42, 43, 44] proffering annotated or estimated skeleton sequences support the development of skeleton-based action recognition. Based on these benchmarks, a significant body of works focus on the design of artificial neural network architecture for more effective skeleton-based action recognition. Early models rely on the basic architecture of Recurrent Neural Network to capture temporal motions [45, 46, 47, 48, 49, 50]. Graph Convolution Network (GCN) shows predominated popularity as various graph convolution operators being proposed [36, 37, 38, 51, 52, 53, 54, 55, 56, 57, 58]. Recent progress of the model design is largely driven by adopting self-attention mechanism and transformer architecture [39, 40, 59, 60, 61, 62, 63, 64].

**Optimization Objectives**. Several works have explored additional optimization objectives beyond the commonly used cross-entropy (CE) loss to ensure robust recognition [37, 65], address challenging open-set problems [16], or integrate supplementary natural language descriptions [66, 67].

However, existing methods are usually developed under the empirical assumption that entities are i.i.d. allowing the backbones to learn representations of actions concerning only one entity [36, 37, 38, 39, 40, 53]. However, when confronted with multi-entity interactions, their common practice of feeding the backbone separately often proves inadequate. Our proposed approach can seamlessly adapt to these existing methods, boosting their performance by minimizing the distribution discrepancies.

### 2.2 Skeleton-based Multi-Entity Action Recognition

**Interactive Actions**. Addressing datasets featuring two-person actions [41, 42, 68, 69, 70, 71, 72] or egocentric hand-object interactions [12, 13, 73, 74] necessitates effective interaction modeling. This spurs the development of various interaction recognition models leveraging human body and hand graph priors [10, 75]. Notably, the introduction of the general interactive action recognition task [35] unifies diverse interactions across various entity types, including person-to-person [10, 11, 14, 35, 76, 77], hand-to-hand [12, 35, 78, 79] and hand-to-object [13, 35, 75, 79, 80, 81] interactions.

**Group Activities**. Another interesting area of study is group activities [82, 83, 84], which involve more entities and may include irrelevant individual motions [85, 86]. To this end, recent works usually leverage compositional reasoning from group skeletons, either alone or in combination with additional modalities, to achieve promising results [87, 88, 89, 90, 91, 92, 93, 94, 95, 96].

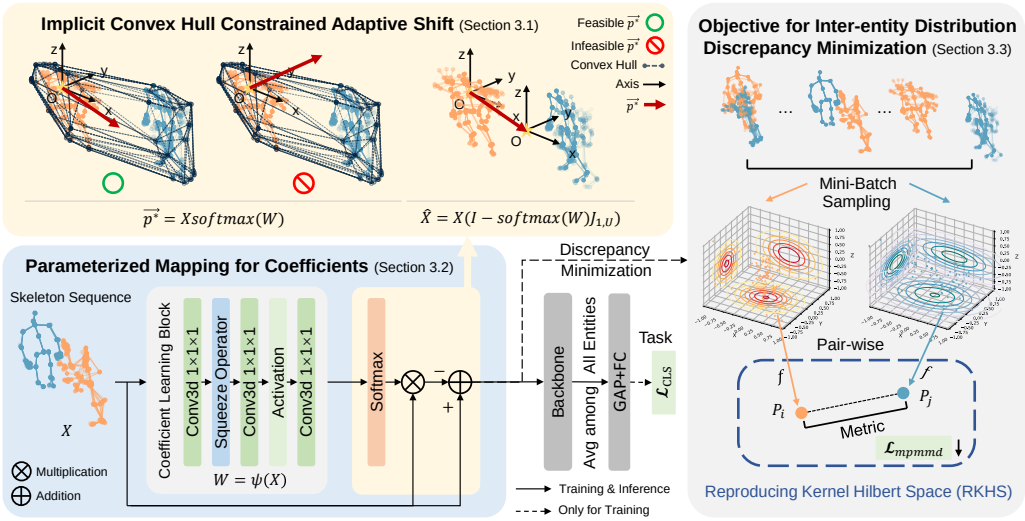



Figure 2: **The overall framework of the proposed CHASE for multi-entity action recognition.** Given a skeleton sequence of multi-entity action as input, CHASE executes an implicit convex hull constrained adaptive shift with the Coefficient Learning Block, implemented as a lightweight backbone wrapper. CHASE also collects pair-wise shifted skeletons within mini-batches, effectively alleviating inter-entity distribution discrepancies by introducing an additional objective.

While these works demonstrate satisfactory performance through interaction modelling, some may encounter model scalability issues when confronted with the factorial growth of inter-entity interactions [10, 13, 35, 75]. Moreover, they usually lack sufficient justification for why multi-entity action modeling significantly diverges from the single-entity one [10, 35, 79, 81]. In this paper, we delve into the inter-entity distribution discrepancy problem and introduce CHASE as a solution to minimize discrepancies. Through our proposed method, we aim to demonstrate that single-entity backbones can work well in multi-entity settings.

## 3 CHASE

Fig. 2 presents the framework of our proposed CHASE for skeleton-based multi-entity action recognition. We begin by presenting the formulation of the implicit convex hull constrained adaptive shift in Section 3.1, followed by the design of a lightweight Coefficient Learning Block in Section 3.2. In Section 3.3, we subsequently introduce an additional objective termed Mini-batch Pair-wise Maximum Mean Discrepancy to further mitigate inter-entity distribution discrepancies.

### 3.1 Implicit Convex Hull Constrained Adaptive Shift

The observed inter-entity distribution discrepancy in multi-entity skeleton sequences stems from the initial configuration of the world coordinate system. To mitigate this discrepancy, we propose an adaptive shift mechanism for each multi-entity skeleton sequence. It guides the origin to a sample-adaptive location, aiming to render each entity approximately i.i.d.. Moreover, based on the empirical assumption that the origin should not be far away from the skeletons, we implicitly constrain the new origin to remain within the skeleton convex hull by proving a simple but crucial proposition.

Consider a scenario where $E$ interactive entities (e.g. persons) engage in purposeful activities over a duration of $T$, and the pose of each entity is indicated by $J$ joints with $C$ Cartesian coordinates. The skeleton sequence of a multi-entity action is defined as $X \in \mathbb{R}^{C\times T\times J\times E}$. For clarity we denote $U = T \times J \times E$. Given points $\vec{p_i} \in \mathbb{R}^{C\times 1}$ in $X \in \mathbb{R}^{C\times U}$, the subtraction $\vec{\hat{p}_i} = \vec{p_i} - \vec{p^*}(1 \leq i \leq U)$ defines a shift of origin for them, where $\vec{\hat{p}_i}, \vec{p^*} \in \mathbb{R}^{C\times 1}$. This can be expressed in matrix form as:

$$\hat{X} = X - \vec{p^*} J_{1,U}, \tag{1}$$

where $J_{1,U} \in \mathbb{R}^{1\times U}$ is a matrix of ones, and $\hat{X}$ is the shifted skeleton sequence. Now the problem is to make the shift vector $\vec{p^*}$ adaptive to $X$. A naive implementation is the linear combination:

$$\hat{X} = X - \vec{p^*} J_{1,U} = X(I - W J_{1,U}), \tag{2}$$

where $I \in \mathbb{R}^{U\times U}$ and the weight matrix $W \in \mathbb{R}^{U\times 1}$.

However, optimizing $W$ can be challenging without constraints, as $\vec{p^*}$ could potentially be any point in $\mathbb{R}^3$. It is therefore reasonable to constrain $\vec{p^*}$ by incorporating the definition of the **Convex Hull**.

**Definition 1** (Convex Hull [97])**.** The convex hull $S$ of a given set $X$ can be defined as: 1) The (unique) minimal convex set containing $X$. 2) The set of all convex combinations of points in $X$. These definitions are equivalent.

We jump to the formulation of the implicit skeleton convex hull constrained adaptive shift vector by proving the following proposition:

**Proposition 1.** *The implicit skeleton convex hull constrained adaptive shift vector is formulated as*

$$\vec{p^*} = X\mathrm{softmax}(W), \tag{3}$$

*where $X \in \mathbb{R}^{C\times U}$, $W \in \mathbb{R}^{U\times 1}$, and $\vec{p^*} \in \mathbb{R}^{C\times 1}$. $\vec{p^*}$ in Eq. 3 is an element in the set of all convex combinations of points in $X$. It is also a point that lies in the minimal convex set containing $X$.*

*Proof.* The first half of this proposition is equivalent to show that the matrix product of $X$ and $\mathrm{softmax}(W)$ is a convex combination of $X$. $X$ is a set of points $\vec{p}_1, \ldots, \vec{p}_U$ with $C$ Cartesian coordinates. We denote $\mathrm{softmax}(W)$ as $\tilde{W}$ with component $\tilde{\alpha}_i$, which is formulated as

$$\tilde{\alpha}_i = \frac{e^{\alpha_i}}{\sum_{j=1}^{U} e^{\alpha_j}} \quad (1 \le i \le U), \tag{4}$$

where $\alpha_i$ is a component of $W$. By applying function $\mathrm{softmax} : \mathbb{R}^U \mapsto (0,1)^U$, each component $\tilde{\alpha}_i$ of $\tilde{W}$ will be in the interval $(0,1)$, and the components will add up to 1. Thus we have $\vec{p^*} = \sum_{i=1}^{U} \tilde{\alpha}_i \vec{p}_i$, where all $\tilde{\alpha}_i \in \mathbb{R}$ satisfy $\tilde{\alpha}_i > 0$ and $\sum_{i=1}^{U} \tilde{\alpha}_i = 1$. This is sufficient for the definition of a convex combination, which only requires $\tilde{\alpha}_i \ge 0$. Then the second half of this proposition is evident with the equivalence of definitions in Def. 1. □

Proposition 1 also implies that all possible $\vec{p^*}$ constitute a subset $\tilde{S}$ of the convex hull $S$ defined by the skeleton joints for all entities during the action period:

$$\tilde{S} = \left\{ \sum_{i=1}^{U} \tilde{\alpha}_i \vec{p}_i \,\middle|\, \vec{p}_i \in X, \sum_{i=1}^{U} \tilde{\alpha}_i = 1, \tilde{\alpha}_i \in (0,1) \right\} \subset S, \tag{5}$$

which specifically is the interior of $S$ (i.e., the open convex hull of $X$). We provide an example of the feasible $\vec{p^*}$ in the interior of $S$, marked by a green circle in Fig. 2. The center of mass (CoM) $\bar{\vec{p}}$ is also in the set $\tilde{S}$, proven by simply taking all $\tilde{\alpha}_i = 1/U (1 \le i \le U)$.

With Eq. 1 and Eq. 3, we introduce Implicit Convex Hull Constrained Adaptive Shift as:

$$\hat{X} = X(I - \mathrm{softmax}(W) J_{1,U}), \tag{6}$$

where $W$ is coefficients needed to be optimized. In Eq. 2, the search space for $\vec{p^*}$ encompasses the entire $R^3$. However, in Eq. 6, it's restricted to the open convex hull $\tilde{S}$. We optimize the weights for each point under the constraint of the skeleton convex hull, subsequently deriving the adaptive shift vector for each sample. Applying a softmax function implicitly constrains $\vec{p^*}$ to remain within the convex hull $S$, while preserving inter-entity measurements. Consequently, the subtraction between the point set and the shift vector repositions the origin to a specific point in the open convex hull.

### 3.2 Parameterized Mapping for Coefficients

In this section, a lightweight Coefficient Learning Block is introduced to parameterize the mapping from the input skeleton sequence to the weight matrix. This parameterization allows CHASE to achieve sample-adaptive coefficients beyond sample-adaptive shifts formulated in Section 3.1.

In Eq. 6, we note that the first-order partial derivative of $\hat{X}$ with respect to $X$ is

$$\frac{\partial \hat{X}}{\partial X} = I - J_{U,1}\mathrm{softmax}(W^T), \tag{7}$$

whose result is constant. This implies that the same learnt weight matrix $W$ is applied to all different $X$s when getting adaptive $\vec{p^*}$s. To make the coefficients $W$ dependent on the input $X$, a mapping $\psi : \mathbb{R}^{C\times U} \mapsto \mathbb{R}^{U\times 1}$ is expected to map $X$ to $W$.

As depicted in Fig. 2, we parameterize the nonlinear mapping $\psi$ as a sequence of learnable layers, termed the Coefficient Learning Block. This lightweight CLB can be formulated as follows:

$$W = \psi(X) = W_3\delta(W_2\phi(W_1X + b)), \tag{8}$$

where $W_1 \in \mathbb{R}^{C_1\times C}, W_2 \in \mathbb{R}^{C_2\times C_1}, W_3 \in \mathbb{R}^{U\times C_2}$ are weight matrices, $b$ is a bias matrix, $\phi : \mathbb{R}^{C_1\times U} \mapsto \mathbb{R}^{C_1\times 1}$ is a squeeze operator [98] and $\delta$ is an activation function. Using a dimensionality-reduction layer and a dimensionality-increasing layer around the non-linearity is a common gating mechanism parameterization [98, 99]. Hence, we ensure $U \geq C_1 > C_2$.

### 3.3 Objective for Inter-entity Distribution Discrepancy Minimization

To facilitate CHASE optimization, we introduce an additional objective aimed at minimizing the inter-entity distribution discrepancies of the shifted skeleton sequences. This objective quantifies the pair-wise discrepancies and employs mini-batch sampling strategies to estimate the expectation.

Maximum mean discrepancy is a metric used to measure the distance between distributions, defined as the distance between their embeddings in the reproducing kernel Hilbert space (RKHS) $\mathcal{H}$:

$$\mathrm{MMD}(P, Q) = \sup_{\|f\|_{\mathcal{H}}\leq 1} \left(\mathbb{E}[f(x)] - \mathbb{E}[f(y)]\right), \tag{9}$$

where $\sup(\cdot)$ denotes the supremum. It is equivalent to finding the RKHS function $f$ that maximizes the difference in expectations between the two probability distributions $P(x)$ and $Q(y)$.

Suppose each entity distribution is denoted as $P^i(1 \leq i \leq E)$ for $E$ entities, we measure the distance of all pair-wise distributions using the empirical mean

$$\mathbb{E}_{r(z)}[\mathrm{MMD}(z)] = \sum_{i=1}^{E-1}\sum_{j=i+1}^{E} \mathrm{MMD}(P^i, P^j)/\mathrm{C}(E, 2), \tag{10}$$

where $z = (P^i, P^j)(1 \leq i, j \leq E, i \neq j)$ with the probability density $r(z)$, and $C(E, 2)$ denotes a combination of $E$ things taken 2 at a time without repetition. We adopt two approximations for computational efficiency. The first involves estimating $\mathbb{E}[f(x)]$ in Eq. 9 using a mini-batch of $x$. The second approximation concerns the right-hand side of Eq. 10, which is impractical due to its complexity of $O(n!)$ in terms of the entity count. Instead, it can be approximated by uniformly sampling a mini-batch of $M$ entity pairs from all possible $\mathrm{C}(E, 2)$ combinations $z$:

$$\mathbb{E}_{r(z)}[\mathrm{MMD}(z)] \approx \frac{1}{M}\sum_{m=1}^{M} \mathrm{MMD}(z_m). \tag{11}$$

We denote Eq. 11 with the above two approximations to be the Mini-batch Pair-wise Maximum Mean Discrepancy Loss $\mathcal{L}_{mpmmd}$, thereby we have the total loss function for training:

$$\mathcal{L} = \mathcal{L}_{CLS} + \lambda\mathcal{L}_{mpmmd}, \tag{12}$$

where $\mathcal{L}_{CLS}$ is the classification loss and $\lambda$ is the trade-off weight factor for $\mathcal{L}_{mpmmd}$.

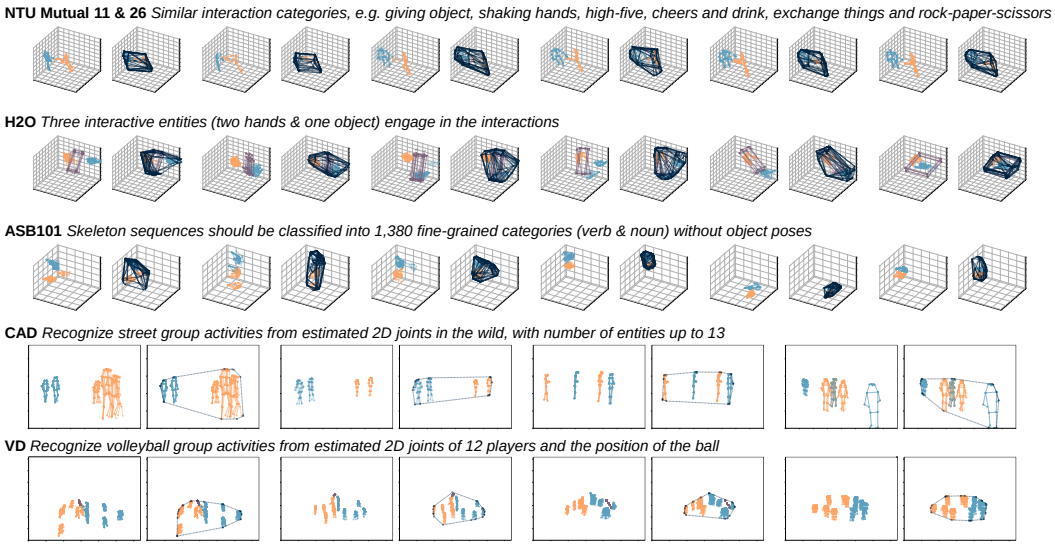



Figure 3: **Visualizations of multi-entity action samples and their skeleton convex hulls.**

# 4 Experiments

## 4.1 Datasets & Settings

We conduct experiments on six multi-entity action recognition datasets. Fig. 3 presents skeletal samples in these datasets and their skeleton convex hulls, showcasing their difficulties.

**NTU Mutual 11** and **NTU Mutual 26**, respectively subsets of **NTU RGB+D** [41] and **NTU RGB+D 120** [42], consist of a variety of inter-person mutual actions. NTU Mutual 11 adopts the widely-used X-Sub and X-View criteria, while NTU Mutual 26 follows the X-Sub and X-Set criteria.

**H2O** [13] proffers 3D poses of human hands and bounding boxes of the manipulated objects, facilitating both hand-to-hand and hand-to-object interactions learning. We follow the training, validation, and test splits outlined in [13] in our experiments.

**Assembly101 (ASB101)** [12] is a large and challenging 3D manual procedural activity dataset, with 1,380 categories of interactive actions. We follow the training, validation, and test splits described in [12] for evaluations. Fine-grained actions (verb & noun) are adopted as labels in experiments.

**Collective Activity Dataset (CAD)** [85] captures people and their behaviors in public using street cameras, categorizing pedestrian collective activities into 4 groups. We adopt the same categories, individual labels, train-test split in [95]. Only 2D joint coordinates are used in our experiments.

**Volleyball Dataset (VD)** [86] consists of video clips from volleyball tournaments and includes 8 group activity classes based on volleyball terminology. We follow the Original split described in [95] for evaluation. Only estimated 2D joint coordinates are used as input features.

**Settings.** Experiments are conducted on the GeForce RTX 3070 GPUs with PyTorch. CTR-GCN [36], InfoGCN [37], STSA-Net [40] and HD-GCN [38] are chosen as our baseline models. To ensure fair comparisons, we adopt single intra-skeleton modality without multi-modality fusion following [35]. For CTR-GCN in NTU Mutual 26, we adopt input shape $X \in \mathbb{R}^{3\times64\times25\times2}$, segment size $(1, 1, 1)$ and $\lambda = 0.1$ in CHASE. SGD optimizer is used with Nesterov momentum of 0.9, a initial learning rate of 0.1 and a decay rate 0.1 at the 80th and 100th epoch. Batch size is set to 64. More detailed configurations for each model are provided in the Appendix.

## 4.2 Experimental Results

Table 1 shows the experimental results on different benchmarks, reporting the averaged top-1 accuracy and its standard deviation in runs with several seed initializations. We compare CHASE with vanilla counterparts (light red background) and the state-of-the-art multi-entity action recognition methods (light yellow background). By adopting our proposed CHASE, we can boost the vanilla counterparts'

Table 1: Comparisons with Skeleton-based Methods on Multi-Entity Action Datasets

| Method | Venue | NTU Mutual 26(%) | | NTU Mutual 11(%) | |
|---|---|---|---|---|---|
| | | X-Sub | X-Set | X-Sub | X-View |
| GDCN [11] | TPAMI'23 | 85.80 | 92.10 | - | - |
| SkeleTR [76] | ICCV'23 | 87.80 | 88.30 | 94.80 | 97.70 |
| ISTA-Net [35] | IROS'23 | $90.56_{(\pm 0.08)}$ | $91.72_{(\pm 0.30)}$ | - | - |
| AHNet-Large [83] | PR'24 | 86.43 | 86.64 | 90.85 | 93.38 |
| me-GCN [77] | arXiv'24 | 90.00 | 90.00 | 95.50 | 98.20 |
| CTR-GCN [36] | ICCV'21 | $89.32_{(\pm 0.06)}$ | $90.19_{(\pm 0.17)}$ | $95.94_{(\pm 0.36)}$ | $98.32_{(\pm 0.29)}$ |
| **+ CHASE (Ours)** | - | $\mathbf{91.30}^{\uparrow 1.98}_{(\pm 0.22)}$ | $\mathbf{92.34}^{\uparrow 2.15}_{(\pm 0.10)}$ | $\mathbf{96.45}^{\uparrow 0.51}_{(\pm 0.05)}$ | $\mathbf{98.83}^{\uparrow 0.51}_{(\pm 0.13)}$ |
| InfoGCN [37](k=1) | CVPR'22 | $90.22_{(\pm 0.13)}$ | $91.13_{(\pm 0.16)}$ | $95.51_{(\pm 0.10)}$ | $97.76_{(\pm 0.22)}$ |
| **+ CHASE (Ours)** | - | $\mathbf{91.86}^{\uparrow 1.64}_{(\pm 0.05)}$ | $\mathbf{92.41}^{\uparrow 1.28}_{(\pm 0.34)}$ | $\mathbf{96.35}^{\uparrow 0.84}_{(\pm 0.18)}$ | $\mathbf{98.25}^{\uparrow 0.49}_{(\pm 0.25)}$ |
| STSA-Net [40] | Neuro.'23 | $88.41_{(\pm 0.01)}$ | $90.19_{(\pm 0.11)}$ | $95.96_{(\pm 0.09)}$ | $98.47_{(\pm 0.09)}$ |
| **+ CHASE (Ours)** | - | $\mathbf{89.77}^{\uparrow 1.36}_{(\pm 0.18)}$ | $\mathbf{91.54}^{\uparrow 1.35}_{(\pm 0.12)}$ | $\mathbf{96.63}^{\uparrow 0.68}_{(\pm 0.10)}$ | $\mathbf{98.73}^{\uparrow 0.26}_{(\pm 0.08)}$ |
| HD-GCN [38](CoM=1) | ICCV'23 | $88.25_{(\pm 0.44)}$ | $90.08_{(\pm 0.12)}$ | $95.58_{(\pm 0.10)}$ | $97.93_{(\pm 0.07)}$ |
| **+ CHASE (Ours)** | - | $\mathbf{90.81}^{\uparrow 2.56}_{(\pm 0.13)}$ | $\mathbf{92.06}^{\uparrow 1.97}_{(\pm 0.21)}$ | $\mathbf{96.22}^{\uparrow 0.64}_{(\pm 0.05)}$ | $\mathbf{98.31}^{\uparrow 0.38}_{(\pm 0.07)}$ |

| Method | Venue | H2O(%) | ASB101(%) | CAD(%) | VD(%) |
|---|---|---|---|---|---|
| AT [26] | CVPR'20 | - | - | - | 92.30 |
| ISTA-Net [35] | IROS'23 | $89.09_{(\pm 1.21)}$ | $28.01_{(\pm 0.06)}$ | $87.16_{(\pm 2.55)}$ | $91.40_{(\pm 0.23)}$ |
| H2OTR [80] | CVPR'23 | 90.90 | - | - | - |
| EffHandEgoNet [81] | arXiv'24 | 91.32 | - | - | - |
| AHNet-Large [83] | PR'24 | - | - | 89.32 | 84.31 |
| CTR-GCN [36] | ICCV'21 | $81.68_{(\pm 0.85)}$ | $27.83_{(\pm 0.45)}$ | $80.45_{(\pm 2.29)}$ | $92.66_{(\pm 0.21)}$ |
| **+ CHASE (Ours)** | - | $\mathbf{91.05}^{\uparrow 9.37}_{(\pm 1.98)}$ | $\mathbf{28.03}^{\uparrow 0.21}_{(\pm 0.30)}$ | $\mathbf{89.61}^{\uparrow 9.16}_{(\pm 0.20)}$ | $\mathbf{92.89}^{\uparrow 0.24}_{(\pm 0.15)}$ |
| InfoGCN [37](k=1) | CVPR'22 | $76.24_{(\pm 3.93)}$ | $27.18_{(\pm 0.10)}$ | $83.07_{(\pm 0.46)}$ | $91.77_{(\pm 0.15)}$ |
| **+ CHASE (Ours)** | - | $\mathbf{83.47}^{\uparrow 7.23}_{(\pm 2.89)}$ | $\mathbf{27.36}^{\uparrow 0.18}_{(\pm 0.12)}$ | $\mathbf{84.18}^{\uparrow 1.11}_{(\pm 2.91)}$ | $\mathbf{92.00}^{\uparrow 0.23}_{(\pm 0.15)}$ |
| STSA-Net [40] | Neuro.'23 | $92.29_{(\pm 0.52)}$ | $27.70_{(\pm 0.19)}$ | $80.20_{(\pm 3.60)}$ | $92.52_{(\pm 0.52)}$ |
| **+ CHASE (Ours)** | - | $\mathbf{94.77}^{\uparrow 2.48}_{(\pm 1.36)}$ | $\mathbf{27.81}^{\uparrow 0.11}_{(\pm 0.13)}$ | $\mathbf{85.93}^{\uparrow 5.73}_{(\pm 2.46)}$ | $\mathbf{92.78}^{\uparrow 0.26}_{(\pm 0.41)}$ |
| HD-GCN [38](CoM=1) | ICCV'23 | $72.73_{(\pm 0.41)}$ | $27.31_{(\pm 0.36)}$ | $76.93_{(\pm 4.38)}$ | $91.32_{(\pm 0.02)}$ |
| **+ CHASE (Ours)** | - | $\mathbf{81.61}^{\uparrow 8.88}_{(\pm 1.03)}$ | $\mathbf{27.50}^{\uparrow 0.19}_{(\pm 0.24)}$ | $\mathbf{82.39}^{\uparrow 5.46}_{(\pm 1.61)}$ | $\mathbf{92.00}^{\uparrow 0.68}_{(\pm 0.07)}$ |

performance by a noticeable margin in most settings. It yields varying degrees of accuracy improvement across different baseline models and benchmarks, owing to differences in model parameter count, training objective, data scale, etc. Compared to models with complicated interaction designs, CHASE can help single action backbones achieve the state-of-the-art performance in interaction recognition by outperforming ISTA-Net [35], AHNet-Large [83], etc. In group activities recognition task, which is more challenging for single-entity backbones, CHASE can help achieve competitive performance. Fig. 4 visualizes that CHASE can effectively alleviate the potential inter-entity distribution discrepancies across a range of data scales, thereby ensuring robust backbone optimization and inference. UMAP [100] visualization in Fig. 5 demonstrates our proposed CHASE differentiate similar multi-entity actions better by assisting backbones to learn more distinctive representations.

### 4.3 Ablation Study

In this section, we conduct ablation studies on the widely-adopted benchmarks NTU Mutual 26 and NTU Mutual 11 with only joint modality.

**Comparison with other alternatives**. We compare our proposed CHASE with several alternatives as follows: 1) Vanilla: Use the raw world coordinates or pixel coordinates. 2) S2CoM: Shift

Table 2: Comparison with Other Alternatives

| Method | Acc (%) | $\Delta$ (%) |
|---|---|---|
| Vanilla | $89.32_{(\pm 0.06)}$ | - |
| S2CoM | $88.66_{(\pm 0.26)}$ | $-0.67$ |
| BatchNorm | $89.06_{(\pm 0.16)}$ | $-0.27$ |
| ER [35] | $89.34_{(\pm 0.15)}$ | $+0.02$ |
| Aug | $89.72_{(\pm 0.04)}$ | $+0.40$ |
| S2CoM†/STD | $90.29_{(\pm 0.06)}$ | $+0.97$ |
| S2CoM† | $90.79_{(\pm 0.10)}$ | $+1.47$ |
| **CHASE (Ours)** | $\mathbf{91.30}_{(\pm 0.22)}$ | $\mathbf{+1.98}$ |

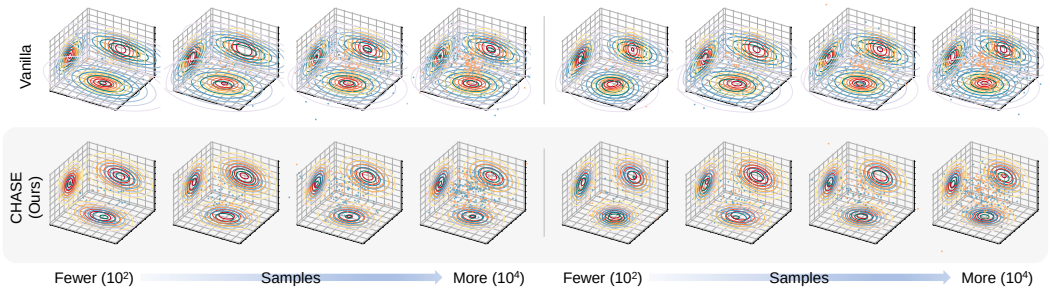



Figure 4: **Qualitative results of CHASE**. Different entity distributions are denoted by blue and orange. CHASE effectively mitigates inter-entity distribution discrepancies, demonstrating its clear effectiveness across a range of data scales, from small to large.

Table 3: Analysis of Inter-entity Distribution Discrepancies

| Set | Method | Avg KLD ↓ | JSD ↓ | BD ↓ | HD ↓ | MMD ↓ |
|---|---|---|---|---|---|---|
| I | Vanilla | $1.07_{(\pm 0.25)}$ | $0.19_{(\pm 0.04)}$ | $0.25_{(\pm 0.06)}$ | $0.46_{(\pm 0.06)}$ | $0.94_{(\pm 0.54)}$ |
| | **CHASE (Ours)** | $\mathbf{0.39}_{(\pm 0.09)}$ | $\mathbf{0.08}_{(\pm 0.02)}$ | $\mathbf{0.10}_{(\pm 0.02)}$ | $\mathbf{0.30}_{(\pm 0.03)}$ | $\mathbf{0.05}_{(\pm 0.02)}$ |
| II | Vanilla | $1.00_{(\pm 0.23)}$ | $0.18_{(\pm 0.04)}$ | $0.23_{(\pm 0.05)}$ | $0.45_{(\pm 0.05)}$ | $1.03_{(\pm 0.60)}$ |
| | **CHASE (Ours)** | $\mathbf{0.45}_{(\pm 0.08)}$ | $\mathbf{0.10}_{(\pm 0.02)}$ | $\mathbf{0.11}_{(\pm 0.02)}$ | $\mathbf{0.32}_{(\pm 0.03)}$ | $\mathbf{0.07}_{(\pm 0.02)}$ |
| III | Vanilla | $0.72_{(\pm 0.14)}$ | $0.14_{(\pm 0.02)}$ | $0.17_{(\pm 0.03)}$ | $0.39_{(\pm 0.04)}$ | $1.25_{(\pm 0.60)}$ |
| | **CHASE (Ours)** | $\mathbf{0.41}_{(\pm 0.08)}$ | $\mathbf{0.08}_{(\pm 0.02)}$ | $\mathbf{0.10}_{(\pm 0.02)}$ | $\mathbf{0.30}_{(\pm 0.03)}$ | $\mathbf{0.05}_{(\pm 0.04)}$ |
| IV | Vanilla | $0.75_{(\pm 0.14)}$ | $0.14_{(\pm 0.03)}$ | $0.17_{(\pm 0.03)}$ | $0.40_{(\pm 0.04)}$ | $1.15_{(\pm 0.56)}$ |
| | **CHASE (Ours)** | $\mathbf{0.41}_{(\pm 0.07)}$ | $\mathbf{0.08}_{(\pm 0.01)}$ | $\mathbf{0.09}_{(\pm 0.02)}$ | $\mathbf{0.30}_{(\pm 0.03)}$ | $\mathbf{0.04}_{(\pm 0.03)}$ |

the individual origins to the spatiotemporal centers of mass for each entity. 3) BatchNorm: Apply an additional BatchNorm operation immediately when batches of samples are fed into the model. 4) ER (Entity Rearrangement [35]): A technique aims to eliminate the orderliness of entities for interaction modelling. 5) Aug: Apply an additional data augmentation by randomly shifting the skeleton sequences. 6) S2CoM†: Shift the origin to the spatiotemporal center of mass. 7) S2CoM†/STD: Scale according to the channel-wise standard deviations after applying S2CoM†. Results in Table 2 indicate that CHASE can outperform these alternatives by bringing the largest accuracy improvement to the vanilla CTR-GCN.

**Analysis of inter-entity distribution discrepancies.** Table 3 presents metrics evaluating the inter-entity distribution discrepancies on test sets, including Averaged Kullback-Leibler Divergence (Avg KLD), Jensen-Shannon Divergence (JSD), Bhattacharyya Distance (BD), Hellinger Distance (HD) and MMD. We measure the pair-wise distributions of sampled data points from different entities in test sets of NTU Mutual 11 X-Sub (I), X-View (II) and NTU Mutual 26 X-Sub (III), X-Set (IV). Table 3 demonstrates that CHASE significantly minimizes discrepancies across all evaluation metrics, thereby benefiting backbone learning for each entity in multi-entity actions.

**Analysis of Key Components**. We validate the effectiveness of each key component in Table 4. When removing the skeleton convex hull constraint (CHC), there is a significant drop in accuracy, exceeding 60%, for initial learning rates (lr) of 0.1 and 0.01. This substantial decline highlights the importance of CHC as a critical constraint for learning the adaptive shift. Additionally, replacing Adaptive Shift (AS) with $\hat{X} = XWJ_{1,U}$ results in a dramatic decrease in accuracy, indicating that simply adding an equivalent number of trainable parameters without an adaptive shift formulation is ineffective. Table 4 further shows CHASE also benefits from CLB and MPMMD.

Table 4: Analysis of Key Components in CHASE

| ICHAS | | CLB | MPMMD | lr | Acc (%) | Δ (%) |
|---|---|---|---|---|---|---|
| AS | CHC | | | | | |
| ✓ | ✓ | ✓ | ✓ | 0.1 | $\mathbf{91.30}_{(\pm 0.22)}$ | - |
| ✓ | | ✓ | ✓ | 0.1 | $22.65_{(\pm 0.35)}$ | −68.65 |
| ✓ | | ✓ | ✓ | 0.01 | $86.99_{(\pm 0.16)}$ | −4.32 |
| ✓ | ✓ | | ✓ | 0.1 | $91.20_{(\pm 0.13)}$ | −0.10 |
| ✓ | | | ✓ | 0.1 | $22.75_{(\pm 0.12)}$ | −68.56 |
| ✓ | | | ✓ | 0.01 | $23.51_{(\pm 0.38)}$ | −67.79 |
| | ✓ | ✓ | ✓ | 0.1 | $20.42_{(\pm 0.09)}$ | −70.88 |
| ✓ | ✓ | ✓ | | 0.1 | $91.17_{(\pm 0.18)}$ | −0.13 |
| | | | | 0.1 | $89.50_{(\pm 0.14)}$ | −1.81 |

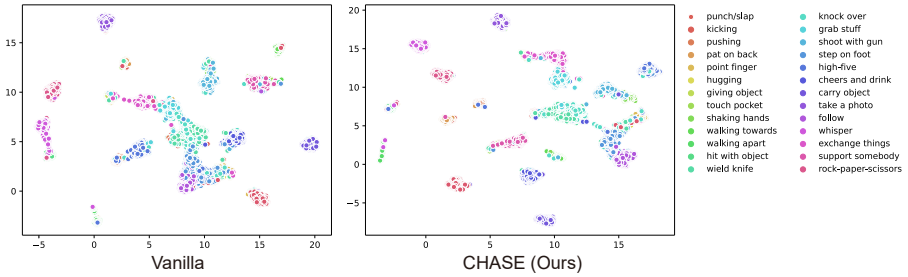



Figure 5: **UMAP [100] visualizations of multi-entity skeleton sequence representations on the test split of NTU Mutual 26 X-Sub**. Compared with Vanilla, our proposed CHASE differentiate similar multi-entity actions better by assisting backbones to learn more distinctive representations.

Table 5: Mixed Recognition on NTU RGB+D 120

| Method | X-Sub (%) | X-Set (%) |
|---|---|---|
| CTR-GCN [36] | $84.95_{(\pm 0.05)}$ | $86.90_{(\pm 0.03)}$ |
| **+ CHASE** | $\mathbf{85.36}_{(\pm 0.05)}$ | $\mathbf{86.95}_{(\pm 0.10)}$ |

**Evaluations on Mixed Recognition of Single- & Multi-Entity Actions**. Table 5 shows a 0.41% improvement in X-Sub accuracy on the entire NTU RGB+D 120. This implies that although CHASE is proposed for multi-entity actions, it is also effective in mixed recognition settings.

**Analysis of Efficiency**. As illustrated in Table 6, the number of trainable parameters of CHASE in NTU Mutual 26 configurations is about 26.37 k, resulting in a mere 1%-2% parameter increase. For computational complexity, FLOPs of CHASE is approximately 2.50 M. These metrics demonstrate that CHASE is both efficient and lightweight.

Table 6: CHASE Trainable Parameters

| Method | # Param. (M) |
|---|---|
| CTR-GCN [36] | 1.44 |
| **+ CHASE** | $1.46^{\uparrow 1.83\%}$ |
| STSA-Net [40] | 4.13 |
| **+ CHASE** | $4.16^{\uparrow 0.60\%}$ |

## 5 Conclusion

This paper proposes the Convex Hull Adaptive Shift for Multi-Entity Action Recognition (CHASE) to address the inter-entity distribution discrepancies. To the best of our knowledge, we are the first to investigate this problem and leverage discrepancy minimization to unbias the classifiers. Our approach can seamlessly adapt to existing backbone architectures and demonstrate performance improvements across six multi-entity action recognition datasets.

## Acknowledgments and Disclosure of Funding

This work was supported by Natural Science Foundation of Shenzhen (No. JCYJ20230807120801002) and National Natural Science Foundation of China (No. 62203476, No. 52105079).

## References


[1] Pengfei Wei, Lingdong Kong, Xinghua Qu, Yi Ren, Zhiqiang Xu, Jing Jiang, and Xiang Yin. Unsupervised video domain adaptation for action recognition: A disentanglement perspective. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[2] Dongho Lee, Jongseo Lee, and Jinwoo Choi. CAST: Cross-attention in space and time for video action recognition. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[3] Filip Ilic, He Zhao, Thomas Pock, and Richard P. Wildes. Selective, interpretable and motion consistent privacy attribute obfuscation for action recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[4] Jiazhou Zhou, Xu Zheng, Yuanhuiyi Lyu, and Lin Wang. Exact: Language-guided conceptual reasoning and uncertainty estimation for event-based action recognition and more. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[5] Duo Peng, Li Xu, Qiuhong Ke, Ping Hu, and Jun Liu. Joint attribute and model generalization learning for privacy-preserving action recognition. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[6] Bruce X.B. Yu, Yan Liu, Xiang Zhang, Sheng-hua Zhong, and Keith C.C. Chan. Mmnet: A model-based multimodal network for human action recognition in rgb-d videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3522–3538, 2023.

[7] Junwei Liang, Enwei Zhang, Jun Zhang, and Chunhua Shen. Multi-dataset training of transformers for robust action recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[8] Jinfu Liu, Chen Chen, and Mengyuan Liu. Multi-modality co-learning for efficient skeleton-based action recognition. In *Proceedings of the ACM Multimedia (ACM MM)*, 2024.

[9] Jinfu Liu, Runwei Ding, Yuhang Wen, Nan Dai, Fanyang Meng, Fang-Lue Zhang, Shen Zhao, and Mengyuan Liu. Explore human parsing modality for action recognition. *CAAI Transactions on Intelligence Technology*, 2024.

[10] Yunsheng Pang, Qiuhong Ke, Hossein Rahmani, James Bailey, and Jun Liu. Igformer: Interaction graph transformer for skeleton-based human interaction recognition. In *Proceedings of the 17th European Conference on Computer Vision (ECCV)*, pages 605–622, 2022.

[11] Shuai Li, Xinxue He, Wenfeng Song, Aimin Hao, and Hong Qin. Graph diffusion convolutional network for skeleton based semantic recognition of two-person actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8477–8493, 2023.

[12] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21064–21074, 2022.

[13] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10138–10148, October 2021.

[14] Mauricio Perez, Jun Liu, and Alex C. Kot. Interaction relational network for mutual action recognition. *IEEE Transactions on Multimedia*, 24:366–376, 2022.

[15] Xinshun Wang, Zhongbin Fang, Xia Li, Xiangtai Li, Chen Chen, and Mengyuan Liu. Skeleton-in-context: Unified skeleton sequence modeling with in-context learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[16] Kunyu Peng, Cheng Yin, Junwei Zheng, Ruiping Liu, David Schneider, Jiaming Zhang, Kailun Yang, M. Saquib Sarfraz, Rainer Stiefelhagen, and Alina Roitberg. Navigating open set scenarios for skeleton-based action recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5):4487–4496, Mar. 2024.

[17] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15085–15099, October 2023.

[18] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2969–2978, June 2022.

[19] Xingzhe He, Bastian Wandt, and Helge Rhodin. Autolink: Self-supervised learning of human skeletons and object outlines by linking keypoints. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[20] Jun Liu, Amir Shahroudy, Gang Wang, Ling-Yu Duan, and Alex C. Kot. Skeleton-based online action prediction using scale selection network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(6):1453–1467, 2020.

[21] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, Christoph Feichtenhofer, and Jitendra Malik. On the benefits of 3d pose and tracking for human action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 640–649, June 2023.

[22] Fan-Yun Sun, Isaac Kauvar, Ruohan Zhang, Jiachen Li, Mykel Kochenderfer, Jiajun Wu, and Nick Haber. Interaction modeling with multiplex attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[23] Simindokht Jahangard, Zhixi Cai, Shiki Wen, and Hamid Rezatofighi. Jrdb-social: A multifaceted robotic dataset for understanding of context and dynamics of human interactions within social groups. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[24] Timur Bagautdinov, Alexandre Alahi, Francois Fleuret, Pascal Fua, and Silvio Savarese. Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[25] Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. *arXiv:2403.08629*, 2024.

[26] Kirill Gavrilyuk, Ryan Sanford, Mehrsan Javan, and Cees G. M. Snoek. Actor-transformers for group activity recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 836–845, 2020.

[27] Hongwei Ren, Yue Zhou, Haotian FU, Yulong Huang, Xiaopeng LIN, Jie Song, and Bojun Cheng. Spikepoint: An efficient point-based spiking neural network for event cameras action recognition. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[28] Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. You2me: Inferring body pose in egocentric video via first and second person interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[29] Fangzhou Hong, Liang Pan, Zhongang Cai, and Ziwei Liu. Versatile multi-modal pre-training for human-centric perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16156–16166, June 2022.

[30] Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, and Xiuyu Sun. Beyond appearance: A semantic controllable self-supervised learning framework for human-centric visual tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15050–15061, June 2023.

[31] Shixiang Tang, Cheng Chen, Qingsong Xie, Meilin Chen, Yizhou Wang, Yuanzheng Ci, Lei Bai, Feng Zhu, Haiyang Yang, Li Yi, Rui Zhao, and Wanli Ouyang. Humanbench: Towards general human-centric perception with projector assisted pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21970–21982, June 2023.

[32] Junkun Yuan, Xinyu Zhang, Hao Zhou, Jian Wang, Zhongwei Qiu, Zhiyin Shao, Shaofeng Zhang, Sifan Long, Kun Kuang, Kun Yao, Junyu Han, Errui Ding, Lanfen Lin, Fei Wu, and Jingdong Wang. HAP: Structure-aware masked image modeling for human-centric perception. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[33] Yuanzheng Ci, Yizhou Wang, Meilin Chen, Shixiang Tang, Lei Bai, Feng Zhu, Rui Zhao, Fengwei Yu, Donglian Qi, and Wanli Ouyang. Unihcp: A unified model for human-centric perceptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17840–17852, June 2023.

[34] Yizhou Wang, Yixuan Wu, Shixiang Tang, Weizhen He, Xun Guo, Feng Zhu, Lei Bai, Rui Zhao, Jian Wu, Tong He, et al. Hulk: A universal knowledge translator for human-centric tasks. *arXiv:2312.01697*, 2023.

[35] Yuhang Wen, Zixuan Tang, Yunsheng Pang, Beichen Ding, and Mengyuan Liu. Interactive spatiotemporal token attention network for skeleton-based general interactive action recognition. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7886–7892, 2023.

[36] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *IEEE International Conference on Computer Vision (ICCV)*, pages 13359–13368, 2021.

[37] Hyung-Gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20154–20164, 2022.

[38] Jungho Lee, Minhyeok Lee, Dogyoon Lee, and Sangyoun Lee. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10444–10453, October 2023.

[39] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition. In *15th Asian Conference on Computer Vision (ACCV)*, page 38–53, 2020.

[40] Helei Qiu, Biao Hou, Bo Ren, and Xiaohua Zhang. Spatio-temporal segments attention for skeleton-based action recognition. *Neurocomputing*, 518:30–38, 2023. ISSN 0925-2312.

[41] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1010–1019, 2016.

[42] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2684–2701, 2020.

[43] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning, and recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, page 2649–2656, 2014.

[44] Liu Chunhui, Hu Yueyu, Li Yanghao, Song Sijie, and Liu Jiaying. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. *ACM Multimedia workshop*, 2017.

[45] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *Proceedings of the 17th European Conference on Computer Vision (ECCV)*, pages 816–833, 2016.

[46] Wentao Zhu, Cuiling Lan, Junliang Xing, Wenjun Zeng, Yanghao Li, Li Shen, and Xiaohui Xie. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 3697–3703. AAAI Press, 2016.

[47] Jun Liu, Gang Wang, Ping Hu, Ling-Yu Duan, and Alex C. Kot. Global context-aware attention lstm networks for 3d action recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3671–3680, 2017.

[48] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2136–2145, 2017.

[49] Jun Liu, Gang Wang, Ling-Yu Duan, Kamila Abdiyeva, and Alex C. Kot. Skeleton-based human action recognition with global context-aware attention lstm networks. *IEEE Transactions on Image Processing*, 27(4):1586–1599, 2018.

[50] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View adaptive neural networks for high performance skeleton-based human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1963–1978, 2019.

[51] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, AAAI'18, 2018.

[52] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3590–3598, 2019.

[53] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12018–12027, 2019.

[54] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 140–149, 2020.

[55] Xiaoke Hao, Jie Li, Yingchun Guo, Tao Jiang, and Ming Yu. Hypergraph neural network for skeleton-based action recognition. *IEEE Transactions on Image Processing*, 30:2263–2275, 2021.

[56] Haodong Duan, Jiaqi Wang, Kai Chen, and Dahua Lin. Pyskl: Towards good practices for skeleton action recognition. In *Proceedings of the 30th ACM International Conference on Multimedia (ACMMM)*, pages 7351–7354, 2022.

[57] Dongjingdin Liu, Pengpeng Chen, Miao Yao, Yijing Lu, Zijie Cai, and Yuxin Tian. Tsgcnext: Dynamic-static multi-graph convolution for efficient skeleton-based action recognition with long-term learning potential. *arXiv:2304.11631*, 2023.

[58] Woomin Myung, Nan Su, Jing-Hao Xue, and Guijin Wang. Degcn: Deformable graph convolutional networks for skeleton-based action recognition. *IEEE Transactions on Image Processing*, 33:2477–2490, 2024.

[59] Yuxuan Zhou, Zhi-Qi Cheng, Chao Li, Yifeng Geng, Xuansong Xie, and Margret Keuper. Hypergraph transformer for skeleton-based action recognition. *arXiv:2211.09590*, 2022.

[60] Qingtian Wang, Shuze Shi, Jiabin He, Jianlin Peng, Tingxi Liu, and Renliang Weng. Iip-transformer: Intra-inter-part transformer for skeleton-based action recognition. In *IEEE International Conference on Big Data (BigData)*, pages 936–945, 2023.

[61] Nguyen Huu Bao Long. Step catformer: Spatial-temporal effective body-part cross attention transformer for skeleton-based action recognition. *arXiv:2312.03288*, 2023.

[62] Yunyao Mao, Jiajun Deng, Wengang Zhou, Yao Fang, Wanli Ouyang, and Houqiang Li. Masked motion predictors are strong 3d action representation learners. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10181–10191, October 2023.

[63] Neel Trivedi and Ravi Kiran Sarvadevabhatla. Psumnet: Unified modality part streams are all you need for efficient pose-based action recognition. In *Proceedings of the 17th European Conference on Computer Vision (ECCV)*, page 211–227, 2022.

[64] Jeonghyeok Do and Munchurl Kim. Skateformer: Skeletal-temporal transformer for human action recognition. *arXiv:2403.09508*, 2024.

[65] Xiaohu Huang, Hao Zhou, Bin Feng, Xinggang Wang, Wenyu Liu, Jian Wang, Haocheng Feng, Junyu Han, Errui Ding, and Jingdong Wang. Graph contrastive learning for skeleton-based action recognition. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[66] Haojun Xu, Yan Gao, Zheng Hui, Jie Li, and Xinbo Gao. Language knowledge-assisted representation learning for skeleton-based action recognition. *CoRR*, abs/2305.12398, 2023.

[67] Wangmeng Xiang, Chao Li, Yuxuan Zhou, Biao Wang, and Lei Zhang. Generative action description prompts for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10276–10285, October 2023.

[68] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L. Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–35, 2012.

[69] Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Multi-person extreme motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13053–13064, June 2022.

[70] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17016–17027, June 2023.

[71] Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, Yunhui Liu, Wenjun Zeng, and Xiaokang Yang. Inter-x: Towards versatile human-human interaction analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[72] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision*, Mar 2024.

[73] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[74] Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. Assemblyhands: Towards egocentric activity understanding via 3d hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12999–13008, June 2023.

[75] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+o: Unified egocentric recognition of 3d hand-object poses and interactions. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4506–4515, 2019.

[76] Haodong Duan, Mingze Xu, Bing Shuai, Davide Modolo, Zhuowen Tu, Joseph Tighe, and Alessandro Bergamo. Skeletr: Towards skeleton-based action recognition in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13634–13644, October 2023.

[77] Mengyuan Liu, Chen Chen, Songtao Wu, Fanyang Meng, and Hong Liu. Learning mutual excitation for hand-to-hand and human-to-human interaction recognition. *arXiv:2402.02431*, 2024.

[78] Yilin Wen, Hao Pan, Lei Yang, Jia Pan, Taku Komura, and Wenping Wang. Hierarchical temporal transformer for 3d hand pose estimation and action recognition from egocentric rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21243–21253, June 2023.

[79] Md Salman Shamil, Dibyadip Chatterjee, Fadime Sener, Shugao Ma, and Angela Yao. On the utility of 3d hand poses for action recognition. *arXiv:2403.09805*, 2024.

[80] Hoseong Cho, Chanwoo Kim, Jihyeon Kim, Seongyeong Lee, Elkhan Ismayilzada, and Seungryul Baek. Transformer-based unified recognition of two hands manipulating objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4769–4778, June 2023.

[81] Wiktor Mucha and Martin Kampel. In my perspective, in my hands: Accurate egocentric 2d hand pose and action recognition. *arXiv:2404.09308*, 2024.

[82] Tian Lan, Yang Wang, Weilong Yang, Stephen N. Robinovitch, and Greg Mori. Discriminative latent models for recognizing contextual group activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(8):1549–1562, 2012.

[83] Guoquan Wang, Mengyuan Liu, Hong Liu, Peini Guo, Ti Wang, Jingwen Guo, and Ruijia Fan. Augmented skeleton sequences with hypergraph network for self-supervised group activity recognition. *Pattern Recognition*, 152:110478, 2024.

[84] Che-Jui Chang, Danrui Li, Deep Patel, Parth Goel, Honglu Zhou, Seonghyeon Moon, Samuel S. Sohn, Sejong Yoon, Vladimir Pavlovic, and Mubbasir Kapadia. M3act: Learning from synthetic human group activities. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[85] Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing? : Collective activity classification using spatio-temporal relationship among people. In *IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 1282–1289, 2009.

[86] Mostafa S. Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1971–1980, 2016.

[87] Sina Mokhtarzadeh Azar, Mina Ghadimi Atigh, Ahmad Nickabadi, and Alexandre Alahi. Convolutional relational machine for group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[88] Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning actor relation graphs for group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[89] Kirill Gavrilyuk, Ryan Sanford, Mehrsan Javan, and Cees G. M. Snoek. Actor-transformers for group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[90] Shuaicheng Li, Qianggang Cao, Lingbo Liu, Kunlin Yang, Shinan Liu, Jun Hou, and Shuai Yi. Groupformer: Group activity recognition with clustered spatial-temporal transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13668–13677, October 2021.

[91] Hangjie Yuan and Dong Ni. Learning visual context for group activity recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):3261–3269, May 2021.

[92] Masato Tamura, Rahul Vishwakarma, and Ravigopal Vennelakanti. Hunting group clues with transformers for social group activity recognition. In *Proceedings of the 17th European Conference on Computer Vision (ECCV)*, page 19–35, 2022.

[93] Mingfei Han, David Junhao Zhang, Yali Wang, Rui Yan, Lina Yao, Xiaojun Chang, and Yu Qiao. Dual-ai: Dual-path actor interaction learning for group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2990–2999, June 2022.

[94] Haritha Thilakarathne, Aiden Nibali, Zhen He, and Stuart Morgan. Pose is all you need: the pose only group activity recognition system (POGARS). *Machine Vision and Applications*, 33 (6):95, October 2022.

[95] Honglu Zhou, Asim Kadav, Aviv Shamsian, Shijie Geng, Farley Lai, Long Zhao, Ting Liu, Mubbasir Kapadia, and Hans Peter Graf. Composer: Compositional reasoning of group activity in videos with keypoint-only modality. In *Proceedings of the 17th European Conference on Computer Vision (ECCV)*, 2022.

[96] Hangjie Yuan, Dong Ni, and Mang Wang. Spatio-temporal dynamic inference network for group activity recognition. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7456–7465, 2021.

[97] Ralph Tyrell Rockafellar. *Convex analysis*. Princeton Landmarks in Mathematics and Physics. Princeton University Press, Princeton, NJ, December 1996.

[98] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018.

[99] Rupesh K Srivastava, Klaus Greff, and Jürgen Schmidhuber. Training very deep networks. In *Twenty-eighth Conference on Neural Information Processing Systems (NeurIPS)*, volume 28, 2015.

[100] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018.

# Appendix

The appendix is organized as follows:


A. Supplementary Analysis of CHASE
B. Code for CHASE
C. Details of Multi-Entity Action Recognition Datasets
D. Evaluation Metrics
E. Implementation & Configuration Details
F. Supplementary Experimental Results
G. Limitations & Broader Impacts
H. Licenses for Used Assets


## A Supplementary Analysis of CHASE

### A.1 Preliminaries

Here we clarify some crucial definitions in the skeleton-based multi-entity action recognition task as follows.

**Definition 2** (Skeleton Sequence of A Multi-Entity Action)**.** Suppose that $E$ interactive entities (e.g. persons) engage in a purposeful act during a period of time $T$, and the pose of each entity is indicated by $J$ joints with $C$ Cartesian coordinates. We can define the skeleton sequence of a multi-entity action as $X \in \mathbb{R}^{C \times T \times J \times E}$.

**Definition 3** (Skeleton-based Multi-Entity Action Recognition)**.** We define the task as finding the optimal estimator $\mathcal{E}_\theta$ of the mapping $\mathcal{E} : X \mapsto Y$, where $X$ is the skeleton sequence of a multi-entity action and $Y$ is its corresponding label.

**Definition 4** (Joints & Bones)**.** A joint within a skeleton is a point in the Cartesian coordinate system, which can also be viewed as a vector $\vec{p}_i (1 \leq i \leq J)$. A bone within a skeleton is a differential vector of two joints $\vec{p}_i$ and $\vec{p}_j$ $(1 \leq i, j \leq J)$, if and only if the two joints are connected or the same one according to a prior graph (e.g. the bone connection of the human body).

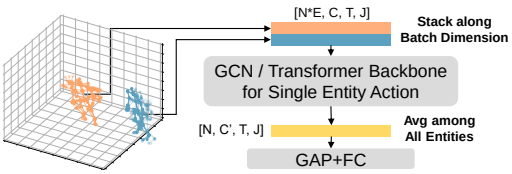



Figure 6: **The common practice in single-entity action recognition models to recognize multi-entity actions.** This late fusion strategy is used in many recent works [36, 37, 38, 39, 40].

We illustrate the common practice [36, 37, 38, 39, 40] when single-entity action recognition models meet multiple entities using Fig. 6. With vanilla world coordinates as input, they usually concatenate each entity along batch dimension, and extract high-dimensional features of each entity separately using GCN or transformer backbone. Subsequently, the individual features get averaged before undergoing global average pooling and full connection layers. Notably, this common practice is based on an empirical assumption that each entity is independent and identically distributed.

### A.2 Adaptive Shift Analysis Under Bone Representation

Most existing research uses Def. 4 for the bone modality definition [36, 53, 54, 56]. In this context, the learned adaptive shift is ineffective because shifting the origin does not affect the Euclidean distance between any two points. However, the adaptive shift could still prove effective if bones are defined differently.

Table 7: Statistics of Multi-Entity Action Recognition Datasets

| Datasets | Annotation | | | #Actions | #Joints | #Clips | #Valid Frames | #Entities | #Participants |
|---|---|---|---|---|---|---|---|---|---|
| | Body | Hand | Object | | | | | | |
| NTU Mutual 11 [41] | ✓ | | | 11 | 25 | 10,347 | 69.18 | 2.00 | 40 |
| NTU Mutual 26 [42] | ✓ | | | 26 | 25 | 24,732 | 59.36 | 2.00 | 106 |
| H2O [13] | | ✓ | ✓ | 36 | 21 | 933 | 97.29 | 3.00 | 4 |
| Assembly101 [12] | | ✓ | | 1,380 | 21 | 85,252 | 105.91 | 2.00 | 53 |
| CAD [85] | ✓ | | | 4 | 17 | 2,511 | 10.00 | 5.22 | - |
| VD [86] | ✓ | | ✓ | 8 | 17 | 4,830 | 10.00 | 13.00 | - |

**Definition 5** (k-hop Bones [37])**.** We denote $X_t$ as the joints at the moment $t$ in a skeleton sequence. The k-hop Bone $\tilde{X}_t^{(k)}$ at moment $t$ can be formulated as

$$\tilde{X}_t^{(k)} = (I - P^k)X_t, \tag{13}$$

where $k \geq 1$, and $P \in \mathbb{R}^{J \times J}$ is a binary adjacency matrix of a directed graph without bi-directional edges.

In the context of human body skeletons, Def. 5 uses the joint $j_r$, representing *the center of the spine*, as the root to construct the directed graph. If $k = max_v d(v) + 1$, where $max_v d(v)$ means the max hop of all joints to the root, then the k-hop Bone $\tilde{X}_t^{(k)}$ aligns with the joint definition in Def. 4. However, when $k = 1$, this equivalence doesn't necessarily hold for the bone definition in Def. 4. The reason is that some joints (such as the root) may have no in-degree, causing them to retain their original joint coordinates in Def. 5. Therefore, the k-hop bones may still contain joint information that can be modified through adaptive shift.

### A.3 Analysis of Gradient

Consider the expectation $\mathbb{E}_{r(z)}[g_\theta(z)]$ in Eq. 11, where $g_\theta$ denotes the composition of MMD (Eq. 9) and ICHAS (Eq. 6) with CLB (Eq. 8). We assume that the gradient of $g_\theta$ with respect to the parameters $\theta$ exists. Since it adopts uniform sampling, we note that the discrete probability density function $r(z)$ of $z$ is independent of the parameters $\theta$. Thus we have

$$\begin{aligned} \nabla_\theta \mathbb{E}_{r(z)}[g_\theta(z)] &= \nabla_\theta[\sum_z r(z) g_\theta(z)] \\ &= \sum_z r(z)[\nabla_\theta g_\theta(z)] \\ &= \mathbb{E}_{r(z)}[\nabla_\theta g_\theta(z)], \end{aligned} \tag{14}$$

which indicates that the gradient of the expectation $\nabla_\theta \mathbb{E}_{r(z)}[g_\theta(z)]$ is equivalent to the expectation of the gradient $\mathbb{E}_{r(z)}[\nabla_\theta g_\theta(z)]$. The latter can be approximated using Monte Carlo methods.

## B Code for CHASE

CHASE is implemented as a wrapper for various skeleton-based action backbones, as illustrated in Algorithm 1. Line 18 indicates input of batch size $N$, channel $C$, number of frames $T$, number of joints $V$, number of entity $M$. Line 19 represents Eq. 8, which is to map $X$ to $W$ with function $\psi$. Line 21 represents the formulation of $\vec{p^*}$ in Eq. 3. Line 25 indicate the subtraction in Eq. 1. Line 26-27 are mini-batch sampling strategy for the Mini-batch Pair-wise Maximum Mean Discrepancy Loss. Line 28 represents the single-entity backbone. Our code is publicly available at `https://github.com/Necolizer/CHASE` with MIT license.

## C Details of Multi-Entity Action Recognition Datasets

We conduct experiments on a range of datasets, including **NTU Mutual 11** (a subset of **NTU RGB+D** [41]), **NTU Mutual 26** (a subset of **NTU RGB+D 120** [42]), **H2O** [13], **Assembly101 (ASB101)** [12], **Collective Activity Dataset (CAD)** [85], and **Volleyball Dataset (VD)** [86]. Table 7

**Algorithm 1** CHASE Wrapper: PyTorch-like Pseudo Code

```
class CHASEWrapper(nn.Module):
    def __init__(self, backbone, in_channels, num_frame, num_point, pooling_seg, num_entity, c1, c2):
        super(CHASEWrapper, self).__init__()

        out_channel = num_frame * num_point * num_entity
        self.pooling_seg = pooling_seg
        self.seg = self.pooling_seg[0]*self.pooling_seg[1]*self.pooling_seg[2]
        self.seg_num_list = [(num_frame//self.pooling_seg[0]), (num_point//self.pooling_seg[1]), (num_entity//self.pooling_seg[2])]
        self.seg_num = self.seg_num_list[0] * self.seg_num_list[1] * self.seg_num_list[2]
        self.shift = nn.Sequential(
            nn.Conv3d(in_channels=in_channels, out_channels=c1, kernel_size=1),
            nn.AdaptiveAvgPool3d((self.pooling_seg[0], self.pooling_seg[1], self.pooling_seg[2])),
            nn.Conv3d(c1, c2, 1, bias=False),
            nn.ReLU(inplace=True),
            nn.Conv3d(c2, out_channel, 1, bias=False),
        )
        self.backbone = backbone

    def forward(self, x):
        N, C, T, V, M = x.size()
        sf = self.shift(x).view(N, T*M*V, -1)
        tx = rearrange(x, 'n c t v m -> n c (t v m)', t=T, m=M, v=V).contiguous()
        sf = (tx @ sf.softmax(dim=1)).unsqueeze(-1).expand(-1, -1, self.seg, self.seg_num)
        sf = rearrange(sf, 'n c (T V M) (t v m) -> n c (T t) (V v) (M m)',
                T=self.pooling_seg[0], V=self.pooling_seg[1], M=self.pooling_seg[2],
                t=self.seg_num_list[0], v=self.seg_num_list[1], m=self.seg_num_list[2]).contiguous()
        x = x - sf
        if self.training:
            pairs = MiniBatchSampling(x)
        out = self.backbone(x)

        if self.training:
            return out, pairs
        else:
            return out
```

provides details about each dataset, including their annotation types, numbers of action categories, numbers of joints, numbers of clips (samples), averaged counts of valid frames, counts of entities engaging in a multi-entity action, and numbers of participants in data collection. For CAD [85] and VD [86], which both capture videos in the wild with a variety of individuals, it is difficult to determine the exact number of participants.

## D Evaluation Metrics

In this section, we provide the detailed formulation for metrics in ablation study. Given two discrete probability distributions $P$ and $Q$ defined on the same sample space $\mathcal{X}$, we can define the following metrics:

**Averaged Kullback-Leibler Divergence (Avg KLD)**:

$$\begin{aligned} Avg\ D_{KL} &= \frac{1}{2}[D_{KL}(P\|Q) + D_{KL}(Q\|P)] \\ &= \frac{1}{2}[\sum_{x\in\mathcal{X}} P(x)\log(\frac{P(x)}{Q(x)}) + \sum_{x\in\mathcal{X}} Q(x)\log(\frac{Q(x)}{P(x)})]. \end{aligned} \tag{15}$$

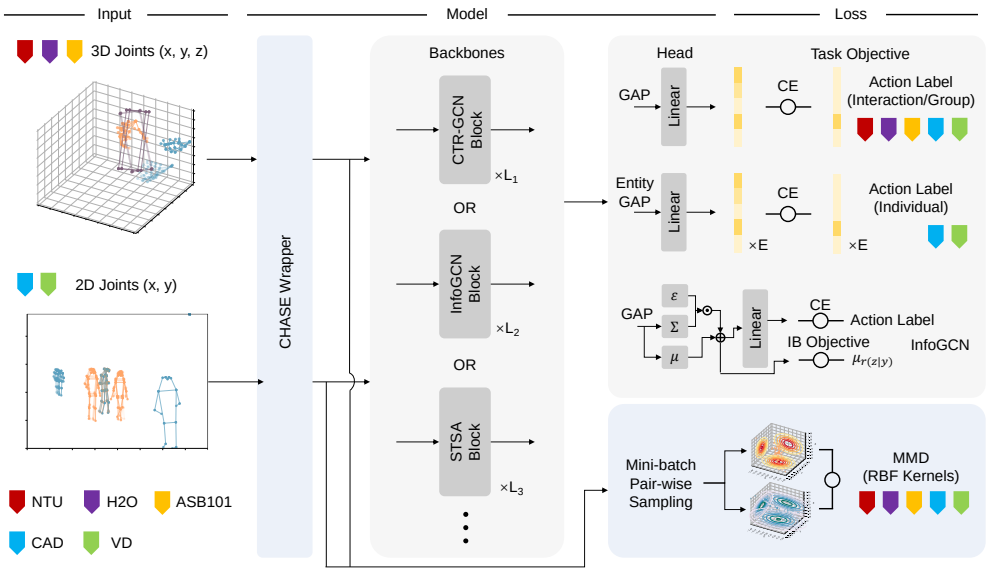



Figure 7: **Implementation details for different models and benchmarks.** CHASE can be adapted to various backbones including CTR-GCN [36], InfoGCN [37], STSA-Net [40], and HD-GCN [38]. We implement linear classification heads for all the models and benchmarks, except for InfoGCN [37]. Individual labels are utilized as an auxiliary classification objective to further improve the group action recognition performance in CAD [85] and ASB101 [12].

**Jensen-Shannon Divergence (JSD)**:

$$JSD(P\|Q) = \frac{1}{2}D(P\|M) + \frac{1}{2}D(Q\|M), \tag{16}$$

where $M = \frac{1}{2}(P+Q)$ is a mixture distribution of $P$ and $Q$.

**Bhattacharyya Distance (BD)**:

$$D_B(P,Q) = -\ln \sum_{x\in\mathcal{X}} \sqrt{P(x)Q(x)}. \tag{17}$$

**Hellinger Distance (HD)**:

$$H(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{x\in\mathcal{X}} (\sqrt{P(x)} - \sqrt{Q(x)})^2}. \tag{18}$$

**Maximum Mean Discrepancy (MMD)**:

$$\text{MMD}(P,Q) = \sup_{\|f\|_{\mathcal{H}}\leq 1} \left(\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]\right), \tag{19}$$

where $\mathbb{E}[f(\cdot)]$ is the expectation of any function $f$ in the RKHS $\mathcal{H}$.

When adopting these metrics in our experiments, the distributions are generated by the kernel density estimation (KDE). We sample the points 30 times with different seed initializations and report the averaged measurements.

# E Implementation & Configuration Details

In this section, we provide more details of our experimental setup and model implementation for each benchmark. Experiments are conducted with 8 GeForce RTX 3070 GPUs (GPU Memory: 8GB), using torch version 1.9.0+cu111, torchvision version 0.10.0+cu111, and CUDA version 11.4. CTR-GCN [36], InfoGCN [37], STSA-Net [40] and HD-GCN [38] are chosen as our baseline models. To ensure fair comparisons, we adopt single intra-skeleton modality without multi-modality fusion, following [35]. Our implementation details are illustrated in Fig. 7.

### E.1 Dataset-related Configurations

**NTU Mutual 11** [41] **& NTU Mutual 26** [42]. X-Sub and X-View criteria [41] are adopted in NTU Mutual 11, while X-Sub and X-Set criteria [42] are used in NTU Mutual 26. We evaluate models with only 3D joint inputs, applying data augmentations such as random rotation and spatial shift. During training and testing, we employ temporal cropping and resizing, adjusting based on the number of valid frames. Notably, we use distinct percentage intervals for training (0.5,1) and testing (0.95). For experiments conducted on the entire NTU-RGB+D 120 dataset, we maintain identical settings as those used for NTU Mutual 26.

**H2O** [13]. We follow the training, validation, and test splits described in [13]. To maintain consistency in GCNs, we use the hand graph structure, originally designed for human hands, for both hand entities and object entities. Models are evaluated with only 3D joint inputs. The same augmentations as NTU Mutual 26 are adopted in this benchmark.

**ASB101** [12]. We follow the training, validation, and test splits outlined in [12] for evaluations. 1,380 Fine-grained actions (verb & noun) are adopted as labels in experiments. We evaluate models with only 3D joint inputs. The same augmentations as NTU Mutual 26 are adopted in this dataset, except that the training percentage interval of the temporal cropping and resizing is set to (0.75,1).

**CAD** [85]. We adopt the same data augmentations, group action categories, individual labels and train-test split in [95]. But different from [95], only 2D joint coordinates are used in our experiments. Individual labels are used as an auxiliary classification objective, as presented in Fig. 7.

**VD** [86]. We follow the same data augmentations, group action categories, individual labels and Original train-test split in [95]. But different from [95], only 2D joint coordinates are used as input features in experiments. Besides, individual labels are leveraged as an auxiliary classification objective to further improve the group action recognition performance, as shown in Fig. 7. The volleyball position is represented as $X \in \mathbb{R}^{T \times C}$. To ensure inter-entity consistency, we apply padding to fit the shape $X \in \mathbb{R}^{C \times T \times J \times 1}$. To maintain consistency in GCNs, we employ human body graph structure for both human body entities and volleyball entities.

### E.2 Model-related Configurations

**CHASE**. We set default segment size $(1, 1, 1)$, $C_1 = 64$, $C_2 = 8$, $M = 1$ and $\lambda = 0.1$ in CHASE. In all experiments, we maintain consistent configurations for baseline models to ensure fair comparisons between models incorporating CHASE and their respective vanilla counterparts. To avoid unnecessary verbosity, we present implementation specifics solely for NTU Mutual 26. For training details pertaining to other benchmarks, please refer to the code repository at `https://github.com/Necolizer/CHASE`.

**CTR-GCN** [36]. Cross entropy is used as the recognition loss function. SGD optimizer is used with Nesterov momentum of 0.9, a initial learning rate of 0.1 and a decay rate 0.1 at the 80th and 100th epoch. Batch size is set to 64. With the first 5 warm-up epochs, the training process is terminated after 110 epochs.

**InfoGCN** [37]. By taking $k = 1$, InfoGCN adopts their definition of 1-hop Bones [37] as input. Cross entropy is used as the loss function with label smoothing factor 0.1 and temperature factor 1.0. The information bottleneck objective [37] is also employed as the auxiliary loss. Following [37], we set $\lambda_1 = 0.1$, $\lambda_2 = 0.0001$, and the $\mu_r(z|y)$ of each action class as random orthogonal vectors with a scale of 3. Diverging from conventional backbones, we substitute the standard classification head with the InfoGCN head, depicted in Fig. 7. SGD optimizer is used with Nesterov momentum of 0.9, a weight decay of 0.0005, a initial learning rate of 0.1 and a decay rate 0.1 at the 90th and 100th epoch. Batch size is set to 120. With the first 5 warm-up epochs, the training process is terminated after 110 epochs.

**STSA-Net** [40]. Cross entropy is used as the loss function with label smoothing factor 0.1 and temperature factor 1.0. SGD optimizer is used with Nesterov momentum of 0.9, a initial learning rate of 0.1 and a decay rate 0.1 at the 60th and 90th epoch. Batch size is set to 64. With the first 5 warm-up epochs, the training process is terminated after 110 epochs.

**HD-GCN** [38]. We set $CoM = 1$ in the hierarchy graph generation and evaluate on this setting. Cross entropy is used as the loss function. SGD optimizer is used with Nesterov momentum of 0.9, a

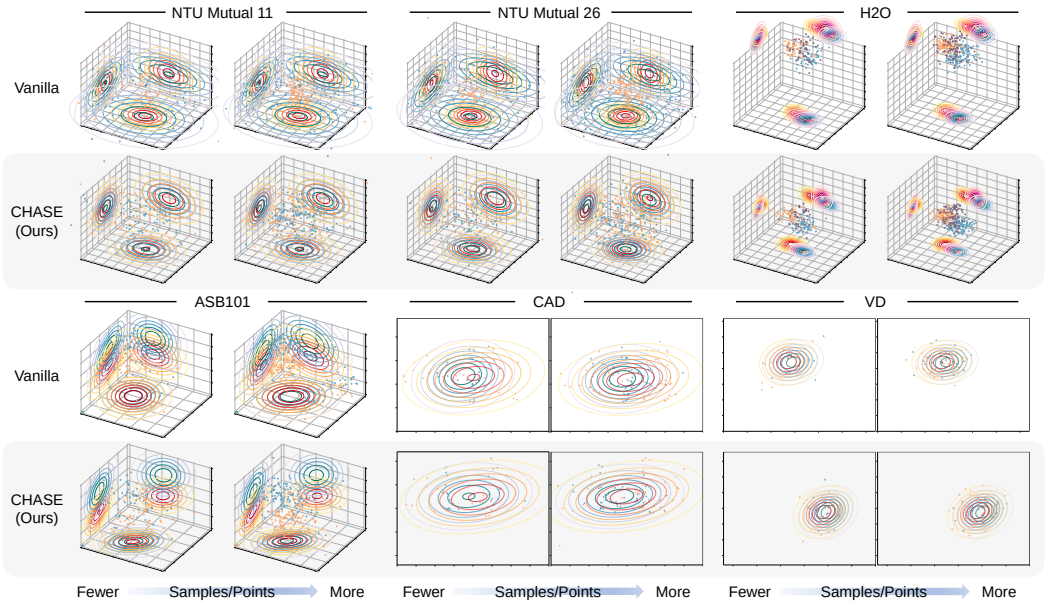



Figure 8: **Qualitative results on multi-entity action recognition datasets**. For visual clarity, we display 2 or 3 estimated entity distributions for each test set. Each subplot shows the projection of the multi-variant normal distributions generated by mean vectors and covariance matrices. Different entity distributions are denoted by distinct colors. CHASE effectively mitigates inter-entity distribution discrepancies while preserving potential entity orderliness in these datasets.

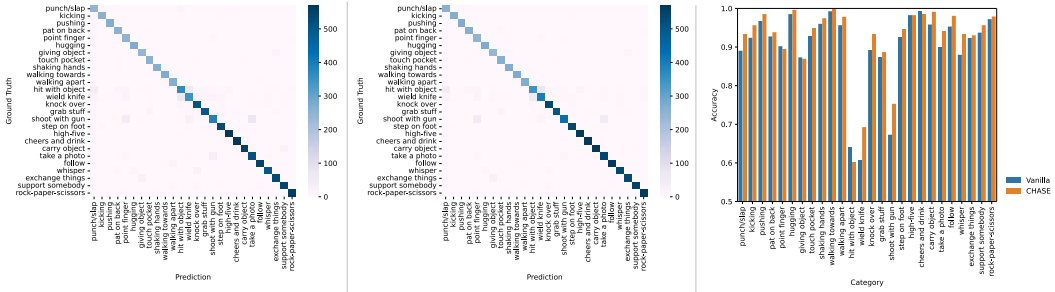
initial learning rate of 0.1 and a decay rate 0.1 at the 80th and 100th epoch. Batch size is set to 64. With the first 5 warm-up epochs, the training process is terminated after 110 epochs.

# F Supplementary Experimental Results

**More Analysis on Table 1.** In Table 1, we observe that CHASE yields varying degrees of accuracy improvement across different baseline models and benchmarks. The performance gains are influenced by both the backbone models and the datasets, as CHASE functions as an additional normalization step that mitigates data bias introduced by inter-entity distribution discrepancies. For baseline backbones, this is owing to differences in their backbone architecture design, parameter count and training objective. For example, STSA-Net [40] is a relatively large backbone based on transformer architecture, which does not rely the prior definition of the skeleton graphs. Its adaptability to different graph structures makes it outweigh GCN-based backbones in H2O benchmark. Another example is InfoGCN [37], which leverages an auxiliary information bottleneck objective in its training. Though this method is proven more effective than the other backbones in some person-to-person interaction settings, it doesn't ensure better performance in hand-to-object interaction and group activity benchmarks. For different benchmarks, it is owing to differences in data scale and label space (see Table 7). For example, ASB101 is an extremely challenging benchmark for its over 80,000 samples and 1,380 target categories. Therefore the accuracy improvement is modest compared with the other benchmarks.

**More Qualitative Results.** Fig. 8 visualizes how CHASE works with multi-entity skeletal sequences. By integrating CHASE, different entity distributions become more similar in both aspects of mean and covariance. It demonstrates that CHASE can lower the inter-entity distribution discrepancy, especially obviously in NTU Mutual 11, NTU Mutual 26, CAD, whose entities have no orderliness. For H2O, ASB101 and VD, whose entities are characterized by an intrinsic order (e.g. left hands, right hands, left-side volleyball players, right-side volleyball players, and objects), CHASE can also preserve their orderliness by letting the distributions be similar but different.

**Evaluations on Mixed Recognition of Single-Entity & Multi-Entity Actions.** Table 5 concludes the action recognition results on the entire NTU RGB+D 120 [42]. By integrating CHASE, the baseline model gets accuracy improvement by 0.41% and 0.05% on X-Sub and X-Set, respectively.




(a) Confusion Matrix (Vanilla) (b) Confusion Matrix (CHASE) (c) Comparison of Category Accuracy

Figure 9: **Confusion Matrices & Comparison of Category-level Accuracy on NTU Mutual 26 X-Sub**. (a) Confusion matrix of the vanilla CTR-GCN. (b) Confusion matrix of CTR-GCN with CHASE. (c) We present a detailed comparison of the category-level accuracy between the vanilla CTR-GCN (blue) and CTR-GCN with CHASE (orange). These results demonstrate the effectiveness of CHASE by improving recognition accuracy for most categories.

This suggests that CHASE is effective even in mixed recognition settings. However, the improvement is modest because single actions are the dominant category in this dataset.

**Analysis on Efficiency**. We analyse CHASE's efficiency on NTU Mutual 26 configurations. Our proposed CHASE is implemented as a backbone wrapper, adaptable to a variety of single-entity action models. As presented in Table 8, the number of trainable parameters is about 26.37 k, which only increases number of backbone's parameters by 1%-2%. We can approximate that the number of trainable parameters is increased by $(U + 1 + C_2) \times C_1 + C_2 \times U$. For computational complexity, the FLOPs of CHASE is approximately 2.50 M. This analysis proves that CHASE is both efficient and lightweight for benefiting multi-entity action learning.

Table 8: CHASE Trainable Parameters

| Method | # Param. (M) |
|---|---|
| CTR-GCN [36] | 1.44 |
| **+ CHASE** | $1.46^{\uparrow 1.83\%}$ |
| InfoGCN [37] | 1.54 |
| **+ CHASE** | $1.57^{\uparrow 1.96\%}$ |
| STSA-Net [40] | 4.13 |
| **+ CHASE** | $4.16^{\uparrow 0.60\%}$ |
| HD-GCN [38] | 1.65 |
| **+ CHASE** | $1.68^{\uparrow 1.60\%}$ |

**Segment Size of Squeeze Operator.** In CLB, which is also the mapping $\psi$ to weight matrix, the squeeze operator squeezes the tensor to a specific shape, denoted as $(T', J', E')$. This segment size determines the point set in a multi-entity action sequence to which ICHAS applies. For example, the segment size $(1, 1, 1)$ indicates that ICHAS applies to all the points, and $(1, 1, 2)$ means two different ICHAS apply to two entities separately. Table 9 evaluates various segment sizes of the squeeze operator, demonstrating that the global ICHAS with the segment size $(1, 1, 1)$ achieves the best performance compared with the other settings. Therefore, we choose the default segment size as $(1, 1, 1)$ in all the experiments. Besides, the reduction ratio between $C_1$ and $C_2$ is determined according to the experimental results in [98].

Table 9: Ablation on Segment Sizes

| $(T', J', E')$ | Acc (%) |
|---|---|
| **(1, 1, 1)** | **91.30**$_{(\pm 0.22)}$ |
| (1, 1, 2) | 91.28$_{(\pm 0.19)}$ |
| (2, 1, 1) | 91.20$_{(\pm 0.04)}$ |
| (4, 1, 1) | 91.03$_{(\pm 0.09)}$ |
| (1, 5, 1) | 91.23$_{(\pm 0.05)}$ |

**Weight $\lambda$ for MPMMD.** Fig. 10 evaluates different values of the trade-off weight factor $\lambda$ in Eq. 12, varying from $10^{-2}$ to $10^{0}$. On NTU Mutual 26, it achieve the best performances when adopting $\lambda = 0.1$ for MPMMD loss. It corresponds to our claim that MPMMD is an auxiliary objective to guide discrepancy minimization, additional to the recognition task objective. We also conduct experiments with MPMMD across a variety of $M$ values but find insignificant differences in performance. Hence, we set $M = 1$ for computational efficiency.

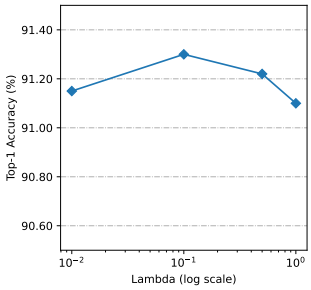



Figure 10: **Ablation study on different $\lambda$s for MPMMD.**

**Analysis on Confusion Matrix & Category-level Accuracy.** We present the confusion matrices of the vanilla CTR-GCN and CTR-GCN with CHASE in Fig. 9 (a) & (b). It indicates that our proposed CHASE is able to assist the backbone to differentiate similar multi-entity actions better. Fig. 9 (c) reports the category-level accuracy for 26 kinds of person-to-person interactions. We observe that in very few categories, their performance slightly drops. One possible

Table 10: Analysis of Performances with Test-Time Skeleton Noises and Masking

| Test-Time | + Noise (%) | | + Mask (%) | |
|---|---|---|---|---|
| | $\sigma = 10^{-3}$ | $\sigma = 10^{-2}$ | $p_m = 10^{-2}$ | $p_m = 10^{-1}$ |
| CTR-GCN [36] | $88.55_{(\pm 0.03)}$ | $80.72_{(\pm 0.03)}$ | $81.15_{(\pm 0.13)}$ | $56.37_{(\pm 0.13)}$ |
| **+ CHASE** | $\mathbf{91.24}_{(\pm 0.01)}$ | $\mathbf{82.53}_{(\pm 0.08)}$ | $\mathbf{88.57}_{(\pm 0.03)}$ | $\mathbf{60.65}_{(\pm 0.07)}$ |

reason is that these actions, like *point finger*, rely heavily on cues from the individual movements of one of the entities, instead of the multi-entity interactions. This might not be addressed by mitigating inter-entity distribution discrepancies. But it could still conclude from Fig. 9 (c) that adopting CHASE can achieve accuracy improvement for most of the categories, e.g. *wield knife* (+8.50%), *shoot with gun* (+8.00%), *whisper* (+5.39%), and *punch/slap* (+4.37%).

**Analysis of Performances with Test-Time Skeleton Noises and Masking.** To evaluate the robustness, we intentionally corrupt the multi-entity skeleton sequences with noise and masking during the inference phase. This aims at resembling possible skeleton occlusions or estimation errors during the test time. The noises $X_n \sim \mathcal{N}(\mu, \sigma^2)$ used in this experiment are normally distributed with mean $\mu = 0$ and standard deviations $\sigma = 10^{-3}, 10^{-2}$. Masking strategies are randomly masking the multi-entity skeleton sequences with probabilities $p_m = 10^{-2}, 10^{-1}$. Table 10 reports the averaged top-1 accuracy and its standard deviation in runs with several seed initializations for noises and masks. For recognizing multi-entity actions with corrupted test-time inputs, CTR-GCN integrating with CHASE consistently outperforms the vanilla counterpart, showcasing its robustness.

## G Limitations & Broader Impacts

This work proposes the Convex Hull Adaptive Shift for Multi-Entity Action Recognition to resolve the inter-entity distribution discrepancies. Although CHASE offers a generic framework for various types of multi-entity actions, such as person-to-person interactions, hand-to-object interactions, hand-to-hand interactions and group activities, its application to single-entity actions warrants further investigation. One potential approach is to consider different parts as multiple entities, such as treating different limbs of a human body as distinct entities. Moreover, it's promising to apply CHASE-like designs to the recently-developed human-centric foundation models [17, 29, 30, 31, 32, 33, 34] to enhance their performances on multi-entity skeletal data. These areas of exploration are left for future research.

This paper focuses on multi-entity action recognition, a field with broad applications in physical human-robot interaction, social scene understanding, multi-agent systems, surveillance, healthcare monitoring, etc [24, 22, 25, 26, 27, 28, 23]. Our work contributes to advancements in these domains by enhancing efficient and effective skeleton-based learning of multi-entity actions. Although the use of human-centered data can pose privacy concerns, our study utilizes only skeletal data estimated from sensors or RGB videos, thus mitigating potential privacy and ethical issues.

## H Licenses for Used Assets

Datasets:

- NTU Mutual 11 / NTU RGB+D [41]): Custom (research-only, non-commercial, attribution) [2]
- NTU Mutual 26 / NTU RGB+D 120 [42]: Custom (research-only) [3]
- H2O [13]: Custom (research-only, non-commercial) [4]
- Assembly101 [12]: Creative Commons Attribution-NonCommercial 4.0 International License [5]
- Collective Activity Dataset [85]: Unknown

[2] `http://rose1.ntu.edu.sg/Datasets/requesterAdd.asp?DS=3`

[3] `http://rose1.ntu.edu.sg/Datasets/actionRecognition.asp`

[4] `https://h2odataset.ethz.ch/`

[5] `http://creativecommons.org/licenses/by-nc/4.0/`

- Volleyball Dataset [86]: BSD 2-Clause license [6]

Models:

- CTR-GCN [36]: Creative Commons Attribution-NonCommercial 4.0 International License [7]
- InfoGCN [37]: Unknown
- STSANet [40]: MIT License [8]
- HD-GCN [38]: MIT License [9]

[6] https://github.com/mostafa-saad/deep-activity-rec/blob/master/LICENSE
[7] https://github.com/Uason-Chen/CTR-GCN/blob/main/LICENSE
[8] https://github.com/heleiqiu/STTFormer/blob/main/LICENSE
[9] https://github.com/Jho-Yonsei/HD-GCN/blob/main/LICENSE

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect this paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our work, provided in Appendix G.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide a proof of our proposition 1 in Section 3.1.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We describe our proposed CHASE clearly and fully in Section 3.1, Section 3.2 and Section 3.3. We illustrate our experimental setting and details in Section 4.1 and Appendix E. Our code is publicly available at `https://github.com/Necolizer/CHASE`.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Our code is publicly available at `https://github.com/Necolizer/CHASE`. We have provided sufficient instructions to reproduce the main experimental results.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We illustrate our experimental setting and details in Section 4.1 and Appendix E. We also provide all the configuration files in our code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Table 1 shows the experimental results on different benchmarks, reporting the averaged top-1 accuracy and its standard deviation in runs with several seed initializations. In Table 3, we sample the points 30 times with different seed initializations and report the averaged measurements with standard deviations.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provided sufficient information on the computer resources in Appendix E. Experiments are conducted with 8 GeForce RTX 3070 GPUs (GPU Memory: 8GB), using torch version 1.9.0+cu111, torchvision version 0.10.0+cu111, and CUDA version 11.4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This research conforms in every respect with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss both potential positive societal impacts and negative societal impacts of our work, provided in Appendix G.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We carefully cited the original papers for existing assets and included their licenses in Appendix H.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our code is publicly available at `https://github.com/Necolizer/CHASE` with MIT license.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.