Bi-Causal: Group Activity Recognition via Bidirectional Causality

Supplementary Material

1. Granger Causality Test

To enhance the reliability of our hypothesis, We employ the Granger Causality Test to investigate the presence and strength of the bidirectional causality. We utilize two feature extractors to derive information from human relations and human-object interactions separately. Both of these feature extractors are consistent with the one used in our Bi-Causal framework. We compare two models: one that includes only the auto-regressive model of human relations feature sequences and another that combines human relations features and human-object interaction features in a multiple regression model. By comparing the performance, we can evaluate whether human-object interactions causally influence human relations. A large disparity in the impacts indicates a robust causality relationship. The same approach can also be employed to ascertain the causal impact of human relations on human-object interactions.

Auto-regressive model. We estimate the self-influence with the residual between the human relation feature and the reconstructed relation feature, which is reconstructed with the historical feature of the human relation itself. We use $\mathbf{R}_i \in \mathbb{R}^D$ (***cls***$_{rela}$ obtained in the main text method) to denote human relation features. Based on the past t instances of human relation features, we employ a transformer encoder $En(\cdot)$ to reconstruct human relation features.

$$En() = Encoder(cat()) \tag{1}$$

$$[\hat{\mathbf{R}}_k, \mathbf{R}_{list}] = En([\textbf{token}, \mathbf{R}_{k-t}, \ldots, \mathbf{R}_{k-2}, \mathbf{R}_{k-1}]) \tag{2}$$

where we utilize $\hat{\mathbf{R}}_k$ as reconstructed feature for comparison with the feature $\mathbf{R}_k$. **token** is a fixed, initialized feature used to learn from relation features and generate $\hat{\mathbf{R}}_k$. The self-influence ssr_1 is estimated with the sum of squares residual (SSR) and we employ SSR as the loss function for training the model.

$$ssr_1 = \sum_k ||\hat{\mathbf{R}}_k - \mathbf{R}_k||_2^2 \tag{3}$$

Multiple-regression model. In this model, apart from human relation features, we introduce human-object interaction features for feature reconstruction, aiming to estimate the causal impact of the latter on the former. We still use a feature sequence of length t as input and compare it with the human relation feature $\mathbf{R}_k$.

$$[\hat{\mathbf{R}}_k^I, \mathbf{R}_{list}] = En([\textbf{token}, \mathbf{R}_{k-t}, \mathbf{I}_{k-t}, \ldots, \mathbf{R}_{k-1}, \mathbf{I}_{k-1}]) \tag{4}$$

$$ssr_2 = \sum_k ||\hat{\mathbf{R}}_k^I - \mathbf{R}_k||_2^2 \tag{5}$$

where $\mathbf{I}_i \in \mathbb{R}^D$ (***cls***$_{inter}$ obtained in the main text method) denotes a human-object interaction features and $\hat{\mathbf{R}}_k^I$ means the reconstructed with $\mathbf{I}$. We still employ ssr_2 as the loss function for training this model.

Since the sum of squares of random variables following a Gaussian distribution follows the chi-squared ($\chi^{\mathbf{2}}$) distribution, the prediction error, being a sum of squares, also follows the $\chi^{\mathbf{2}}$ distribution. Therefore, we can construct an F-distribution for hypothesis testing.

$$F_{\mathcal{I}\to\mathcal{R}} = \frac{(ssr_1 - ssr_2)/td}{ssr_2/(n_t - m)} \tag{6}$$

Once we establish the significance level α, the value of $F_{\mathcal{I}\to\mathcal{R}}$ can be used to validate the causal relationship. The causal impact of human relations on human-object interactions can also be validated using the methods described above. According to Equation 6, we obtain two F-statistics, $F_{\mathcal{R}\to\mathcal{I}}$ and $F_{\mathcal{I}\to\mathcal{R}}$, which indicate the level of confidence in the bidirectional causality between human relations and human-object interactions.

2. Comparision with SOTA in Weakly Supervised Group Activity Recognition

In real-world applications, Group Activity Recognition (GAR) faces numerous obstacles, particularly with the scarcity of extensive annotations such as bounding boxes and individual actions. Consequently, Yan *et al.* [14] introduced Weakly Supervised GAR (WSGAR), which eliminates the need for actor-level labels in both training and inference, thereby further reducing annotation costs. Our approach demonstrated excellent efficacy in the fully supervised GAR. To more rigorously validate the effectiveness of our Bi-Causal model, we conducted tests within a weakly supervised scenario.

NBA dataset. The NBA dataset [14] comprises a total of 9,172 labeled clips from 181 NBA videos, with 7,624 clips designated for training and 1,548 clips earmarked for testing. Presently, it stands as the singular dataset posited for WSGAR and the most extensive dataset for group activity recognition. Each clip is annotated with one of nine group activities, without information on individual actions or bounding boxes. In terms of evaluation, we employ the metrics of Multi-class Classification Accuracy (MCA) and Mean Per Class Accuracy (MPCA), the latter chosen to address the class imbalance inherent in the dataset.

The NBA dataset lacks annotations such as bounding boxes and individual action labels, making it challenging to

Method	Backbone	NBA	W-Volleyball
AT [2]	ResNet-18	47.1 / 41.5	84.3 / 89.6
SAM [14]	Incep-v3	49.1 / 47.5	– / 94.0
SAM [14]	ResNet-18	54.3 / 51.5	86.3 / 93.1
Dual [3]	Incep-v3	51.5 / 44.8	– / 95.8
SACRF [7]	ResNet-18	56.3 / 52.8	83.3 / 86.1
ARG [12]	ResNet-18	59.0 / 56.8	87.4 / 92.9
DIN [16]	ResNet-18	61.6 / 51.0	86.5 / 93.1
Ours	ResNet-18	**70.3 / 64.5**	**93.4 / 96.7**

Table 1. Comparision with state-of-the-art methods on the NBA and the Weakly supervised VOLLEYBALL dataset following metrics adopted in [14]. For the NBA dataset, we employ MCA (left) and MPCA (right) for evaluation. For the Weakly supervised VOLLEYBALL, we adopt MCA (left) and Merged MCA (right) for evaluation.

extract corresponding keypoint information. Due to the design of our method for fully supervised tasks, it proves challenging for us to conduct experiments on the NBA dataset. Inspired by [4], we situate a Transformer encoder on the convolutional neural network backbone. Employing the encoder on the feature map directs attention to entities involved in group activities, thereby circumventing the need for explicit object detection and pose estimation. Specifically, we define learnable tokens as inputs to the encoder. These tokens, guided by the attention mechanism, learn and localize the local context of group activities, encompassing both individuals and objects. Through these steps, we attain the required feature inputs (person features and object features), enabling the application of our method in unsupervised scenarios.

The experimental results on the NBA dataset are presented in Table 1. In comparison to methods such as ARG [12] and DIN [16], our approach achieves the optimal outcomes, exhibiting a nearly 10% lead in both MCA and MPCA metrics. Dual [3] and our Bi-Causal both exhibit commendable performance on the VOLLEYBALL dataset. However, a substantial disparity is observed between them on the NBA dataset, highlighting the exceptional efficacy of our method in the context of sports group activities.

Weakly supervised VOLLEYBALL dataset. To ensure a fair comparison with existing methods, we also validate our method on the Weakly Supervised setting of the VOLLEYBALL dataset. We adopt Multi-class Classification Accuracy (MCA) and Merged MCA for evaluation, Merged MCA means to merge the classes right set and right pass into right pass-set, and left set and left pass into left pass-set as in SAM [14] for a fair comparison.

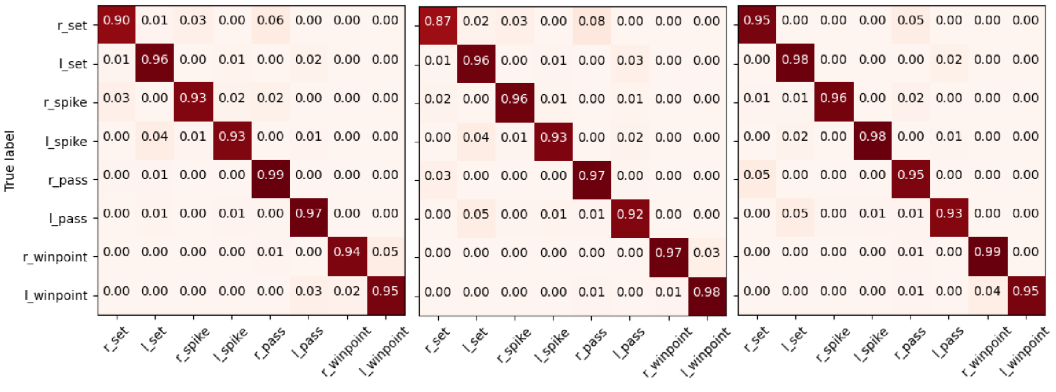
In the pursuit of equitable comparison, we refrained from utilizing bounding boxes and individual action label information within the dataset and adopted a processing approach akin to that of the NBA dataset. As indicated in Table 1, our method surpasses all other approaches in both MCA and Merged MCA on the VOLLEYBALL dataset, achieving the highest performance. It is worth mentioning that [14] and [1] previously pointed out the confusion in labeling the pass and set categories in the VOLLEYBALL dataset. Consequently, we merged these two categories and trained the model in a fully supervised manner, resulting in a final group activity recognition accuracy of **98.4%**.

In contrast to the NBA dataset, the human relations in the VOLLEYBALL dataset exhibit more fixed patterns. The causal relationships between human relations and human-object interactions are clearer, and there are more easily discernible interactions between individuals and objects. This facilitates our method in recognizing human-object interactions and modeling bidirectional causality, resulting in superior performance of Bi-Causal on this dataset.

3. Efficacy Analysis of Classification Results

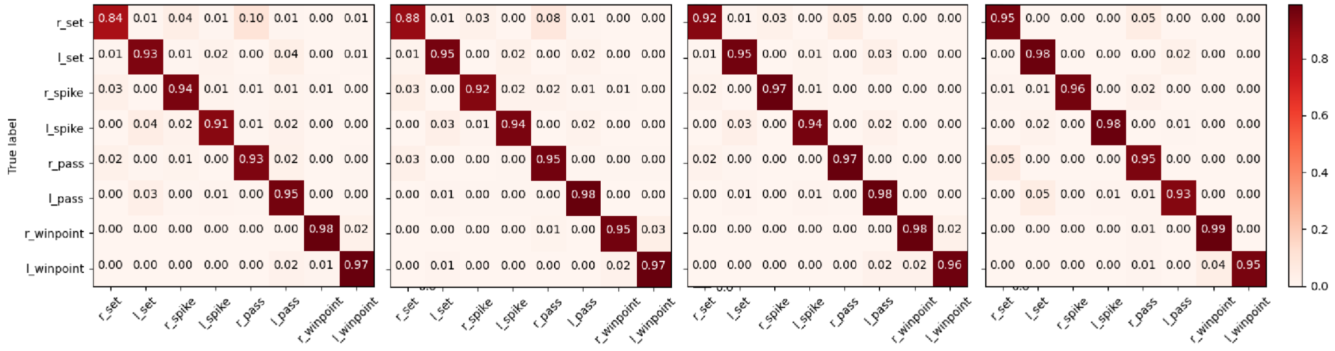
The confusion matrices in Fig 1 show the comparison result of our method with the COMPOSER [17] and GroupFormer [5]. One of the main advantages of our approach lies in the identification of two classes, pass and set. The recognition accuracy of COMPOSER and GroupFormer in the r-set class is 87% and 90% respectively, and the main source of their errors is the r-pass class. A similar situation still occurs in the l-set and l-pass classes. At the same time, our method performs relatively well with accuracy rates above 93% for all categories and 95% for the r-set class. Our Bi-Causal not only manifests superior performance but also demonstrates stability across various categories. In group activity pass and set, human actions and relations are more similar, but there are large differences in human-object interaction. During the passing activity, the interaction between individuals and objects is manifested by an upward wrist strike, aiming to prevent the ball from touching the ground. In contrast, within the setting activity, the interaction between individuals and objects is characterized by the palm supporting the ball, with the purpose of facilitating an opportunity for a spike. These differences can be captured by our Interaction Module and the bidirectional causality between human relations and human-object interactions can also make the Relation Module aware of these differences, making our Bi-Causal better distinguish the pass and set class.

As depicted in Figure 2, solely considering the interaction between individuals and objects yields suboptimal results. This is attributed to the fact that human-object interaction constitutes merely a fraction of group activities, neglecting the collaborative relationships and mutual influences among individuals, posing challenges in comprehending group activities. Similarly, exclusively focusing on interpersonal relationships makes it challenging for the model to pinpoint the essence of group activities, rendering group



(a) GroupFormer (b) COMPOSER (c) Bi-Causal

Figure 1. Comparison of confusion matrix of different methods. (a) Illustration of GroupFormer with RGB and keypoint modality. (b) Illustration of COMPOSER with keypoint modality. (c) Illustration of our proposed Bi-Causal with keypoint modality. The ordinate denotes the actual labels, while the abscissa signifies the predicted labels.



(a) Interaction (b) Relation (c) Fusion (d) Bi-Causal

Figure 2. Comparison of confusion matrix of different methods using human relations or human-object interactions. (a) Illustration of model based on human relations. (b) Illustration of model based on human-object interactions. (c) Illustration of relation and interaction fusion model. (d) Illustration of our bidirectional causality model.

behavior susceptible to the impact of similar member actions. However, a simplistic fusion of human relations and human-object interactions without considering their interdependence leads to their segregation, hindering the comprehensive utilization of vital information provided by both. Therefore, unifying the two through causal relationships allows for the full exploitation of the strengths of human relations and human-object interactions, offering more comprehensive support for the identification of group activities.

4. Feature Distributions

Fig 3 visualizes the feature distributions learned from the test set of the VOLLEYBALL dataset. Our method adds inter-class differences compared to the COMPOSER [17] and GroupFormer [5], making the classification boundaries more visible, especially between pass and set activity (both right and left). This also coincides with what we discussed in Section 3. Because of the similarities in the actions of people in the pass and set categories, it is difficult for relation-based methods to perceive the subtle differences. However, these two categories have large differences in human-object interaction, and these differences can be captured by our Interaction Module and causality communication channel, making our method better distinguish between these categories.

5. HOI features

Our objective for the IM is to extract HOI features from object features and human features. The HOI features are correctly retrieved both in terms of method and effect. In terms of the method, on the one hand, we utilize dot product and graph structure to calculate the association among entities, which is consistent with numerous HOI recognition methods [8, 11, 13]; on the other hand , we employ $\mathcal{L}_{person}$ and $\mathcal{L}_{inter}$ to constrain HOI features extraction with labels of human actions and group activities, rather than relying

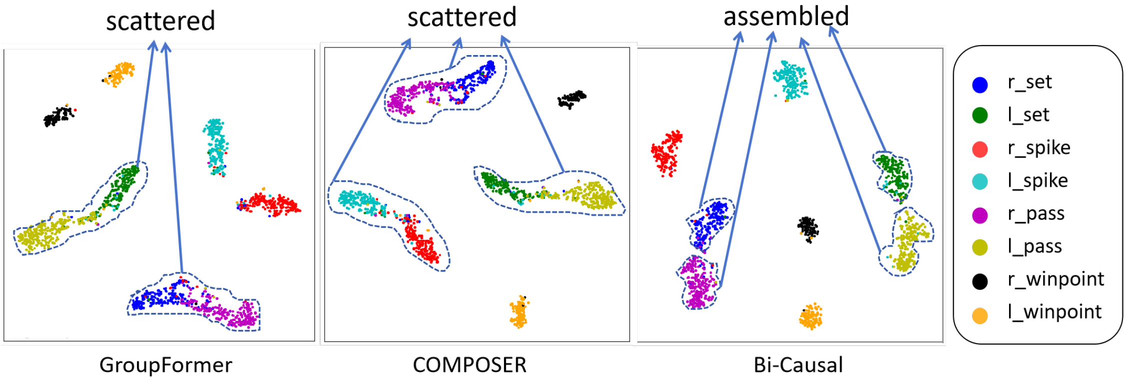


Figure 3. **t-SNE visualization of activity representation on the VOLLEYBALL dataset.**

on the identity, position, or category of individuals/objects. The human action labels in the datasets (such as spiking, and passing in volleyball data) encompass rich interaction information and can semantically replace the verb labels in HOI.

In terms of the effect, we adapt our IM to HOI recognition tasks (metrics are consistent with [8]). The results in table 2 below show the effectiveness of IM.

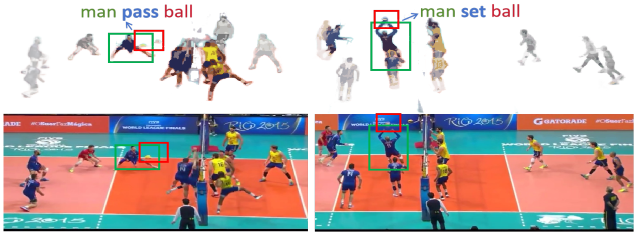


Figure 4. Interaction matrix visualization.

We also visualize the interaction matrix (edges) of IM in 4, with lighter colors indicating weaker attention. It shows that the module focuses more on parts with greater interaction.

6. Efficiency Analysis

In this section, we compare our Bi-Causal with the current SOTA methods based on execution efficiency. In addition to recognition accuracy, the recognition efficiency of a model is also an important metric for evaluating a model. This performance metric is inevitable if a model is to be truly usable. A model that uses only keypoint information has a greater advantage in recognition efficiency since fewer data and processing steps are used. Table 3 illustrates the efficiency analysis of the current state-of-the-art methods on the VOLLEYBALL dataset. The reported numbers exclude the parameters from the backbone and embedding layer to ensure comparability with prior work (e.g., Inception-v3). Our Bi-Causal demonstrates the best results with lower computational cost, requiring only 0.808 GFLOPs for a forward pass. In contrast to Groupformer, which utilizes additional RGB information and optical flow information, our method not only exhibits a precision improvement of 0.4% but also significantly reduces FLOPs from 10.99G to 0.808G. Both COMPOSER and our Bi-Causal take keypoint-only modality as input. Despite incurring 0.03 GFLOPs increase compared to COMPOSER, our method surpasses it by 1.5% in group activity classification accuracy.

7. Implementation details

The human keypoint information encompasses various types of information, including absolute and relative coordinates, absolute and relative velocities, normalized coordinates, and keypoint types in two-dimensional space. We concatenate these pieces of information and incorporate temporal and spatial embeddings, resulting in a feature dimensionality of D, which equals 256. To ensure a consistent evaluation, we adopt a comparable approach to related studies [2, 12, 15–17] by utilizing a fixed input size of $T = 10$ frames for training and testing on the VOLLEYBALL dataset and the COLLECTIVE ACTIVITY dataset. While poses are estimated for all 41 frames in the VOLLEYBALL dataset, not all frames are utilized in constructing the input. We employ an alternating sampling strategy, filtering 10 frames from the total of 41 frames for input. The number of spatial and temporal encoder layers in our relation module is set as 1. The dimension of the FFN layer in all Transformer encoders is set as 1024, and the non-linear activation function is ReLU. The dropout rate of the Transformer encoder at each scale is set as 0.3. We utilize HRnet [10] to obtain human keypoints following [2]. The keypoints we use have 17 different types, and the person number of the VOLLEYBALL dataset is 12 while the Collective Activity is 13. When using the VOLLEYBALL dataset, the object's keypoints annotations are from [6]. To reduce the problem caused by noisy estimated keypoints, we use the temporal

Method	MPHOI-72			CAD-120		
	F1@10	F1@25	F1@50	F1@10	F1@25	F1@50
Relational BiRNN	–	–	–	79.2 ± 2.5	75.2 ± 3.5	62.5 ± 5.5
ASSIGN [Ref4]	59.1 ± 12.1	51.0 ± 16.7	33.2 ± 14.0	88.0 ± 1.8	84.8 ± 3.0	73.8 ± 5.8
2G-GCN [Ref1]	68.6 ± 10.4	60.8 ± 10.3	45.2 ± 6.5	89.5 ± 1.6	87.1 ± 1.8	76.2 ± 2.8
Ours	68.5 ± 9.2	61.2 ± 10.4	43.2 ± 9.6	89.6 ± 2.1	87.2 ± 3.2	74.2 ± 4.1

Table 2. Effectiveness of IM in HOI recognition tasks.

Model	#Params ↓	FLOPs ↓	Accuracy ↑
ARG [12]	25.18M	5.44G	91.0
AT [2]	5.24M	1.26G	92.8
SACRF	29.42M	74.75G	93.3
Dual [3]	4.29M	2.81G	94.4
COMPOSER [17]	11.10M	0.777G	94.6
GroupFormer [5]	81.52M	10.99G	95.7
Ours	28M	0.808G	96.1

Table 3. Efficiency comparison with the state-of-the-art method on the VOLLEYBALL dataset in terms of FLOPs.

Object Keypoint Similarity (OKS) proposed in [9]. During the training process, we employ the Adam optimizer to update the network parameters, using a learning rate of 0.001. A weight decay of 0.002 is applied, and a batch size of 128 is utilized for all datasets. The network is implemented using PyTorch and trained for 80 epochs on a single NVIDIA Tesla V100 GPU with 16GB memory capacity.

8. Failure Case Analysis

In comparison to other methods, we observe improvements in the recognition accuracy across multiple categories, particularly in the l-set category, reaching 95%. This enhancement is attributed to our proposed causal model, which effectively leverages the bidirectional causality between human relations and human-object interactions. It adeptly captures distinctions between set and pass activities, and distinctions between set-pass and other categories. According to the confusion matrix of Bi-Causal in Figure 1, nearly all misclassifications in the set category result from categorizing it as pass, and the same pattern is observed for errors in the pass category. This indicates that Bi-Causal effectively distinguishes set and pass from other categories (such as spike). Misclassifications between set and pass might be attributed to potential label errors within the VOLLEYBALL dataset itself [1, 14] (confusion exists in the labeling of the pass and set categories). After merging the pass and set categories, our recognition accuracy reaches **98.4%**.

As shown in Figures 1 and 5, our method does not exhibit a prominent performance in the l-winpoint category. This is attributed to the transient nature of winning activities, often lacking interactions between individuals and objects. In such instances, group activities lack a unified tactical purpose, and the bidirectional causality between human relations and human-object interactions is unclear, resulting in a decline in model performance.

References

[1] Berker Demirel and Huseyin Ozkan. Decompl: Decompositional learning with attention pooling for group activity recognition from a single volleyball image. *arXiv:2303.06439*, 2023. 2, 5

[2] Kirill Gavrilyuk, Ryan Sanford, Mehrsan Javan, and Cees GM Snoek. Actor-transformers for group activity recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 839–848, 2020. 2, 4, 5

[3] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2990–2999, 2022. 2, 5

[4] Dongkeun Kim, Jinsung Lee, Minsu Cho, and Suha Kwak. Detector-free weakly supervised group activity recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 20083–20093, 2022. 2

[5] 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 *Proc. IEEE Int. Conf. Comput. Vis.*, pages 13668–13677, 2021. 2, 3, 5

[6] Mauricio Perez, Jun Liu, and Alex C Kot. Skeleton-based relational reasoning for group activity analysis. *Pattern Recognit.*, 122:108360, 2022. 4

[7] Rizard Renanda Adhi Pramono, Yie Tarng Chen, and Wen Hsien Fang. Empowering relational network by self-attention augmented conditional random fields for group activity recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 71–90. Springer, 2020. 2

[8] Tanqiu Qiao, Qianhui Men, Frederick WB Li, Yoshiki Kubotani, Shigeo Morishima, and Hubert PH Shum. Geometric features informed multi-person human-object interaction recognition in videos. In *European Conference on Computer Vision*, pages 474–491. Springer, 2022. 3, 4

[9] Michael Snower, Asim Kadav, Farley Lai, and Hans Peter Graf. 15 keypoints is all you need. In *Proceedings of*

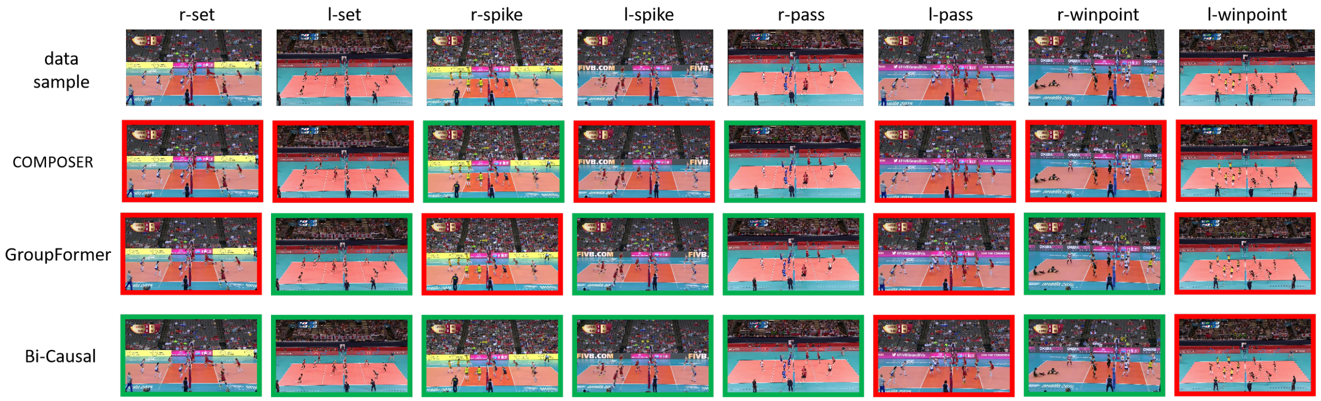


Figure 5. **Accuracy diagram for sample recognition of the VOLLEYBALL dataset. Comparing the predictions of the proposed method with other methods. Green boxes denote correct predictions, while red boxes indicate prediction errors.**

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6738–6748, 2020. 5

[10] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 5693–5703, 2019. 4

[11] Ning Wang, Guangming Zhu, Liang Zhang, Peiyi Shen, Hongsheng Li, and Cong Hua. Spatio-temporal interaction graph parsing networks for human-object interaction recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4985–4993, 2021. 3

[12] Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning actor relation graphs for group activity recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 9964–9974, 2019. 2, 4, 5

[13] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S Kankanhalli. Learning to detect human-object interactions with knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 3

[14] Rui Yan, Lingxi Xie, Jinhui Tang, Xiangbo Shu, and Qi Tian. Social adaptive module for weakly-supervised group activity recognition. In *Proc. Eur. Conf. Comput. Vis.*, pages 208–224, 2020. 1, 2, 5

[15] Hangjie Yuan and Dong Ni. Learning visual context for group activity recognition. In *Proc. AAAI Conf. Artif. Intell.*, pages 3261–3269, 2021. 4

[16] Hangjie Yuan, Dong Ni, and Mang Wang. Spatio-temporal dynamic inference network for group activity recognition. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 7476–7485, 2021. 2

[17] 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 *Proc. Eur. Conf. Comput. Vis.*, 2022. 2, 3, 4, 5