Expert Systems With Applications 290 (2025) 128344

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Geometric visual fusion graph neural networks for multi-person human-object interaction recognition in videos

Tanqiu Qiao [a], Ruochen Li [a], Frederick W.B. Li [a], Yoshiki Kubotani [b], Shigeo Morishima [c], Hubert P. H. Shum [a,*]

[a] Durham University, Department of Computer Science, Durham, DH1 3LE, UK
[b] cvpaper.challenge, Tokyo, Japan
[c] Waseda University, Waseda Research Institute for Science and Engineering, Tokyo, 169-8555, Japan

ARTICLE INFO

Keywords:
Human-object interaction
Multi-person interaction
Feature fusion

ABSTRACT

Human-Object Interaction (HOI) recognition in videos requires understanding both visual patterns and geometric relationships as they evolve over time. Visual and geometric features offer complementary strengths. Visual features capture appearance context, while geometric features provide structural patterns. Effectively fusing these multimodal features without compromising their unique characteristics remains challenging. We observe that establishing robust, entity-specific representations before modeling interactions helps preserve the strengths of each modality. Therefore, we hypothesize that a bottom-up approach is crucial for effective multimodal fusion. Following this insight, we propose the Geometric Visual Fusion Graph Neural Network (GeoVis-GNN), which uses dual-attention feature fusion combined with interdependent entity graph learning. It progressively builds from entity-specific representations toward high-level interaction understanding. To advance HOI recognition to real-world scenarios, we introduce the Concurrent Partial Interaction Dataset (MPHOI-120). It captures dynamic multi-person interactions involving concurrent actions and partial engagement. This dataset helps address challenges like complex human-object dynamics and mutual occlusions. Extensive experiments demonstrate the effectiveness of our method across various HOI scenarios. These scenarios include two-person interactions, single-person activities, bimanual manipulations, and complex concurrent partial interactions. Our method achieves state-of-the-art performance.

1. Introduction

Human-Object Interaction (HOI) recognition aims to interpret the intricate relationships between humans and the objects they interact with. While traditional video analysis tasks can achieve strong performance using visual features alone, HOI recognition demands additional geometric reasoning through human poses and object spatial configurations. In video-based scenarios, this complexity intensifies as systems track dynamic spatial relationships across frames while handling occlusions and viewpoint changes. This complexity goes beyond the pixel-level understanding required for coarse actions like cooking. Instead, it includes geometric analysis of fine-grained interactions, such as specific hand poses for holding objects or spatial configurations needed for cutting. These interactions often occur concurrently or in sequence.

Significant efforts have focused on image-based HOI detection, which combines object localization and interaction classification within static frames. Recent advances have leveraged transformer architectures with specialized mechanisms (Kim, Jung, & Cho, 2023; Li, Wei, Wang, & Yang, 2024; Ma, Wang, Wang, & Wei, 2023; Zhu, Ho, Chen, Yang, & Shum, 2024). While these methods are effective for static scenes, they are inadequate for capturing the temporal dynamics and motion complexities inherent in video scenarios.

Video-based HOI recognition is a relatively less-explored area, which requires understanding not only the spatial relationships between humans and objects but also how these interactions evolve over time. Existing methods primarily rely on visual features (Morais, Le, Venkatesh, & Tran, 2021; Tu, Sun, Min, Zhai, & Shen, 2022; Wang et al., 2023), which encode rich appearance and contextual cues but are vulnerable to occlusions common in real-world scenarios. In contrast, geometric features, derived from human pose estimations and object spatial configurations, provide explicit structural details crucial for interaction understanding (Das, Sharma, Dai, Bremond, & Thonnat, 2020; Wan, Zhou, Liu, Li, & He, 2019; Zhu et al., 2024). While some recent approaches (Qiao, Li, Li, & Shum, 2024; Qiao et al., 2022) have attempted to integrate

* Corresponding author at: Department of Computer Science, Durham University, Durham, DH1 3LE, United Kingdom.
E-mail addresses: tanqiu.qiao@durham.ac.uk (T. Qiao), ruochen.li@durham.ac.uk (R. Li), frederick.li@durham.ac.uk (F.W.B. Li), yoshikikubotani.lab@gmail.com (Y. Kubotani), shigeo@waseda.jp (S. Morishima), hubert.shum@durham.ac.uk (H.P.H. Shum).

https://doi.org/10.1016/j.eswa.2025.128344
Received 6 December 2024; Received in revised form 2 May 2025; Accepted 24 May 2025
Available online 3 June 2025

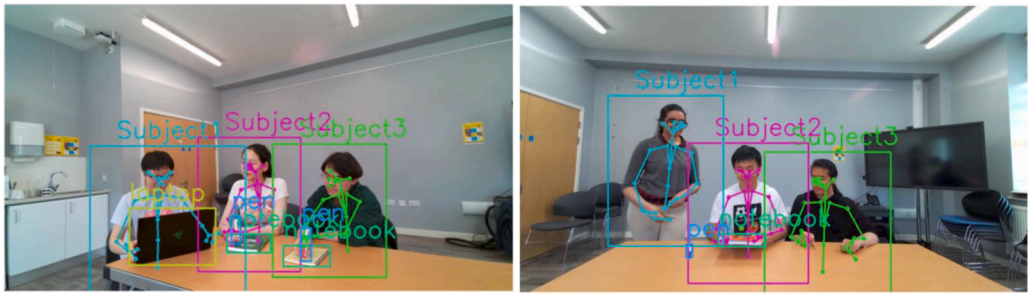

Fig. 1. Two examples (*Teaching* and *Signing*) of our collected concurrent partial interaction datasets. Geometric features such as skeletons and bounding boxes are annotated.

visual and geometric features to leverage their complementary advantages, these methods typically fuse both modalities directly. This simplistic fusion neglects the distinct and valuable characteristics inherent to each modality, resulting in suboptimal interaction representations. Consequently, effectively integrating visual and geometric features remains challenging, limiting the capability to fully exploit their complementary strengths for robust interaction recognition.

Another challenge is that current approaches often fail to preserve fine-grained entity characteristics when integrating multimodal features, resulting in the loss of important interaction details. Effectively fusing geometric and visual features to fully leverage their potential for enhancing entity interaction recognition remains underexplored. A mixed-fusion approach (Qiao et al., 2022) that combines feature fusion and entity interaction learning in a unified graph suffers from entangled entity representations, limiting explicit HOI understanding. A top-down approach (Zheng et al., 2023) that prioritizes entity-level relationships over specific features may miss critical interaction details and misalign overarching patterns with individual nuances. An alternative bottom-up approach (Wang, Zhou, Chen, Tang, & Wang, 2022a; Wang, Li, Cai, Chen, & Han, 2022b) starts with fundamental entity features, enabling detailed integration before addressing complex interactions, resulting in more effective entity interaction analysis. Although the bottom-up approach has potential benefits, it faces challenges in accurately fusing multimodal features. Specifically, it struggles to preserve fine-grained entity characteristics when transitioning from individual entity representations to modeling complex interactions.

In this paper, we introduce a novel Geometric Visual Fusion Graph Neural Network (GeoVis-GNN) to address the two critical challenges identified in previous research. To overcome the inadequate fusion of geometric and visual features, we propose a dual-attention mechanism operating at the feature level. This mechanism first utilizes graph attention to produce structured geometric embeddings and then employs channel attention to adaptively fuse these embeddings with visual features, effectively leveraging the complementary strengths of both modalities. To preserve fine-grained entity characteristics, we adopt a bottom-up approach. Specifically, we first establish robust entity-specific representations and then progressively build up to higher-level interaction understanding through an interdependent entity graph. This graph models explicit spatial interactions and implicit contextual dependencies among entities, ensuring that detailed entity characteristics are retained even when modeling complex interactions.

Video-based HOI datasets have primarily focused on single-person activities or limited two-person interactions, leaving a notable gap in capturing the complexity of real-world multi-person scenarios. Existing datasets, such as UCLA HHOI (Shu, Gao, Ryoo, & Zhu, 2017; Shu, Ryoo, & Zhu, 2016) and MPHOI-72 (Qiao et al., 2022), include interactions with up to two participants and a few objects. They assume all participants are continuously active, which limits their ability to represent scenarios with idle or waiting individuals. To address this limitation, we introduce the Concurrent Partial Interaction Dataset (MPHOI-120), which captures dynamic multi-person interactions where some participants are engaged while others are idle Fig. 1. This dataset incorporates diverse interactions, high variability, and challenging dynamics such as simultaneous actions, mutual dependencies, and occlusions. It offers a richer and more realistic benchmark for advancing HOI recognition in complex real-world scenarios.

We demonstrate the effectiveness of our approach across a comprehensive spectrum of real-world HOI scenarios. These include single-person interactions in CAD-120 (Koppula, Gupta, & Saxena, 2013), bimanual manipulations in Bimanual Actions (Dreher, Wächter, & Asfour, 2020), two-person collaborative activities in MPHOI-72 (Qiao et al., 2022), and concurrent partial interactions in our proposed MPHOI-120 dataset. These diverse datasets collectively represent the full range of human-object interactions, from individual tasks to simultaneous multi-person collaborations. Our main contributions are:

- A novel bottom-up framework (GeoVis-GNN) for multi-person HOI recognition. It preserves fine-grained entity characteristics and progressively builds from entity-level representations to interaction-level understanding.[1]
- A dual-attention fusion mechanism that first employs graph attention to learn structured geometric embeddings, followed by channel attention-based adaptive fusion with visual features, resulting in rich entity-specific representations.
- An interdependent entity graph that leverages the enriched entity representations to simultaneously model explicit spatial interactions and implicit contextual dependencies among multiple entities.
- A novel Concurrent Partial Interaction Dataset (MPHOI-120)[2] captures complex multi-person interactions with concurrent actions and partial engagement, providing a more realistic benchmark to advance HOI recognition.

2. Related work

2.1. HOI recognition

HOI recognition is divided into two primary areas: HOI detection in images and HOI recognition in videos. HOI detection in images focuses on identifying interactions within a single static picture, combining object localization with interaction classification. It aims to detect triplets $\langle human, verb, object \rangle$, providing a spatially grounded understanding of HOIs in a single image (Cheng, Duan, Wang, & Chen, 2024b; Kim et al., 2023; Li et al., 2024). These methods are not directly applicable to HOI recognition in videos, as the task introduces a temporal dimension, requiring models to capture interactions as they evolve over time. Video-based HOI recognition demands the ability to process dynamic, sequential data to understand interaction context more comprehensively. While

[1] Code is available in the Supplementary Materials.

[2] Data collection performed in the UK, under Durham University Ethics Approval Ref: COMP-2020-10-01T19_29_22-cbmw62.

some image-based methods provide valuable spatial insights using geometric and visual features (Park, Park, & Lee, 2023; Wu et al., 2022; Zhu et al., 2024), they lack the temporal modeling capabilities necessary to capture motion patterns, action progression, and continuity. This limitation results in an incomplete understanding of the evolving interactions critical for accurate recognition in video scenarios.

HOI recognition in videos encompasses human action analysis (Cob-Parro, Losada-Gutiérrez, Marrón-Romera, Gardel-Vicente, & Bravo-Muñoz, 2024; Hu, Xiao, Li, Liu, & Ji, 2024; Tan, Lim, Lee, & Kwek, 2022) and skeleton-based activity recognition (Cheng, Cheng, Liu, Ren, & Liu, 2024a; Setiawan, Yahya, Chun, & Lee, 2022; Yu, Tanaka, & Fujiwara, 2024) by integrating the detection of human movements and postures with the contextual understanding of interactions between humans and objects, thereby offering a more holistic approach to activity recognition in complex environments. Deep Neural Networks (DNNs) and graphical models are combined in recent works. Wang et al. (2021) utilize the parsed graphs to directly model the global relationship between the human and object, capturing the state change of the interacting objects across frames. ASSIGN (Morais et al., 2021) presents a visual feature attention model to learn asynchronous and sparse HOIs in videos. TUTOR (Tu et al., 2022) employs a reinforced tokenization strategy that jointly learns instance tokens through selective attention and aggregation in the spatial domain and links them across frames to generate tubelet tokens, serving as highly-abstracted spatio-temporal representations for HOI recognition. Xing and Burschka (2022) introduce a spatial attention mechanism that can enhance action recognition by adaptively generating a spatial-relation graph during HOIs. STIGPN (Wang et al., 2023) exploits spatio-temporal graph convolutions to enhance the detection of salient human-object interactions and efficiently modeling long-term dynamics.

Based on prior visual-based approaches, 2G-GCN (Qiao et al., 2022) firstly proposes the multi-person HOI recognition problem and incorporates geometric features into the Graph Convolutional Network (GCN). However, 2G-GCN merges the collective geometric features of all entities with individual visual features, leading to potential hierarchical misalignment. The high-level spatial information from geometric features may not align well with detailed, entity-specific visual data. As a result, the model may struggle to correctly distinguish between different entities and their interactions, leading to impaired performance and a focus on less relevant objects. CATS (Qiao et al., 2024) learns HOIs from multimodal feature fusion of different categories, such as humans and objects, to the scenery interactive graph. However, it neglects the entity concept and entity relationships within the same category, which is particularly limiting in multi-person HOI scenarios. Therefore, in this paper, we follow a bottom-up approach that first establishes fine-grained entity-specific features before capturing entity-level interactions, ensuring precise entity representations and facilitating accurate interaction modeling in complex multi-entity scenarios.

2.2. Geo-vis fusion in human activities

Combining diverse data modalities offers unique, complementary insights that lead to a more holistic understanding of a subject. In multimodal research of human action recognition, attention has been directed towards key areas of the human body, particularly the hands (Baradel, Wolf, & Mille, 2017, 2018a; Baradel, Wolf, Mille, & Taylor, 2018b). These studies employ attention-based methods to improve the overall accuracy of models that integrate skeletal and visual modalities. Building on this, Bruce, Liu, and Chan (2021) expand the focus to include additional regions of the body such as the head, hands, and feet by adopting a temporal approach. They generate a fused representation by multiplying spatial attention weights with appearance features. TSMF (Bruce et al., 2021) fuses skeleton and RGB data at the model level using teacher-student networks to learn enriched representations. However, these model-based fusion models often lack transparency, making it difficult to interpret how individual features contribute to recognition. Besides, Boulahia, Amamra, Madi, and Daikh (2021) investigate the integration of various image modalities (RGB, Depth, Skeleton, and InfraRed) at different stages of the action recognition pipeline, encompassing early, intermediate, and late fusion techniques, to enhance the robustness of recognition.

In human interaction analysis, Wan et al. (2019) concatenate human skeletal embeddings with visual embeddings from other branches like human, object and union to obtain the final holistic feature in the HOI scene. Zhou et al. (2022) combine embedded visual and human pose features through element-wise addition. Wang et al. (2023) directly concatenate multimodalities to output visual-spatial and spatial-semantic feature sequences, which are then input into a two-stream network. CATS (Qiao et al., 2024) also concatenates geometric and visual features for different categories. However, their direct operations may dilute distinct contributions of visual and geometric features, often amplifying dominant features while undervaluing subtle geometric cues, which can reduce accuracy in fine-grained interaction recognition. These challenges indicate that the fusion of geometric and visual features still has design intricacies that require further optimization. Therefore, we propose a dual-attention fusion mechanism to integrate geometric and visual features. This mechanism combines graph attention and channel attention to preserve the complementary strengths of both modalities. As a result, it produces enriched entity representations, enabling more robust and precise interaction modeling.

2.3. Video-based HOI datasets

There are various datasets available for the investigation of HOI in videos for multiple tasks. For single-person HOI recognition, datasets like CAD-120 (Koppula et al., 2013), Bimanual Actions (Dreher et al., 2020), Bimanual Manipulation (Krebs, Meixner, Patzer, & Asfour, 2021), *etc.* are effective, with the latter two also encompassing bimanual HOI tasks due to their focus on interactions involving both hands. There are several datasets available for single-hand HOI recognition tasks, including Something-Else (Materzynska et al., 2020), VLOG (Fouhey, Kuo, Efros, & Malik, 2018), EPIC Kitchens (Damen et al., 2021). Since EPIC Kitchens records both hands in the cooking process, it can also be utilized for bimanual HOI recognition. Besides, a full-body HOI dataset called BEHAVE (Bhatnagar et al., 2022) includes multi-view RGBD frames, associated 3D SMPL and object fits. HOI4D (Liu et al., 2022) is a large-scale 4D egocentric dataset aimed to facilitate research on category-level HOIs. The UCLA HHOI Dataset (Shu et al., 2017, 2016) focuses on human-human-object interaction with a maximum of two humans and one object involved. The MPHOI-72 dataset (Qiao et al., 2022) is specifically proposed for the multi-person HOI recognition task but is constrained to interactions between two individuals and 2–4 objects, reducing its applicability to complex real-world scenarios.

3. Concurrent partial interaction dataset

The majority of video-based HOI datasets primarily focus on single-person HOIs, albeit from various perspectives (Bhatnagar et al., 2022; Damen et al., 2021; Koppula et al., 2013; Liu et al., 2022). Efforts to encompass multiple human interactions are still in their infancy. For instance, the UCLA HHOI dataset (Shu et al., 2017, 2016) captures interactions involving up to two people and one object, while MPHOI-72 (Qiao et al., 2022) slightly broadens this scope to include two people and several objects. However, these datasets assume that all participants are continuously active throughout the activity. In contrast, real-world multi-person HOIs often include scenarios where some individuals are not interacting, such as sitting or standing idle while waiting for their turn. This gap highlights the need for datasets that better represent the complexity and variability of real-world multi-person and multi-object interactions.

To bridge this critical gap, we introduce the Concurrent Partial Interaction Dataset (MPHOI-120), which captures dynamic interactions involving multiple people and objects. In our context, “concurrent” refers to scenarios where multiple interactions occur simultaneously, while “partial interaction" highlights moments when not all individuals are actively engaging - some may remain idle or waiting during certain moments of the activity. For example, in the *Signing* activity, while two people are passing a notebook and pen, the other person is standing or sitting idle. Similarly, when one individual is signing, the other two are not interacting. Such scenarios, which reflect real-world interaction patterns, are extensively captured in our dataset, providing a richer benchmark for advancing multi-person HOI recognition methods. In addition, increasing the number of people and objects introduces an exponential increase in complexity. It expands the range of human-human, human-object, and object-object interactions, while also intensifying challenges such as simultaneous actions, mutual dependencies, and significant occlusions.

3.1. Dataset details

MPHOI-120 is a dataset of 120 high-resolution videos of three participants interacting with 2 to 5 objects. All annotations are performed frame-by-frame by a single trained annotator using a predefined list of sub-activities to ensure consistency and avoid inter-annotator variability. Sample video screenshots with annotated sub-activities for all activities are shown in Fig. 2. Each main activity captures unique interaction patterns. *Signing* highlights turn-taking behaviors amid potential occlusions, while *Cheering* features synchronous and sequential human-object actions. *Teaching* depicts fine-grained states (e.g., noting vs. listening) between a teacher and students, and *Snooker* focuses on strategic turn-taking with frequent body occlusions around the table.

Leveraging the Azure Kinect SDK along with the Body Tracking SDK (Microsoft, 2022), we acquire RGB-D videos to capture the comprehensive dynamics of multiple individual skeletons. We offer 2D human skeletal data and bounding boxes for both subjects and objects within each video, serving as geometric characteristics. The integration of depth information within our dataset further broadens its utility, such as versatile benchmarks for 3D human pose estimation (You et al., 2023; Zhai, Nie, Ouyang, Li, & Yang, 2023) and 3D object estimation (Fan, Chen, Hu, & Zhou, 2023; Heitzinger & Kampel, 2023), among others.

Table 1
A statistical comparison between MPHOI-120 and popular HOI datasets. CPI denotes Concurrent Partial Interactions.

Datasets	MPHOI-120	MPHOI-72	CAD-120	Bimanual Actions
No. people interacting	3	2	1	1
Total videos	120	72	120	540
Total frames	53,604	26,383	61,585	221,000
Total frames of CPI	20,100	0	0	0
Video average length	15s	12s	17s	15s
No. sub-activities	17	13	10	14
No. subjects/objects	7/6	5/6	4/10	6/12
Total activities	4	3	10	9
Fps	30	30	30	30
Resolution	1920×1080	3840×2160	640×480	640×480

3.2. Statistical comparison of datasets

We perform a statistical comparison between MPHOI-120 and existing popular HOI datasets, as shown in Table 1. MPHOI-120 includes scenarios with three people interacting and 17 sub-activities, which is higher than any other listed dataset, standing out for its complexity and richness. With a total of 53,604 frames across 120 videos, nearly half (20,100 frames) capture concurrent partial interactions, offering a unique focus on dynamic multi-person interactions absent in other datasets. Additionally, the high video resolution (1920×1080) ensures detailed feature capture, essential for advanced HOI analysis. In contrast, although Bimanual Actions is large, it is limited to dual-hand movements of an individual, leading to a more monotonic data distribution.

4. Methodology

We propose a bottom-up approach to design GeoVis-GNN, which (1) preserves fine-grained entity characteristics during feature fusion, and (2) progressively builds from entity-specific representations to interaction-level understanding. The bottom-up approach has been widely used in pose estimation (Kresović & Nguyen, 2021; Wang et al., 2022a,b) and object detection (Samet, Hicsonmez, & Akbas, 2020; Wang, Shen, Cheng, & Shao, 2019; Zhou, Zhuo, & Krahenbuhl, 2019) tasks with considerable performances. It ensures a thorough

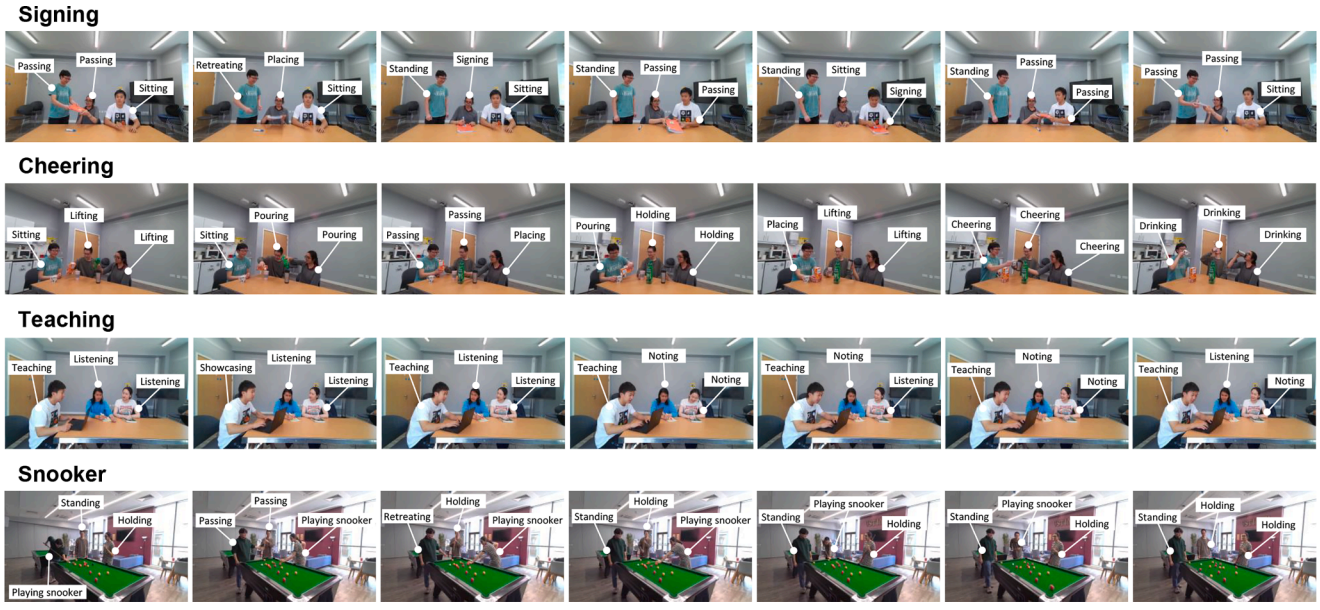

Fig. 2. Sample video screenshots from our new MPHOI-120 dataset, displaying concurrent partial interactions along the timelines of four multi-person HOI activities in daily life.

understanding of the fundamental aspects of each entity before delving into complex entity-level interactions. This approach, starting from basic features and building upwards, enables detailed feature integration to achieve more effective entity interaction analysis.

Alternative designs perform suboptimally. A top-down approach (Zheng et al., 2023), which prioritizes a broad view of entity-level relationships before refining specific entity features, often overlooks crucial interaction details and misaligns overarching patterns with individual interaction nuances. Besides, a mixed-fusion method (Qiao et al., 2022) that integrates feature fusion and entity interaction learning within a single graph entangles entity concepts, lacking a specific feature to represent each entity, which fails to learn HOIs explicitly. We compare these alternative architectures with our method in Experimental Results 5.

4.1. Dual-attention fusion for feature optimization

Previous HOI recognition approaches primarily rely on CNN or 3D-CNN models to process visual inputs. These models extract spatio-temporal features that capture rich appearance and contextual cues from humans and objects (Le, Sahoo, Chen, & Hoi, 2020; Maraghi & Faez, 2019; Morais et al., 2021). While effective in clean settings, these methods are highly sensitive to occlusions and struggle when visual cues are incomplete or ambiguous. Without explicit spatial reasoning, they often fail to capture the structural context of interactions. Incorporating geometric information is therefore critical for improving robustness and enabling accurate recognition in real-world HOI scenarios. Advanced methods such as 2G-GCN (Qiao et al., 2022) attempt to integrate geometric features within a GCN framework to augment visual data. However, their fusion of collective geometric features with individual visual features risks hierarchical misalignment, fusion inefficiencies, and difficulties in entity distinction. CATS (Qiao et al., 2024) also employs GCN to model geometric features but directly combines them with visual features, which may dilute their distinct contributions.

We propose a dual-attention fusion mechanism to optimize multi-modal feature integration for entity representations (Fig. 3). We first apply a graph attention mechanism to geometric features, enabling the model to learn structured spatial representations by capturing the varying importance of neighboring entities. With these enriched geometric embeddings, we then employ a channel attention module to adaptively fuse geometric and visual features, selectively emphasizing informative channels while suppressing less relevant ones. This sequence ensures that spatial reasoning is established before feature fusion and allows the model to balance modality contributions more effectively. If channel attention is applied before relational modeling, it would risk fusing less informative geometric features and weaken the spatial reasoning capability. As a result, we obtain a well-contextualized entity representation that effectively blends geometric and visual cues, providing a robust foundation for subsequent entity interaction graph learning.

4.1.1. Graph attention-based feature embedding

Previous research (Qiao et al., 2024, 2022; Zhou et al., 2022) learns geometric features using GCNs, which typically apply the same convolution operation to all neighbors of a node. This approach fails to account for the different roles or importance that neighbors may have in the context of multi-person HOIs. This may lead to a homogenization of features that fails to capture the complexity of multi-entity dynamics.

We propose a Graph Attention Network (GAT) (Brody, Alon, & Yahav, 2021) based embedding to capture the evolving significance of interactions. It learns multi-entity geometric features, adaptively weighting the importance of each entity's geometric features through an attention mechanism. This enables the model to expertly handle occlusions and dynamic environments for multi-person HOI recognition.

For feature representation, we concatenate the position and velocity of all entities into keypoint channels, forming geometric features $\mathcal{G} = \{g_t^{e,k}\}_{t=1,e=1,k=1}^{T,E,K} \in \mathbb{R}^4$ with $g_t^{e,k}$ as the k-th type features for entity e at frame t, where T denotes the total number of frames in the video, E and K denote the total number of entities and keypoints of an entity in a frame, respectively. Human joints and object bounding box diagonals are extracted as keypoints.

We adaptively infer spatial correlations with our GAT among keypoints k_1 and k_2 for a single timestep among entities as follows:

$$\mathbf{g}_t^s = \alpha_{k_1,k_1}\mathbf{\Theta}\mathbf{g}_{t,k_1} + \sum_{k_2\in\mathcal{K}} \alpha_{k_1,k_2}\mathbf{\Theta}\mathbf{g}_{t,k_2}, \tag{1}$$

and the attention coefficients α_{k_1,k_2} are computed as:

$$\alpha_{k_1,k_2} = \frac{\exp\left(\Gamma\left(\mathbf{a}^\top[\mathbf{\Theta}\mathbf{g}_{k_1} \,\|\, \mathbf{\Theta}\mathbf{g}_{k_2}]\right)\right)}{\sum_{k_3\in\mathcal{K}\cup\{k_3\}} \exp\left(\Gamma\left(\mathbf{a}^\top[\mathbf{\Theta}\mathbf{g}_{k_1} \,\|\, \mathbf{\Theta}\mathbf{g}_{k_3}]\right)\right)}, \tag{2}$$

where $\mathbf{\Theta}$ and Γ are the transformation function and LeakyReLU activation, respectively.

To efficiently integrate spatial and temporal information, we further process the attention-enhanced geometric features. In particular, $\mathbf{g}_t^s$ is then fused with a 1×1 convolution along the temporal channel to form spatial-temporal geometric features $\mathbf{g}_t^{st} \in \mathbb{R}^{T\times N K\times C_1}$, effectively summarizing temporal dynamics while avoiding the complexities of 3D convolutions. It is then reshaped to $\mathbf{g}_t^{st} \in \mathbb{R}^{T\times N\times KC_1}$ and embedded by a Multi-Layer Perceptron (MLP) to get entity geometric features $\mathbf{g}_t' \in \mathbb{R}^{T\times N\times C_2}$.

Unlike geometric features, visual features in videos contain rich contextual information and fundamental feature representations. Following

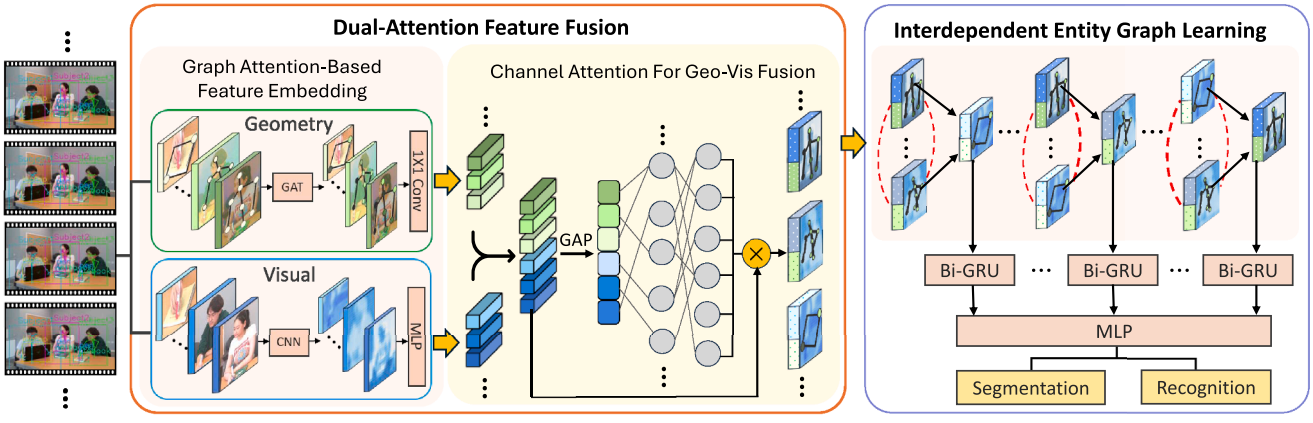

Fig. 3. Overview of our bottom-up framework GeoVis-GNN. We first design a dual-attention fusion for entity feature optimization, which embeds and fuses visual and geometric features in a graph attention-based mechanism and channel attention module, respectively. The enriched entity-specific representations are then inputted into the interdependent entity graph to further model explicit interactions and implicit interdependencies. Finally, we apply a BiGRU to capture the temporal dependencies to obtain segmentation and recognition results.

(Morais et al., 2021; Qiao et al., 2024), we extract entity visual features $\mathbf{v}_{t,n} \in \mathbb{R}^{2048}$ from ROI pooled 2D bounding boxes of humans and objects in videos, utilizing a pre-trained Faster R-CNN (Ren, He, Girshick, & Sun, 2016) module on the Visual Genome (Krishna et al., 2017). They are subsequently aligned dimensionally with geometric features to $\mathbf{v}'_t \in \mathbb{R}^{T\times N\times C_2}$ through an MLP with learnable embeddings.

4.1.2. Geo-vis channel attention-based feature fusion

Incorporating geometric and visual features poses a significant challenge due to their inherent representation and scale discrepancies. Prior approaches have attempted multimodal fusion by element-wise addition (Zhou et al., 2022) or feature concatenation (Qiao et al., 2024; Wan et al., 2019). However, such direct operations are infeasible for our task as they do not account for the disparate nature of feature spaces, leading to suboptimal learning outcomes.

We propose a novel geometry-visual channel attention-based feature fusion to effectively integrate geometric and visual features of all humans and objects, which achieves selective feature enhancement and encourages complementarity between multimodal features. We exploit channel attention mechanisms (Hu, Shen, & Sun, 2018) in geometry-visual channels of all entities. This allows the model to adaptively emphasize informative features while suppressing less relevant ones, which is especially beneficial for learning more representative visual and geometric features in diverse HOI scenarios. For instance, visual features often suffer in noisy backgrounds but thrive in scenarios with small backgrounds. Geometric features demonstrate strength in addressing partial occlusions (Qiao et al., 2022), which is a common situation in multi-person HOI scenarios.

Specifically, as shown in Fig. 3, our channel attention based feature fusion module first concatenates $\mathbf{g}'_t$ and $\mathbf{v}'_t$ along the entity dimension to entity geometry-visual features $\mathbf{gv}_t \in \mathbb{R}^{T\times 2N\times C_2}$, and compute a channel attention A as:

$$A = \sigma\big(\mathbf{W}_2\delta\big(\mathbf{W}_1\big(GAP\big(\mathbf{gv}_t\big)\big)\big)\big), \tag{3}$$

where GAP denotes Global Average Pooling (Lin, Chen, & Yan, 2013), δ and σ represent the ReLU and Sigmoid activation. $\mathbf{W}_1$ and $\mathbf{W}_2$ are weights of Fully-Connected (FC) layers, shared across all entities and timesteps to ensure consistent transformation and improved generalization with fewer parameters. Apply these values to original features for attended geometry-visual fusion features:

$$\mathbf{gv}'_t = A \cdot \mathbf{gv}_t. \tag{4}$$

Finally, we enhance the feature representation of each entity. In particular, after assigning distinct weights to each geometry-visual channel of an entity, the weighted features are strategically split into separate geometric and visual streams. These are then adeptly fused back together, producing a new enriched entity representation $\widetilde{\mathbf{gv}}_t \in \mathbb{R}^{T\times N\times C_3}$. This refined feature fusion set, being a weighted and well-contextualized blend of geometric and visual cues, sets the stage for more discerning entity interaction graph learning.

Compared to our attention-based feature fusion, Zhang et al. (2022), Tu, Sun, Zhai, and Shen (2023) apply Transformer to fuse geometric and visual features in image-based HOI detection, which is constrained in processing video data due to memory inefficiency. Graph-based feature fusion treats multimodal features as graph nodes (Gao et al., 2020; Liang et al., 2020), which is heavily reliant on the design of graph representation. As HOI is a dynamic process, it is non-trivial to manually define an appropriate representation.

4.2. Interdependent entity graph

In HOI analysis, the majority of approaches (Morais et al., 2021; Wang et al., 2023, 2021) construct an independent entity graph that assumes a fixed structure to decipher spatial interactions between entities focusing solely on visual features. For example, 2G-GCN (Qiao et al., 2022) represents geometric features of all entities as a single entity linked with visual features of object entities, failing to explicitly model interactions between all entities. CATS (Qiao et al., 2024) learns interactions between human and object categories but neglects relationships between entities within the same category, which is particularly limiting in multi-person HOI scenarios.

Our insight is that an effective entity interaction graph should not only capture explicit interactions among independent entities, but also concurrently discern the implicit interdependencies that exist among neighboring entities surrounding a specific entity. These complementary focuses are crucial for understanding the intricate graph network of relations that exist around any specific entity within the scene.

To this end, we propose an interdependent entity graph to capture the interdependencies among all neighboring nodes around a particular entity with fused geometric and visual features. To improve the precision of interaction modeling and the representation of relational dynamics, we further refine it by employing attention weights between the entity in focus and its neighbors (Fig. 3 right). This entity-level graph offers a richer representation of spatial interactions in multi-person HOI scenarios, advancing the understanding of complex behavioral patterns beyond the reach of previous methods.

Specifically, as illustrated in Fig. 4, given a specific entity e at each frame t, we first calculate the features from its neighbor u to itself as follows:

$$S_t^u = \lambda \times \widetilde{\mathbf{gv}}_t^u + (1-\lambda) \times \frac{(\mathrm{GAP}(\mathbf{W}_3(\widetilde{\mathbf{gv}}_t^u)))}{N-1}, \tag{5}$$

where λ controls the contextual fusion threshold and is fixed to 0.5. This value is selected based on preliminary validation experiments, and we find that the model's performance remains stable despite small variations in this setting. $\mathbf{W}_3$ is the weight of a FC layer. These neighboring features are then aggregated into a robust representation that encapsulates the collective attributes of the neighboring group:

$$S_t^e = STACK_{u\in N, u\neq e}(S_t^u \odot M(S_t^u)), \tag{6}$$

where $M(\cdot)$ is the mask indicator for valid neighbors and $\odot$ denotes element-wise multiplication. Meanwhile, we employ a dot-product attention mechanism (Morais et al., 2021; Vaswani et al., 2017) to obtain the attention weights between node e and its neighbors as:

$$W_t^e = \sum_{u\in N, u\neq e} Softmax\left(\frac{S_t^e (S_t^u)^T}{\sqrt{d}}\right), \tag{7}$$

where d is the feature dimension. Finally, the refined feature representation of the entity is $F_t^e = W_t^e \odot S_t^e$, ensuring a contextually aware integration of features that strengthens the entity's representation within its surroundings.

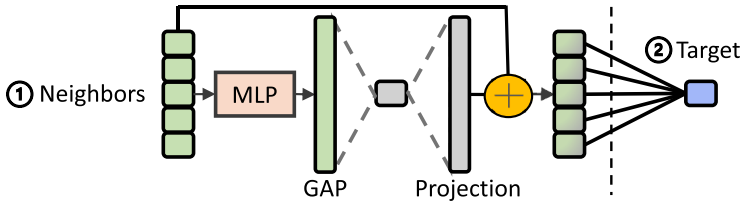

Fig. 4. In the interdependent entity graph, we first model neighbor features ① before aggregating them to the target entity ②.

To enable precise and adaptable delineation of sub-event lengths in video sequences, after obtaining the fused features of each entity at each time step, we employ a Gumbel-Softmax module (Jang, Gu, & Poole, 2016) to F_t^e. It efficiently facilitates gradient-based learning and ensures probabilistically coherent segmentation, essential for handling the dynamic nature of video data. Finally, we apply a Bi-directional Gated Recurrent Unit (BiGRU) (Chung, Gulcehre, Cho, & Bengio, 2014) to capture the temporal dependencies between each sub-action and then use the output features to recognize sub-activities for humans and object affordances for objects, varying according to the dataset.

5. Experimental results

5.1. Datasets

We evaluate GeoVis-GNN on multiple datasets: MPHOI-120, MPHOI-72 (Qiao et al., 2022), CAD-120 (Koppula et al., 2013), and Bimanual Actions (Dreher et al., 2020), showcasing the superior results on concurrent partial HOI, two-person, single-person and two-hand HOI recognition.

The MPHOI-72 dataset is valuable for two-person HOI recognition tasks. It contains 72 videos of 8 pairs of people performing 3 distinct activities (*Cheering*, *Hair cutting* and *Co-working*) with 13 human sub-activities (*e.g.*, *Sit*, *Approach*, *Pour*). Each video showcases two participants interacting with 2–4 objects from 3 unique angles. Geometric features and human sub-activity labels are frame-wise annotated.

CAD-120 is a prominent dataset for single-person HOI recognition. It contains 120 RGB-D videos, capturing 10 distinct activities executed by 4 participants, each repeated three times. In each video, a participant interacts with 1–5 objects. The dataset provides frame-wise annotations for 10 human sub-activities (*e.g.*, *opening*, *cleaning*, *placing*) and 12 object affordances (*e.g.*, *openable*, *cleanable*, *placeable*).

The Bimanual Actions dataset is a large-scale collection of 540 RGB-D videos capturing HOIs using both hands. It documents the actions of 6 subjects who engage in 9 varied bimanual tasks, with each task performed 10 times. The dataset assigns 14 unique action labels to each hand, with frame-wise annotations for each entity within the videos.

5.2. Implementation details

We follow Morais et al. (2021) and Qiao et al. (2024, 2022) to evaluate GeoVis-GNN on two tasks: joint segmentation and label recognition, and label recognition given known segmentation. The first task involves segmenting the timeline of each entity and classifying segment labels in a video. The second task, an extension of the first, requires labeling pre-existing segments with known ground-truth segmentation. We utilize the F1@k metric (Lea, Flynn, Vidal, Reiter, & Hager, 2017) for evaluation, applying standard thresholds of $k = 10\,\%$, $25\,\%$, and $50\,\%$. This metric considers a predicted action segment correct if it achieves a minimum Intersection over Union (IoU) overlap of k with the ground truth. It is widely adopted in temporal segmentation research (Farha & Gall, 2019; Lea et al., 2017; Morais et al., 2021), particularly for its ability to handle short or partial actions commonly found in HOI scenarios by requiring a certain overlap for each segment. As a result, it offers a more fine-grained evaluation of segmentation quality, capturing both the correctness of segment boundaries and the overall alignment with the ground truth.

For dataset evaluation, we use different cross-validation protocols tailored to the characteristics of each dataset to ensure subjects in the training set do not appear in the test set. For the single-person HOI datasets, CAD-120 and Bimanual Actions, we use leave-one-subject-out cross-validation, treating each individual as a separate fold. For the two-person HOI dataset MPHOI-72, we employ leave-two-subjects-out to preserve the same principle while accounting for pairs of interacting subjects. This ensures a strict separation of subjects (or subject pairs) between training and testing. For MPHOI-120, our cross-validation scheme specifies three subjects not present in the training set as the test set.

Table 2
Joined segmentation and label recognition results on MPHOI-120.

Model	Sub-activity		
	F_1@10	F_1@25	F_1@50
ASSIGN	58.0 ± 8.5	53.7 ± 7.9	39.1 ± 7.4
2G-GCN	60.7 ± 6.5	55.3 ± 6.9	39.6 ± 6.5
CATS	62.8 ± 2.7	56.7 ± 4.2	42.8 ± 3.9
GeoVis-GNN	**65.1** ± 5.2	**59.8** ± 4.7	**46.6** ± 5.1

The GeoVis-GNN framework is implemented in PyTorch and trained in two stages using the AdamW optimizer. A batch size of 16 is used across all datasets. The learning rate is set to 0.0001 for both the MPHOI datasets and the CAD-120 and Bimanual Actions datasets. Training MPHOI-120, MPHOI-72, CAD-120 and Bimanual Actions on four Nvidia Titan RTX GPUs take 6, 4, 8 hours and 7 days respectively, while testing the entire set takes approximately 2, 2, 6 and 20 minutes respectively.

To capture increasingly complex features while keeping computational cost reasonable, we adopt an incremental increase in dimensionality. Specifically, we set $C_1 = 128$, $C_2 = 256$, and $C_3 = 512$ based on empirical experimentation to balance model capacity and efficiency. C_1 and C_2 serve as mid-level embeddings for spatio-temporal transformations, while C_3 enables deeper representations for modeling high-level interactions. As Bimanual Actions has a significantly more monotonic data distribution, we set $C_2 = 32$, $C_3 = 64$ and $[S_t^u = 0]$.

5.3. Quantitative and qualitative comparison with SOTAs

5.3.1. Concurrent partial HOIs

In the MPHOI-120 dataset, GeoVis-GNN beats ASSIGN (Morais et al., 2021), 2G-GCN (Qiao et al., 2022) and CATS (Qiao et al., 2024) by a considerable gap (Table 2). Especially under multi-person HOI conditions, ASSIGN drops below $60\,\%$ in F_1 metrics due to occlusions affecting visual features in HOI tasks. GeoVis-GNN shows an improvement of about $2\,\%$ to $4\,\%$ in $F_1@\{10, 25, 50\}$ over SOTA, demonstrating its ability to effectively handle concurrent partial interactions. This highlights the strength of its dual-attention fusion strategy and interdependent entity graph in capturing essential features and modeling stable interactions, even in the presence of unexpected occlusions and complex multi-person dynamics.

Fig. 5 illustrates the visualization results of GeoVis-GNN and CATS on MPHOI-120 comparing with Ground-truth for the *Signing* activity, where red dashed boxes highlight major segmentation errors. Although both GeoVis-GNN and CATS make errors compared to Ground-truth, GeoVis-GNN can contribute relatively plausible segmentation results in all three subjects. For example, in subject 3, CATS oversegments *sit* in the beginning and then completely misses *pass* and *lift* before *note*, while our GeoVis-GNN can accurately segment *sit* and *pass* but miss *lift*. This is likely due to the *lift* action of the subject being very fast and closely resembles the *note* action, leading our model to misclassify *lift* as *note*. Incorporating temporal attention mechanisms could potentially enhance performance in the short duration of the action and its overlapping features with subsequent actions.

5.3.2. Two-person HOIs

GeoVis-GNN achieves an impressive performance on the MPHOI-72 dataset (Table 3), with an $F_1@10$ score of $84.3\,\%$, significantly outstripping the $71.3\,\%$ scored by CATS (Qiao et al., 2024). Across all F_1 configurations, GeoVis-GNN exhibits substantial improvements of $13.0\,\%$, $10.8\,\%$, and $10.6\,\%$, respectively. The advanced technique for fusing geometric and visual features allows to capture more complex patterns in the data, while CATS and 2G-GCN cannot leverage it due to its inefficient fusion.

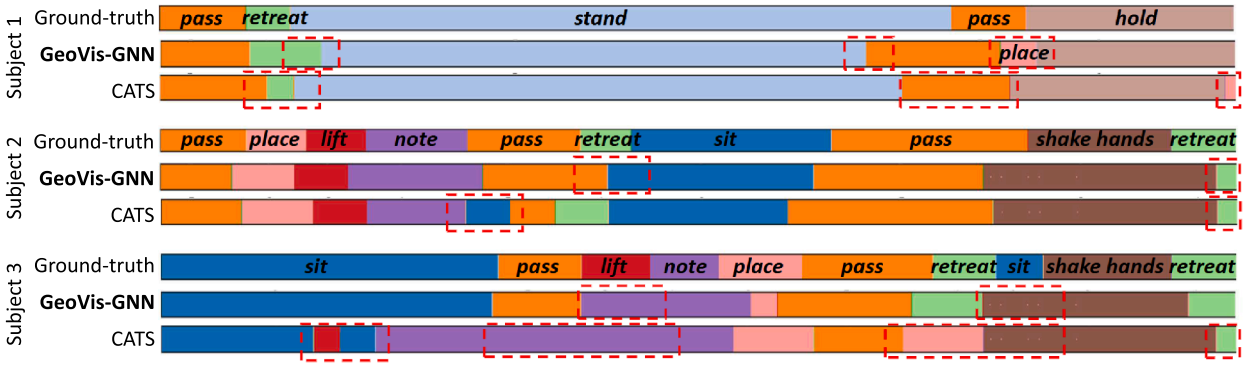

Fig. 5. Visualization of segmentation on MPHOI-120 for *Signing* activity. Red dashed boxes highlight major segmentation errors.

Table 3
Joined segmentation and label recognition results on MPHOI-72.

Model	Sub-activity		
	F_1@10	F_1@25	F_1@50
ASSIGN	59.1 ± 12.1	51.0 ± 16.7	33.2 ± 14.0
2G-GCN	68.6 ± 10.4	60.8 ± 10.3	45.2 ± 6.5
CATS	71.3 ± 5.0	65.8 ± 3.9	48.8 ± 5.3
GeoVis-GNN	**84.3** ± 5.5	**76.6** ± 4.5	**59.4** ± **4.9**

Fig. 6 shows the visualization of segmentation and labeling on the MPHOI-72 dataset with the two advanced models for the *Cheering* activity comparing with Ground-truth. GeoVis-GNN presents more reasonable and robust segmentation results in all sub-activities, while CATS provides some unexpected abnormal results in certain sub-activities, such as *pour* and *place*. Interestingly, CATS directly recognizes the static action *sit* rather than the ongoing action *retreat* following *place* at the end of the activity for subject 1. This may result from the dominant role of visual features, as these two actions appear similar in the front view.

5.3.3. Single-person HOIs

Table 4 shows the effectiveness of GeoVis-GNN in CAD-120 evaluated by sub-activity and object affordance labels. GeoVis-GNN beats previous visual-based (Koppula & Saxena, 2016; Morais et al., 2021; Sener & Saxena, 2015) and geometry-informed (Qiao et al., 2024, 2022) networks for both labels and achieves the highest F_1 scores of mean in every configuration. Notably, the two geometry-informed networks show comparable performance in human sub-activity recognition, but CATS performs poorly in object affordance recognition. This may be due to two main factors: an imbalance in feature representation, with fewer keypoints for objects than humans, reducing object emphasis in the scene graph, while the dual-attention feature fusion in GeoVis-GNN helps mitigate this. Additionally, our task requires both segmentation and label recognition, a two-stage process that does not align well with the end-to-end framework of CATS, which may struggle with such distinct processing stages. Although CATS performs well in multi-person HOI scenarios, empirical results indicate that it is less suited for single-person HOI tasks. Therefore, in the subsequent HOI recognition comparisons involving a single individual, we use 2G-GCN as the state-of-the-art benchmark.

Fig. 7 presents the visualization outcomes for the *Cleaning Objects* activity in CAD-120, depicting a scene where a person uses a cloth to clean a microwave. The qualitative analysis shows that GeoVis-GNN surpasses 2G-GCN in recognizing human sub-activities and object affordances, notably *reachable* and *movable* for the microwave, closely matching the Ground-truth.

5.3.4. Two-hand HOIs

GeoVis-GNN achieves the superior performance on the large-scale Bimanual Actions dataset (Table 5), with near 1 % improvement in the same standard deviation at F_1@10. The slight improvement is partly due to the limited hand pose estimation that OpenPose (Cao, Hidalgo, Simon, Wei, & Sheikh, 2018) uses for the hand skeleton of the dataset, which may introduce noise, especially in occlusions. Fig. 8 presents the visualization outcomes for the *Pouring* activity in Bimanual Actions. The qualitative analysis demonstrates that GeoVis-GNN has outstanding performance in segmenting and recognizing actions of both hands, which almost overlaps the Ground-truth, while 2G-GCN oversegments some sub-activities like *pour*.

5.4. Scenario-Based performance and error analysis

Table 6 summarizes GeoVis-GNN's performance across different HOI scenarios with performance gaps of MPHOI-120 at F_1@10. It arranges the datasets from single-person to multi-person partial, revealing a progressive increase in complexity. In single-person scenarios, CAD-120 shows the largest gap relative to MPHOI-120 at +24.8 %, indicating that single-participant tasks with clear sub-activity boundaries are relatively straightforward. Similarly, Bimanual Actions follows with a +20.7 % gap, reflecting simpler interactions than multi-person scenarios. GeoVis-GNN generally distinguishes between left- and right-hand movements

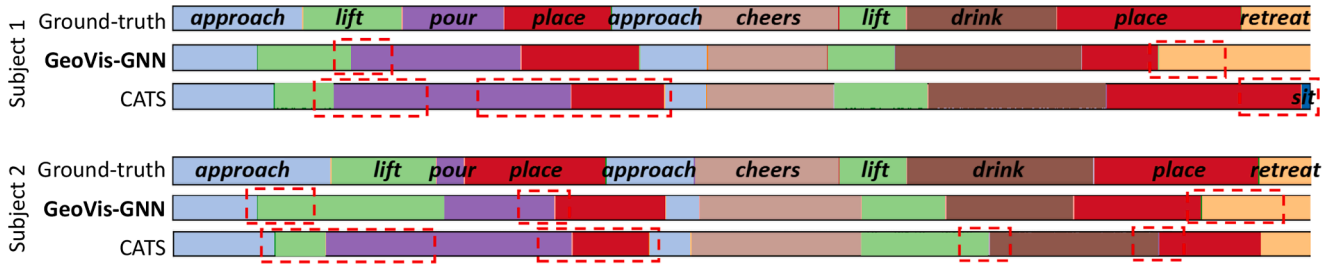

Fig. 6. Visualization of segmentation on MPHOI-72 for *Cheering* activity. Red dashed boxes highlight major segmentation errors.

Table 4
Joined segmentation and label recognition results on CAD-120.

Model	Sub-activity			Object Affordance		
	F_1@10	F_1@25	F_1@50	F_1@10	F_1@25	F_1@50
rCRF	65.6 ± 3.2	61.5 ± 4.1	47.1 ± 4.3	72.1 ± 2.5	69.1 ± 3.3	57.0 ± 3.5
Independent BiRNN	70.2 ± 5.5	64.1 ± 5.3	48.9 ± 6.8	84.6 ± 2.1	81.5 ± 2.7	71.4 ± 4.9
ATCRF	72.0 ± 2.8	68.9 ± 3.6	53.5 ± 4.3	79.9 ± 3.1	77.0 ± 4.1	63.3 ± 4.9
Relational BiRNN	79.2 ± 2.5	75.2 ± 3.5	62.5 ± 5.5	82.3 ± 2.3	78.5 ± 2.7	68.9 ± 4.9
ASSIGN	88.0 ± 1.8	84.8 ± 3.0	73.8 ± 5.8	92.0 ± 1.1	90.2 ± 1.8	82.4 ± 3.5
2G-GCN	89.5 ± 1.6	87.1 ± 1.8	76.2 ± 2.8	92.4 ± 1.7	**90.4** ± 2.3	82.7 ± 2.9
CATS	89.6 ± 2.1	87.3 ± 1.5	76.0 ± 3.5	90.2 ± 1.5	89.1 ± 2.4	80.5 ± 2.8
GeoVis-GNN	**89.9** ± 2.0	**87.8** ± 1.9	**76.7** ± 3.1	**92.7** ± 0.4	**90.4** ± 0.6	**83.3** ± 1.8

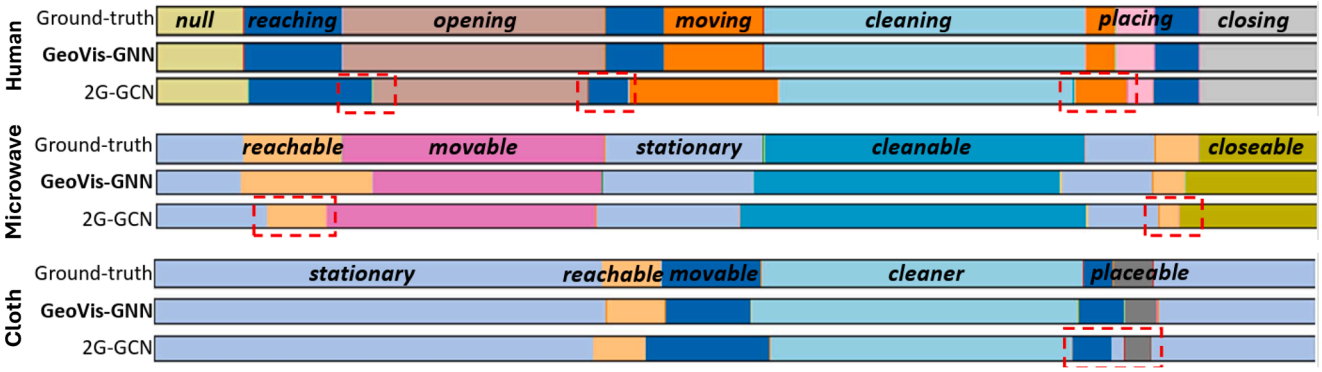

Fig. 7. Visualization of segmentation on CAD-120 for *Cleaning objects* activity. Red dashed boxes highlight major segmentation errors.

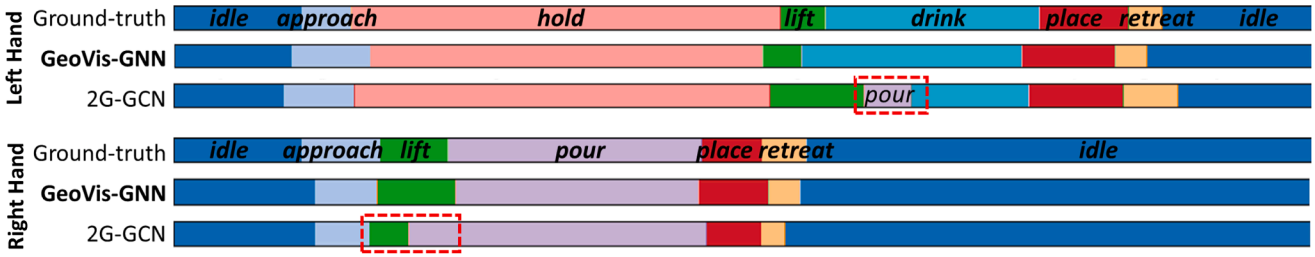

Fig. 8. Visualization of segmentation on Bimanual Actions for *Pouring* activity. Red dashed boxes highlight major segmentation errors.

Table 5
Joined segmentation and label recognition results on Bimanual Actions.

Model	Sub-activity		
	F_1@10	F_1@25	F_1@50
Dreher et al. (2020)	40.6 ± 7.2	34.8 ± 7.1	22.2 ± 5.7
Independent BiRNN	74.8 ± 7.0	72.0 ± 7.0	61.8 ± 7.3
Relational BiRNN	77.7 ± 3.9	75.0 ± 4.2	64.8 ± 5.3
ASSIGN	84.0 ± 2.0	81.2 ± 2.0	68.5 ± 3.3
2G-GCN	85.0 ± 2.2	82.0 ± 2.6	69.2 ± 3.1
GeoVis-GNN	**85.8** ± 2.2	**82.7** ± 2.8	**69.7** ± 3.0

Table 6
GeoVis-GNN performance across different HOI scenarios. The rightmost column indicates the performance gap relative to MPHOI-120.

Dataset	Scenario	Difference (F_1@10)
CAD-120	Single-person (General)	+24.8 %
Bimanual Actions	Single-person (Bimanual)	+20.7 %
MPHOI-72	Two-person (Full)	+19.2 %
MPHOI-120	Multi-person (Partial)	–

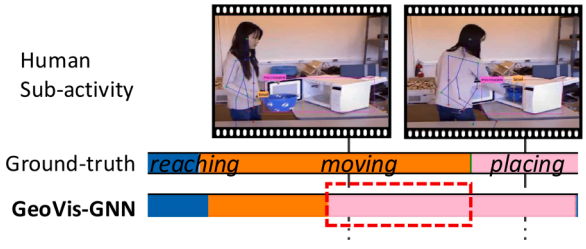

Fig. 9. Example of sub-activity segmentation error on the CAD-120 dataset for *Microwaving Food* activity. Corresponding RGB frames are provided for visual context.

effectively. Furthermore, Fig. 9 shows an example of sub-activity segmentation by GeoVis-GNN compared to the ground-truth, with corresponding RGB screenshots for visual reference. The segmentation error in the red box occurs because the model misclassifies part of the *moving* phase as *placing*, likely due to the smooth transition and similar motion patterns between the two sub-activities. This suggests that the model lacks sensitivity to subtle temporal boundaries. Improving temporal modeling or introducing boundary-aware supervision could help address this issue.

In multi-person settings, the challenges are more pronounced. MPHOI-72 focuses on two fully engaged participants and has a +19.2 % advantage. Although moderate occlusions and overlapping actions are present, the model generally maintains good performance. In contrast,

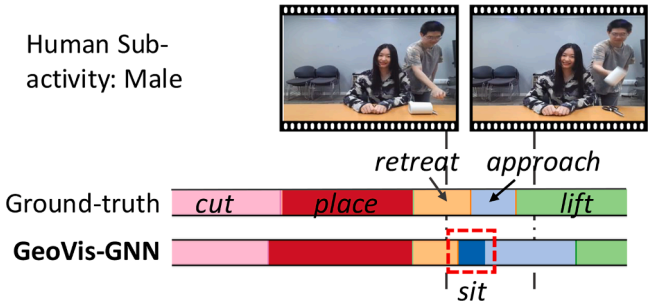

Fig. 10. Example of sub-activity segmentation error on the MPHOI-72 dataset for *Hair Cutting* activity. Corresponding RGB frames are provided for visual context.

MPHOI-120, which features partial engagements, idle participants, and concurrent interactions, yields a significantly lower score of 65.1 % due to heavy occlusions and ambiguous sub-activity boundaries. These results highlight the difficulty of accurately segmenting short actions and managing overlapping activities in crowded, dynamic scenes. Fig. 10 shows an over-segmentation error by GeoVis-GNN on the MPHOI-72 dataset during the *Hair Cutting* activity. As highlighted in the red box, the model incorrectly inserts a *sit* action between *place* and *approach*. This likely results from short-term pose ambiguity, causing the model to misinterpret a brief motion pause as a distinct sub-activity. This suggests the need for improved temporal smoothing to reduce false segment boundaries.

5.5. Cross-dataset zero-shot study

In real-world applications, models usually perform reliably on unseen data distributions without the luxury of extensive retraining or domain-specific adaptations. To demonstrate the robustness and generalization capabilities of our proposed GeoVis-GNN, we conduct a cross-dataset zero-shot evaluation, as detailed in Table 7. This study involves training GeoVis-GNN exclusively on the concurrent partial interaction dataset and subsequently testing it on the two-person HOI dataset.

Our results show that GeoVis-GNN significantly outperforms the existing baselines, ASSIGN, 2G-GCN and CATS, achieving an improvement of 3.6 % in the F_1@10 score. This substantial performance gain underscores the stronger generalization ability of GeoVis-GNN compared to state-of-the-art methods. The ability to effectively transfer learned features from a more complex concurrent partial HOI scenario to a simpler two-person setting highlights the model's adaptability and transferability across diverse multi-person HOI datasets.

Additionally, in many real-world scenarios, target domain fine-tuning or transfer learning is often employed to adapt models to specific environments. However, our zero-shot results, while not reaching the performance levels achievable when training and testing on the same two-person dataset, are achieved without any such fine-tuning, relying solely on training with one dataset and testing on another that do not necessarily share a direct relationship. This suggests that GeoVis-GNN has the potential to generalize across different datasets with varying characteristics, even without extensive retraining. Although there is room for improvement, the results are promising and indicate that our approach can still be valuable in scenarios where labeled data for every possible situation may not be readily available.

Table 7
Zero-shot results of training on concurrent partial interaction dataset (MPHOI-120) and testing on two-person HOI dataset (MPHOI-72).

Model	Sub-activity		
	F_1@10	F_1@25	F_1@50
ASSIGN	33.7	31.5	28.2
2G-GCN	36.2	33.3	30.4
CATS	38.5	35.6	33.2
GeoVis-GNN	42.1	40.3	34.5

Table 8
Results of different strategies in channel attention-based feature fusion on MPHOI-120.

Model	Sub-activity		
	F_1@10	F_1@25	F_1@50
(a) ho feature-channel attention	58.2 ± 4.0	50.7 ± 4.2	38.4 ± 3.6
(b) ho entity-channel attention	61.4 ± 5.7	56.5 ± 5.3	40.4 ± 4.7
(c) vg entity-channel attention	59.1 ± 5.3	50.4 ± 6.0	39.9 ± 4.8
d) GeoVis-GNN (ours)	65.1 ± 5.2	59.8 ± 4.7	46.6 ± 5.1

5.6. Ablation study and alternative architecture

We extensively evaluate the design of channel attention-based feature fusion. Fig. 11 shows four design strategies, in which: **(a)**: Separately concatenate human features *hv, hg* and object features *ov, og* on feature-channel with attentions; **(b)**: Separately concatenate human features *hv, hg* and object features *ov, og* on entity-channel with attentions; **(c)**: Separately concatenate visual features *hv, ov* and geometric features *hg, og* on entity-channel with attentions; **(d) Ours**: Concatenate all features *hv, hg, ov, og* on entity-channel with a unified attention. The results of the comparison are shown in Table 8. Our design **(d)** presents the highest F_1 score with a significant improvement gap *w.r.t.* other designs. Notably, design **(a)** shows the lowest score, indicating the importance of entity-channel fusion. Although **(b)** and **(c)** contribute relatively high score, they still show 3.7 % and 6 % performance degradation in F_1@10, respectively. This demonstrates the efficiency of our holistic entity-channel attention in selectively enhancing the most crucial visual or geometric features among all entities.

To further validate the effectiveness and complementary roles of each module, we conduct ablation studies on MPHOI-120, where CAF and IEG refer to the channel attention-based fusion and the interdependent entity graph, respectively (Table 9). Specifically, variant (1) removes IEG, variant (2) removes both CAF and IEG, variant (3) removes CAF and IEG while replacing the GAT-based geometric embedding with GCN, and variant (4) adopts an alternative top-down design instead of our bottom-up architecture.

Our results show that removing any module leads to a significant performance drop, confirming that each component not only addresses a specific challenge but also enhances the entire pipeline. For instance, variant (1) sees a 3.9 % decline in F_1@10, highlighting the critical role of IEG in modeling complex entity interactions. Likewise, variants (2) and (3) drop by 5.8 % and 6.5 %, respectively, underscoring the importance of CAF for effectively merging geometric and vi-

Table 9
Architecture alternative and ablation study on MPHOI-120. CAF and IEG denote the channel attention-based fusion and the interdependent entity graph, respectively.

Model	Sub-activity		
	F_1@10	F_1@25	F_1@50
(1) GAT, w CAF, w/o IEG	61.2 ± 6.0	55.7 ± 5.2	45.4 ± 4.6
(2) GAT, w/o CAF&IEG	59.3 ± 6.1	52.5 ± 5.7	39.4 ± 4.3
(3) GCN, w/o CAF&IEG	58.6 ± 6.4	51.5 ± 5.3	38.3 ± 5.7
(4) Top-down architecture	62.8 ± 5.7	56.7 ± 5.2	42.8 ± 4.9
(5) GeoVis-GNN (ours)	65.1 ± 5.2	59.8 ± 4.7	46.6 ± 5.1

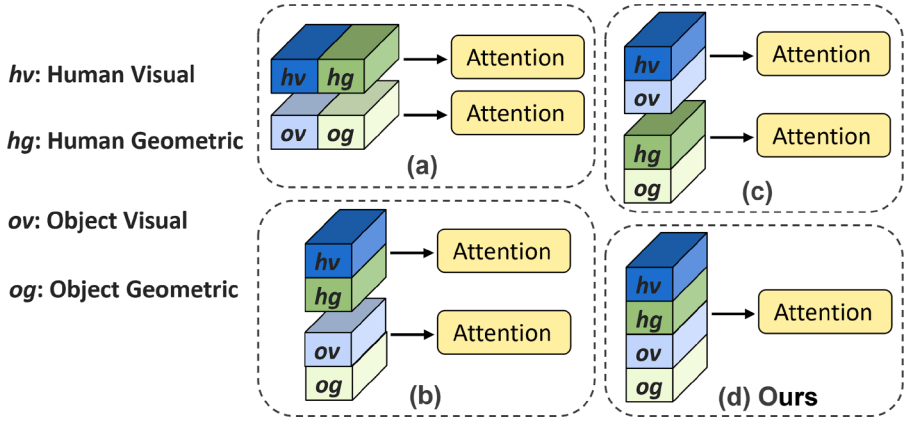

Fig. 11. Different designs to combine geometric and visual features in channel attention-based feature fusion.

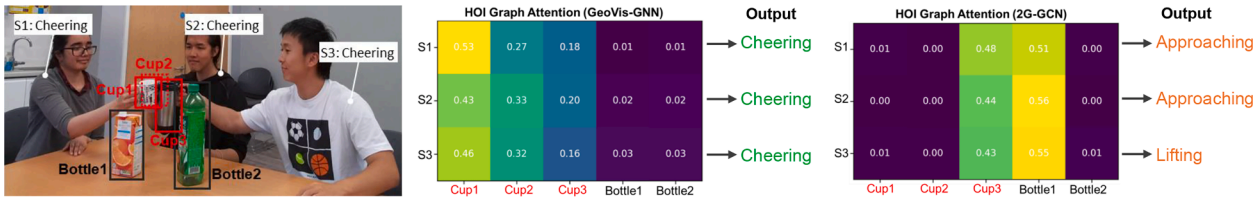

Fig. 12. Visualization of HOI attention maps for GeoVis-GNN and 2G-GCN during a "Cheering" activity. Correct and incorrect recognition results are highlighted in green and orange, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

sual features, and the GAT for generating expressive geometric embeddings. Moreover, comparing the top-down approach (variant (4)) to our final method (5) reveals that the bottom-up framework better integrates multimodal information and preserves fine-grained entity details. These findings collectively demonstrate the synergy among GAT, CAF, and IEG, where each module contributes to robust HOI recognition by providing refined features that the subsequent modules further leverage, resulting in more accurate segmentation and interaction understanding.

5.7. HOI attention analysis

To enhance the interpretability of our model, we deep into the attention analysis in the HOI graph. We compare GeoVis-GNN with the recent advanced method that constructs entity-level HOI graphs. Fig. 12 presents a comparative analysis of HOI attention maps in entity-level graphs generated by GeoVis-GNN and 2G-GCN for a *Cheering* activity involving three subjects, each holding a cup, with two bottles placed on the table. In the left attention map, our GeoVis-GNN model demonstrates its superior interpretability by accurately focusing on all three cups, even effectively handling occlusions, such as Cup2 being partially hidden behind Cup1. This targeted attention enables the model to correctly recognize the *Cheering* sub-activity for all three subjects (highlighted in green).

In contrast, the 2G-GCN model exhibits less precise attention, incorrectly focusing on Cup3 and Bottle1, leading to erroneous sub-activity predictions such as *Approaching* and *Lifting* (highlighted in orange). This comparison highlights GeoVis-GNN's ability to maintain robust attention across relevant entities, even in occluded or cluttered environments, thereby ensuring more accurate HOI recognition. The clear distinction in attention focus between the two models underscores the effectiveness of our bottom-up approach in capturing the essential elements of complex interactions, which is critical for accurate activity recognition in multi-person scenarios.

Table 10
Results of different number of object usage on MPHOI-120.

Model	Sub-activity		
	F_1@10	F_1@25	F_1@50
2 objects only	61.4 ± 3.4	55.4 ± 2.0	40.1 ± 3.2
3 objects only	62.6 ± 6.9	56.2 ± 8.2	41.8 ± 9.1
4 objects only	63.1 ± 6.4	56.7 ± 7.5	43.2 ± 8.7
GeoVis-GNN (5 objects)	65.1 ± 5.2	59.8 ± 4.7	46.6 ± 5.1

5.8. Analysis of varying number of objects

Table 10 presents a comprehensive analysis of our model's performance when varying the number of objects considered on the MPHOI-120 dataset. Notably, MPHOI-120 contains 2–5 objects in total, even when using only 2 objects, our model outperforms the 2G-GCN baseline, demonstrating its robustness and highlighting its capability to extract meaningful interactions even from a limited set of objects.

Increasing the number of objects from 2 to 5 improves performance across all F_1 metrics, but also increases memory cost. This trade-off suggests that while more objects provide richer interaction contexts, leading to better recognition accuracy, the memory requirements scale with the number of objects included. However, in highly cluttered environments with potentially hundreds of objects, our design offers an advantage by enabling the selection of a fixed number of objects to avoid a linear increase in memory consumption.

5.9. Parameter size and inference time analysis

To verify the efficiency of our approach, Table 11 compares GeoVis-GNN against 2G-GCN, CATS, and ASSIGN, all of which represent recent state-of-the-art HOI recognition frameworks, on the MPHOI-120 dataset. Specifically, 2G-GCN and CATS combine geometric and vi-

Table 11
Comparison of parameter size (M), inference time (millisecond per frame) and performance between GeoVis-GNN and state-of-the-arts on MPHOI-120.

Model	Param (M)	Time (ms/f)	Sub-activity F_1@10
ASSIGN	121	79	58.0 ± 8.5
2G-GCN	148	84	60.7 ± 6.5
CATS	132	137	62.8 ± 2.7
GeoVis-GNN	130	119	**65.1** ± 5.2

sual cues, while ASSIGN only focuses on visual features. GeoVis-GNN demonstrates greater efficiency with a smaller parameter size (130M) compared to 2G-GCN (148M) and CATS (132M), while achieving competitive interactive times of 119 ms/f. Additionally, GeoVis-GNN achieves a notable performance improvement, underscoring its capability to balance efficiency and accuracy effectively in multi-person HOI recognition tasks.

6. Conclusion and discussion

Our bottom-up GeoVis-GNN framework for video-based multi-person HOI recognition introduces a novel dual-attention fusion mechanism. It optimizes feature integration by embedding and fusing visual and geometric features using a graph attention mechanism followed by a channel attention module. These enhanced entity-specific representations are then fed into an interdependent entity graph, enabling the modeling of both explicit interactions and implicit interdependencies for a more comprehensive understanding of multi-person HOI. Additionally, we propose a challenging concurrent partial interaction dataset and GeoVis-GNN sets new benchmarks across various HOI scenarios.

Our attention-based feature fusion effectively handles scenes with multiple entities by discerning dynamic relevance and underlying connections among individuals. In highly cluttered environments - where dozens of people or objects may overlap - the root issue is that key interactions, whether contact-based or not, risk being overwhelmed by irrelevant visual clutter. For instance, a person watching TV in a room filled with other objects and people may go unnoticed if the system cannot separate important cues from background noise. This interplay between partial interactions and large-scale clutter underscores a deeper need for efficient extraction of both in-contact and non-contact interactions (Hassan, Ghosh, Tesch, Tzionas, & Black, 2021; Jiang, Koppula, & Saxena, 2013; Nie, Dai, Han, & Nießner, 2022). Identifying the most probable HOIs in such scenarios requires robust methods for filtering out extraneous information and focusing on contextually meaningful entities.

While our concurrent partial interaction dataset closely reflects real-world multi-person HOIs, its controlled indoor settings do not fully mirror the unpredictability of in-the-wild situations. The underlying cause is that real-world environments often introduce variables like inconsistent lighting, unpredictable occlusions, diverse camera angles, and partially missing objects (Tripathi et al., 2023; Yang, Zhai, Luo, Cao, & Zha, 2024; Ye, Wang, Li, & Zhang, 2023). These factors, compounded by more fluid participant behaviors, lead to greater data ambiguity and annotation difficulty. Although we capture significant variation in our dataset, future work will extend to in-the-wild HOI videos. Tackling these unstructured real-world contexts requires innovative strategies to handle sudden motion, incomplete viewpoints, and other complexities beyond the scope of indoor, well-annotated data.

Contemporary HOI recognition often depends on precise, frame-level annotations (Li, Du, Torralba, Sivic, & Russell, 2021), which become costly and inconsistent when interactions are frequent and subtle - common traits in multi-person environments. The core problem is that a large volume of overlapping sub-activities escalates labeling complexity, amplifying human errors and making the labeling process time-consuming. Moreover, ambiguous transitions (e.g., partial engagement or fleeting interactions) make it hard for annotators to decide when a sub-activity starts or ends. Weakly-supervised learning (Ren, Yang, Zhang, & Zhang, 2023; Rizve et al., 2023) mitigates this challenge by using approximate or high-level labels, allowing models to generalize without requiring every frame to be manually annotated. As a result, this approach offers a scalable pathway for handling diverse, real-world HOI data, where precise and exhaustive annotations may be neither feasible nor reliable.

CRediT authorship contribution statement

Tanqiu Qiao: Conceptualization, Methodology, Software, Validation, Investigation, Writing – original draft, Writing – review & editing; **Ruochen Li:** Methodology, Software, Validation, Investigation, Writing – original draft, Writing – review & editing; **Frederick W.B. Li:** Conceptualization, Methodology, Writing – original draft; **Yoshiki Kubotani:** Writing – original draft, Software, Data curation, Visualization; **Shigeo Morishima:** Conceptualization, Supervision, Project administration, Funding acquisition; **Hubert P. H. Shum:** Conceptualization, Methodology, Writing – original draft, Supervision, Project administration, Funding acquisition.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research is supported in part by the EPSRC NorthHFutures project (ref: EP/X031012/1).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.eswa.2025.128344.

References

Baradel, F., Wolf, C., & Mille, J. (2017). Human action recognition: Pose-based attention draws focus to hands. In *Proceedings of the IEEE/CVF international conference on computer vision workshops (ICCVW)* (pp. 604–613).

Baradel, F., Wolf, C., & Mille, J. (2018a). Human activity recognition with pose-driven attention to RGB. In *British machine vision conference (BMVC)* (pp. 1–14).

Baradel, F., Wolf, C., Mille, J., & Taylor, G. W. (2018b). Glimpse clouds: Human activity recognition from unstructured feature points. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 469–478).

Bhatnagar, B. L., Xie, X., Petrov, I. A., Sminchisescu, C., Theobalt, C., & Pons-Moll, G. (2022). Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 15935–15946).

Boulahia, S. Y., Amamra, A., Madi, M. R., & Daikh, S. (2021). Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Machine Vision and Applications*, *32*(6), 121.

Brody, S., Alon, U., & Yahav, E. (2021). How attentive are graph attention networks? arXiv preprint arXiv:2105.14491.

Bruce, X. B., Liu, Y., & Chan, K. C. C. (2021). Multimodal fusion via teacher-student network for indoor action recognition. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 3199–3207). (*vol. 35*).

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., & Sheikh, Y. (2018). OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *arXiv e-prints*, (pp. 1812).

Cheng, Q., Cheng, J., Liu, Z., Ren, Z., & Liu, J. (2024a). A dense-sparse complementary network for human action recognition based on RGB and skeleton modalities. *Expert Systems with Applications*, *244*, 123061.

Cheng, Y., Duan, H., Wang, C., & Chen, Z. (2024b). Parallel disentangling network for human–object interaction detection. *Pattern Recognition*, *146*, 110021.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

Cob-Parro, A. C., Losada-Gutiérrez, C., Marrón-Romera, M., Gardel-Vicente, A., & Bravo-Muñoz, I. (2024). A new framework for deep learning video based human action recognition on the edge. *Expert Systems with Applications*, *238*, 122220.

Damen, D., Doughty, H., Farinella, G. M., Furnari, A., Ma, J., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., & Wray, M. (2021). Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. *International Journal of Computer Vision (IJCV)*. *30*, 33–55.

Das, S., Sharma, S., Dai, R., Bremond, F., & Thonnat, M. (2020). VPN: Learning video-pose embedding for activities of daily living. In *European conference on computer vision (ECCV)* (pp. 72–90).

Dreher, C. R. G., Wächter, M., & Asfour, T. (2020). Learning object-action relations from bimanual human demonstration using graph networks. *IEEE Robotics and Automation Letters*, *5*(1), 187–194.

Fan, M., Chen, M., Hu, C., & Zhou, S. (2023). Occˆ 2Net: Robust image matching based on 3D occupancy estimation for occluded regions. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 9652–9662).

Farha, Y. A., & Gall, J. (2019). MS-TCN: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 3575–3584).

Fouhey, D. F., Kuo, W.-c., Efros, A. A., & Malik, J. (2018). From lifestyle Vlogs to everyday interactions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 4991–5000).

Gao, J., Sun, C., Zhao, H., Shen, Y., Anguelov, D., Li, C., & Schmid, C. (2020). VectorNet: Encoding HD maps and agent dynamics from vectorized representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 11525–11533).

Hassan, M., Ghosh, P., Tesch, J., Tzionas, D., & Black, M. J. (2021). Populating 3D scenes by learning human-scene interaction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 14708–14718).

Heitzinger, T., & Kampel, M. (2023). A fast unified system for 3D object detection and tracking. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 17044–17054).

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 7132–7141).

Hu, Z., Xiao, J., Li, L., Liu, C., & Ji, G. (2024). Human-centric multimodal fusion network for robust action recognition. *Expert Systems with Applications*, *239*, 122314.

Jang, E., Gu, S., & Poole, B. (2016). Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:1611.01144.

Jiang, Y., Koppula, H., & Saxena, A. (2013). Hallucinated humans as the hidden context for labeling 3d scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 2993–3000).

Kim, S., Jung, D., & Cho, M. (2023). Relational context learning for human-object interaction detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2925–2934).

Koppula, H. S., Gupta, R., & Saxena, A. (2013). Learning human activities and object affordances from RGB-D videos. *The International Journal of Robotics Research*, *32*(8), 951–970.

Koppula, H. S., & Saxena, A. (2016). Anticipating human activities using object affordances for reactive robotic response. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(1), 14–29.

Krebs, F., Meixner, A., Patzer, I., & Asfour, T. (2021). The KIT bimanual manipulation dataset. In *IEEE/RAS International conference on humanoid robots (humanoids)* (pp. 0–0).

Kresović, M., & Nguyen, T. D. (2021). Bottom-up approaches for multi-person pose estimation and it's applications: A brief review. arXiv preprint arXiv:2112.11834.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A. et al. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, *123*(1), 32–73.

Le, H., Sahoo, D., Chen, N. F., & Hoi, S. C. H. (2020). BiST: Bi-directional spatio-temporal reasoning for video-grounded dialogues. arXiv preprint arXiv:2010.10095.

Lea, C., Flynn, M. D., Vidal, R., Reiter, A., & Hager, G. D. (2017). Temporal convolutional networks for action segmentation and detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 156–165).

Li, L., Wei, J., Wang, W., & Yang, Y. (2024). Neural-logic human-object interaction detection. *Advances in Neural Information Processing Systems*, *36*, 21158–21171.

Li, S., Du, Y., Torralba, A., Sivic, J., & Russell, B. (2021). Weakly supervised human-object interaction detection in video via contrastive spatiotemporal regions. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1845–1855).

Liang, M., Yang, B., Hu, R., Chen, Y., Liao, R., Feng, S., & Urtasun, R. (2020). Learning lane graph representations for motion forecasting. In *European conference on computer vision (ECCV)* (pp. 541–556). Springer.

Lin, M., Chen, Q., & Yan, S. (2013). Network in network. arXiv preprint arXiv:1312.4400.

Liu, Y., Liu, Y., Jiang, C., Lyu, K., Wan, W., Shen, H., Liang, B., Fu, Z., Wang, H., & Yi, L. (2022). HOI4D: A 4D egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 21013–21022).

Ma, S., Wang, Y., Wang, S., & Wei, Y. (2023). FGAHOI: Fine-grained anchors for human-object interaction detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. *46*, 2415–2429.

Maraghi, V. O., & Faez, K. (2019). Zero-shot learning on human-object interaction recognition in video. In *Iranian conference on signal processing and intelligent systems (ICSPIS)* (pp. 1–7).

Materzynska, J., Xiao, T., Herzig, R., Xu, H., Wang, X., & Darrell, T. (2020). Something-else: Compositional action recognition with spatial-temporal interaction networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 1049–1059).

Microsoft (2022). Quickstart: Set up azure kinect body tracking. https://docs.microsoft.com/en-us/azure/kinect-dk/body-sdk-setup.

Morais, R., Le, V., Venkatesh, S., & Tran, T. (2021). Learning asynchronous and sparse human-object interaction in videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 16041–16050).

Nie, Y., Dai, A., Han, X., & Nießner, M. (2022). Pose2Room: Understanding 3D scenes from human activities. In *European conference on computer vision (ECCV)* (pp. 425–443). Springer.

Park, J., Park, J.-W., & Lee, J.-S. (2023). ViPLO: Vision transformer based pose-conditioned self-loop graph for human-object interaction detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 17152–17162).

Qiao, T., Li, R., Li, F. W. B., & Shum, H. P. H. (2024). From category to scenery: An end-to-end framework for multi-person human-object interaction recognition in videos. In *International conference of pattern recognition*.

Qiao, T., Men, Q., Li, F. W. B., Kubotani, Y., Morishima, S., & Shum, H. P. H. (2022). Geometric features informed multi-person human-object interaction recognition in videos. In *European conference on computer vision (ECCV)*.

Ren, H., Yang, W., Zhang, T., & Zhang, Y. (2023). Proposal-based multiple instance learning for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2394–2404).

Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster r-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149.

Rizve, M. N., Mittal, G., Yu, Y., Hall, M., Sajeev, S., Shah, M., & Chen, M. (2023). Pivotal: Prior-driven supervision for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 22992–23002).

Samet, N., Hicsonmez, S., & Akbas, E. (2020). HoughNet: Integrating near and long-range evidence for bottom-up object detection. In *European conference on computer vision (ECCV)* (pp. 406–423). Springer.

Sener, O., & Saxena, A. (2015). rCRF: Recursive belief estimation over CRFs in RGB-d activity videos. In *Robotics: Science and systems*.

Setiawan, F., Yahya, B. N., Chun, S.-J., & Lee, S.-L. (2022). Sequential inter-hop graph convolution neural network (SIhGCN) for skeleton-based human action recognition. *Expert Systems with Applications*, *195*, 116566.

Shu, T., Gao, X., Ryoo, M. S., & Zhu, S.-C. (2017). Learning social affordance grammar from videos: transferring human interactions to human-robot interactions. In *2017 IEEE International conference on robotics and automation (ICRA)* (pp. 1669–1676).

Shu, T., Ryoo, M. S., & Zhu, S.-C. (2016). Learning social affordance for human-robot interaction. arXiv preprint arXiv:1604.03692.

Tan, K. S., Lim, K. M., Lee, C. P., & Kwek, L. C. (2022). Bidirectional long short-term memory with temporal dense sampling for human action recognition. *Expert Systems with Applications*, *210*, 118484.

Tripathi, S., Chatterjee, A., Passy, J.-C., Yi, H., Tzionas, D., & Black, M. J. (2023). Deco: Dense estimation of 3D human-scene contact in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 8001–8013).

Tu, D., Sun, W., Min, X., Zhai, G., & Shen, W. (2022). Video-based human-object interaction detection from tubelet tokens. *Advances in Neural Information Processing Systems*, *35*, 23345–23357.

Tu, D., Sun, W., Zhai, G., & Shen, W. (2023). Agglomerative transformer for human-object interaction detection. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 21614–21624).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008). (*vol. 30*).

Wan, B., Zhou, D., Liu, Y., Li, R., & He, X. (2019). Pose-aware multi-level feature network for human object interaction detection. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 9469–9478).

Wang, H., Zhou, L., Chen, Y., Tang, M., & Wang, J. (2022a). Regularizing vector embedding in bottom-up human pose estimation. In *European conference on computer vision* (pp. 107–122). Springer.

Wang, N., Zhu, G., Li, H., Feng, M., Zhao, X., Ni, L., Shen, P., Mei, L., & Zhang, L. (2023). Exploring spatio–temporal graph convolution for video-based human–object interaction recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *33*(10), 5814–5827.

Wang, N., Zhu, G., Zhang, L., Shen, P., Li, H., & Hua, C. (2021). Spatio-temporal interaction graph parsing networks for human-object interaction recognition. In *Proceedings of the ACM international conference on multimedia (ACMM)* (pp. 4985–4993).

Wang, W., Shen, J., Cheng, M.-M., & Shao, L. (2019). An iterative and cooperative top-down and bottom-up inference network for salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 5968–5977).

Wang, Y., Li, M., Cai, H., Chen, W.-M., & Han, S. (2022b). Lite pose: Efficient architecture design for 2D human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13126–13136).

Wu, X., Li, Y.-L., Liu, X., Zhang, J., Wu, Y., & Lu, C. (2022). Mining cross-person cues for body-part interactiveness learning in HOI detection. In *European conference on computer vision* (pp. 121–136). Springer.

Xing, H., & Burschka, D. (2022). Understanding spatio-temporal relations in human-object interaction using pyramid graph convolutional network. In *2022 IEEE/RSJ International conference on intelligent robots and systems (IROS)* (pp. 5195–5201).

Yang, Y., Zhai, W., Luo, H., Cao, Y., & Zha, Z.-J. (2024). Lemon: Learning 3D human-object interaction relation from 2D images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16284–16295).

Ye, Q., Wang, X., Li, R., & Zhang, Y. (2023). Human object interaction detection based on feature optimization and key human-object enhancement. *Journal of Visual Communication and Image Representation*, *93*, 103824.

You, Y., Liu, H., Wang, T., Li, W., Ding, R., & Li, X. (2023). Co-evolution of pose and mesh for 3D human body estimation from video. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 14963–14973).

Yu, Q., Tanaka, M., & Fujiwara, K. (2024). Exploring vision transformers for 3D human motion-language models with motion patches. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 937–946).

Zhai, K., Nie, Q., Ouyang, B., Li, X., & Yang, S. (2023). HopFIR: Hop-wise graphformer with intragroup joint refinement for 3D human pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 14985–14995).

Zhang, Y., Pan, Y., Yao, T., Huang, R., Mei, T., & Chen, C.-W. (2022). Exploring structure-aware transformer over interaction proposals for human-object interaction detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 19548–19557).

Zheng, C., Wu, W., Chen, C., Yang, T., Zhu, S., Shen, J., Kehtarnavaz, N., & Shah, M. (2023). Deep learning-based human pose estimation: A survey. *ACM Computing Surveys*, *56*(1), 1–37.

Zhou, J., Wang, Z., Meng, J., Liu, S., Zhang, J., & Chen, S. (2022). Human interaction recognition with skeletal attention and shift graph convolution. In *International joint conference on neural networks (IJCNN)* (pp. 1–8). IEEE.

Zhou, X., Zhuo, J., & Krahenbuhl, P. (2019). Bottom-up object detection by grouping extreme and center points. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 850–859).

Zhu, M., Ho, E. S. L., Chen, S., Yang, L., & Shum, H. P. H. (2024). Geometric features enhanced human-object interaction detection. *IEEE Transactions on Instrumentation and Measurement. 73*, 1–14.