Language-Instructed Reasoning for Group Activity Detection via Multimodal Large Language Model

Jihua Peng[1], Qianxiong Xu[2], Yichen Liu[3], Chenxi Liu[2], Cheng Long[2], Rui Zhao[3], Ziyue Li[4]

[1]The Hong Kong Polytechnic University
[2]Nanyang Technological University
[3]SenseTime Research
[4]Technical University of Munich, Germany

Abstract

Group activity detection (GAD) aims to simultaneously identify group members and categorize their collective activities within video sequences. Existing deep learning-based methods develop specialized architectures (e.g., transformer networks) to model the dynamics of individual roles and semantic dependencies between individuals and groups. However, they rely solely on implicit pattern recognition from visual features and struggle with contextual reasoning and explainability. In this work, we propose LIR-GAD, a novel framework of language-instructed reasoning for GAD via Multimodal Large Language Model (MLLM). Our approach expand the original vocabulary of MLLM by introducing an activity-level `<ACT>` token and multiple cluster-specific `<GROUP>` tokens. We process video frames alongside two specially designed tokens and language instructions, which are then integrated into the MLLM. The pretrained commonsense knowledge embedded in the MLLM enables the `<ACT>` token and `<GROUP>` tokens to effectively capture the semantic information of collective activities and learn distinct representational features of different groups, respectively. Also, we introduce a multi-label classification loss to further enhance the `<ACT>` token's ability to learn discriminative semantic representations. Then, we design a Multimodal Dual-Alignment Fusion (MDAF) module that integrates MLLM's hidden embeddings corresponding to the designed tokens with visual features, significantly enhancing the performance of GAD. Both quantitative and qualitative experiments demonstrate the superior performance of our proposed method in GAD taks.

Introduction

Group activity detection (GAD) is a critical task in visual understanding that aims to simultaneously identify group members and classify their collective activities within dynamic scenes. Compared to group activity recognition (GAR) tasks (Li et al. 2021; Yuan and Ni 2021; Kim et al. 2022; Zhang et al. 2024), which rely on restrictive assumptions such as assigning a single activity label to an entire video clip and requiring actors to be pre-identified, GAD addresses the complex reality of crowded scenes by simultaneously localizing multiple groups and classifying each group's activity. Hence, the task has broad applications in
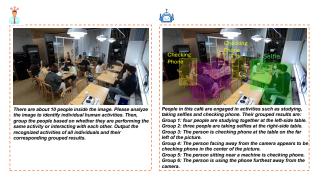
Figure 1: We present LIR-GAD, an end-to-end instruction-based group activity detection model that fully leverages the reasoning capabilities of MLLM, achieving accurate, interpretable, and explicit group localization, along with flexible activity representation, particularly for outlier actors who fall outside the predefined activity categories.

public safety monitoring, crowd behavior analysis, and intelligent video systems, where understanding group-level semantics beyond individual actions is essential.

Recent GAD approaches (Ehsanpour et al. 2020, 2022; Tamura, Vishwakarma, and Vennelakanti 2022; Kim et al. 2024) have developed specialized model architectures to address the inherent complexity of modeling interactions among multiple actors and distinguishing heterogeneous activity patterns across groups. Several approaches (Ehsanpour et al. 2020, 2022; Han et al. 2022b) employ graph neural networks (GNNs) to capture inter-individual relationships, followed by off-the-shelf clustering algorithms to group participants. HGC (Tamura, Vishwakarma, and Vennelakanti 2022) and GADFormer (Kim et al. 2024) leverage attention modules in transformers (Vaswani et al. 2017) to identify and then aggregate features relevant to social group activities, employing an end-to-end training framework. However, these methods still primarily depend on implicit pattern recognition from visual features, struggling with high-level contextual reasoning and semantic understanding. While these methods can identify group members and classify the activity of each group, they lack interpretability for these classification results and assign only predefined and rigid activity labels to groups and individual actors.

To address the aforementioned challenges of GAD, we propose LIR-GAD, a novel framework of language-instructed reasoning for GAD via Multimodal Large Language Model (MLLM) (e.g., LLaVA (Liu et al. 2023)). Our approach, LIR-GAD, leverages the remarkable reasoning and intention comprehension capabilities of MLLMs to facilitate instruction-based GAD task. Specifically, we expand the original vocabulary of MLLM by introducing an activity-level <ACT> token and multiple cluster-specific <GROUP> tokens. Then, we process video frames together with two specially designed tokens and language instructions, which are subsequently integrated into the MLLM. The pretrained commonsense knowledge within the MLLM enables the <ACT> token and <GROUP> tokens to effectively capture the semantic information of collective activities and learn distinct representational features of different groups, respectively. Moreover, the <GROUP> tokens can also capture semantic information of activities from the preceding <ACT> token through the autoregressive learning process of MLLM. Meanwhile, we extract fine-grained visual features of actors from the input video frames and introduce a learnable query vector to attend to these actor features, thereby obtaining initial group representations.

Thereafter, we introduce a Multimodal Dual-Alignment Fusion (MDAF) module which consists of two distinct cross-attention blocks followed by a feed-forward network (FFN) with residual connections. The first cross-attention block integrates the MLLM's hidden embedding of the <ACT> token with the visual features of actors, while the second aligns the <GROUP> token embeddings with the visual features of groups. The subsequent FFN acts as a feature fusion module, implicitly modeling the interactions between the enhanced group and actor representations. Through this simple yet effective design, the MDAF module ensures that both group and actor representations are enhanced with high-level textual and rich visual information, leading to more accurate and interpretable group activity detection.

Furthermore, to further enhance the <ACT> token's ability to learn discriminative semantic representations, we introduce a multi-label classification loss to supervise the learning of the <ACT> embedding. The <ACT> embedding is transformed from a high-dimensional space to a logit space, generating activity-specific logits that represent the probability of each activity occurring within a scene. The multi-label classification loss computes the discrepancy between the predicted logits for each activity class and a multi-hot encoded ground-truth label. By optimizing this loss, the model is trained to predict all activities occurring simultaneously within a scene, including those of outliers performing activities beyond predefined categories.

As shown in Figure 1, our method can achieve accurate, interpretable, and explicit group localization, along with flexible activity representation, particularly for outlier actors. Also, our method achieves state-of-the-art performance on both Café (Kim et al. 2024) and JRDB-Act (Ehsanpour et al. 2022) benchmarks. Our contributions are as follows:

- We leverage the reasoning capabilities of MLLM to simultaneously achieve group localization and per-group activity classification, facilitated by our proposed MDAF module, which effectively integrates textual and visual features.

- We propose a multi-label classification loss, enabling the MLLM to achieve more accurate detection of all concurrent activities in a scene, including those of outliers performing activities outside predefined categories.

- We present LIR-GAD by combining the above designs, demonstrating state-of-the-art performance on Café and JRDB-Act benchmarks. Moreover, LIR-GAD can achieve accurate, interpretable, and explicit group localization, along with flexible activity representation.

Related Work

In this section, we first review the related work on group activity detection (GAD), followed by an overview of multimodal LLMs (MLLMs) in visual understanding tasks.

Group Activity Detection. In computer vision, group activity detection (GAD) has emerged as a crucial advancement in the field of social scene understanding, addressing fundamental limitations of traditional group activity recognition (GAR) approaches (Yuan, Ni, and Wang 2021; Han et al. 2022a; Kim et al. 2022; Zhang et al. 2024). Current approaches of GAD can be broadly categorized into two methodological streams. The first approach (Ehsanpour et al. 2020, 2022; Han et al. 2022b) leverages graph neural networks (GNNs) to model interactions between individuals and subsequently partition them into groups using off-the-shelf clustering algorithms (Ng, Jordan, and Weiss 2001; Zelnik-Manor and Perona 2004). The second approach (Tamura, Vishwakarma, and Vennelakanti 2022; Kim et al. 2024) directly predicts the final set of groups and their activities through an end-to-end process, eliminating the need for the non-differentiable step such as clustering. HGC (Tamura, Vishwakarma, and Vennelakanti 2022) performs localization and group-to-member matching in the 2D coordinate space using Deformable DETR (Zhu et al. 2020), while GADFormer (Kim et al. 2024) conducts the matching in a semantic embedding space to exploit contextual and semantic clues by transformer networks (Vaswani et al. 2017). However, these methods rely solely on implicit pattern recognition from visual features, thereby exhibiting limitations in contextual reasoning and lacking inherent explainability.

Multimodal Large Language Model. Driven by the powerful reasoning abilities of LLMs, the research community has been advancing the development of multimodal LLMs (MLLMs) to extend these foundational capabilities into the visual domain. BLIP-2 (Li et al. 2023) encodes image features using a visual encoder, which are subsequently fed into the LLM alongside text embeddings. LLaVA (Liu et al. 2023) and MiniGPT-4 (Zhu et al. 2023) first conduct image-text feature alignment followed by instruction tuning. These pioneering works have significantly advanced subsequent visual understanding tasks. Recently, several studies have injected region-level image understanding and grounding capabilities into MLLMs. Kosmos-2 (Peng et al. 2023)
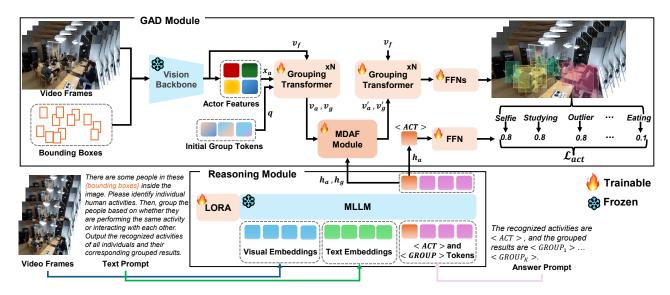
Figure 2: The overall framework of LIR-GAD. It mainly consists of three parts: (1) Reasoning Module. Given video frames and the instruction embedded with <ACT> and <GROUP> tokens, it leverages MLLM to comprehend the GAD task. (2) MDAF Module. It integrates the hidden states of the <ACT> and <GROUP> tokens from the MLLM with the corresponding visual features. (3) GAD Module. It primarily processes visual features and employs grouping transformers to perform group localization and recognize the activity categories of the groups. Our multi-label classification loss $\mathcal{L}_{act}$ is applied in this module.

achieves the visual grounding task by incorporating coordinate information into the training data, enabling MLLMs to understand spatial locations within images. Grounding-GPT (Li et al. 2024) perform fine-grained grounding tasks for image, video and audio. LISA (Lai et al. 2024) integrates a MLLM with the SAM (Kirillov et al. 2023) to achieve reasoning segmentation. It expand the original vocabulary with a special token to produce fine-grained segmentation masks. VideoLISA (Bai et al. 2024) generates temporally consistent segmentation masks in videos based on language instructions. VRS-HQ (Gong et al. 2025) design frame-level <SEG> and temporal-level <TAK> tokens that utilize MLLM's autoregressive learning to capture both local and global information, improving the quality of reasoning segmentation. SmartEdit (Huang et al. 2024) and InsightEdit (Xu et al. 2025) both leverage MLLMs to enhance their understanding and reasoning capabilities in instruction-based image editing. Inspired by the aforementioned works, we propose a novel framework to enhance the reasoning capabilities of MLLM for the GAD task.

Proposed Method

The overall framework of LIR-GAD is depicted in Figure 2. It mainly consists of a reasoning module, a MDAF module, and a GAD module. This section is organized as follows: we first introduce the first half of the GAD module, which focuses on extracting visual features from the input video frames. Then, we describe the reasoning module for leveraging MLLM to perceive and comprehend the GAD task. Next, we delve into the Multimodal Dual-Alignment Fusion (MDAF) module. Finally, we introduce the proposed multi-label classification loss for the <ACT> token.

GAD Module

Vision Backbone   Similar to recent GAD methods (Ehsanpour et al. 2022; Kim et al. 2024), we adopt an pretrained ResNet-18 (He et al. 2016) as the vision backbone to extract frame-level features v_f from the input video and use RoIAlign (He et al. 2017) with a 5×5 crop size to extract actor features x_a based on the given actor bounding boxes. For simplicity, we incorporate RoIAlign into the vision backbone, as shown in Figure 2.

Grouping Transformer   To deal with a varying number of groups in each video clip, we introduce a learnable query vector q as the initial group tokens, whose number K is supposed to be larger than the possible maximum number of groups in a clip. Then, we employ an N-layer grouping transformers (Kim et al. 2024), to process the initial group tokens q, actor features x_a, and frame features v_f, producing the first encoded group features v_g and actor features v_a. The mentioned process is represented as:

$$v_g, v_a = G_F(q, x_a, v_f) \qquad (1)$$

where G_F represents N-layer grouping transformers. The features v_g and v_a are concatenated and then fed into the MDAF module for enhancement.

Reasoning Module

In this module, we use LLaVA-Phi-3-V (Rasheed et al. 2024) as our vision-language foundation model. The MLLM receives the original video frames and the instruction to comprehend the GAD task. Specifically, we first expand the original vocabulary of MLLM by introducing an activity-level <ACT> token and multiple cluster-specific <GROUP>

tokens. The number of <GROUP> tokens is equal to the predefined number K of learnable group tokens in GAD module. As shown in Figure 2, the two types of tokens are embedded within a textual answer instruction. Also, we embed the bounding box coordinates of each frame into a corresponding textual question instruction and associate this instruction with its corresponding video frame. Then, the text instructions are tokenized as text embeddings c and the video frames are encoded into visual embeddings v_e. These embeddings v_e and c are fed into the LLaVA decoder to produce the hidden states of <ACT> and <GROUP> tokens:

$$h_a, h_g = LLaVA(v_e, c) \qquad (2)$$

where the hidden states h_a and h_g correspond to the <ACT> and <GROUP> tokens, respectively. The model is trained by minimizing the negative log-likelihood of predicting <ACT> token and K <GROUP> tokens, conditioned on the previously generated tokens. For the <ACT> token, the loss function can be represented below:

$$\mathcal{L}_{MLLM}(c) = -\sum \log p_{\{\theta \cup \omega_a\}}(\text{<ACT>} \mid v_e, c) \qquad (3)$$

where the trainable matrix ω_a denotes the <ACT> token embedding. The majority of parameters θ in the MLLM are kept frozen and we use LoRA (Hu et al. 2022) for efficient fine-tuning. The pretrained commonsense knowledge embedded in the MLLM enables the <ACT> token to effectively learn the semantic information of collective activities. For the <GROUP> tokens, the loss function can be formulated as:

$$\mathcal{L}_{MLLM}(c) = -\sum_{i=1}^{K} \log p_{\{\theta \cup \omega_g\}}(\text{<GROUP}_i\text{>} \mid v_e, c,$$
$$\text{<ACT>}, c, \text{<GROUP}_1\text{>}, \ldots, \text{<GROUP}_{i-1}\text{>}) \qquad (4)$$

where ω_g denotes the <GROUP> token embeddings. These <GROUP> tokens can learn distinct representational features of different groups from the knowledge of the MLLM. Particularly, the <GROUP> tokens can also capture semantic information of collective activities from the preceding <ACT> token via the autoregressive learning process.

MDAF Module

The textual features from the MLLM provide high-level semantic information, while the visual features from the GAD module offer fine-grained spatial and temporal details. To comprehensively integrate features from these two modalities, we design a Multimodal Dual-Alignment Fusion (MDAF) module, as shown in Figure 3.

The MDAF module receives the concatenated visual features of groups and actors from the first N-layer grouping transformers and disentangles them into separate group features v_g and actor features v_a. Then, the group features v_g serve as queries to interact with the textual features h_g of <GROUP> tokens, which act as both keys and values, through a cross-attention block. Similarly, the actor features v_a act as queries to attend to the textual features h_a of
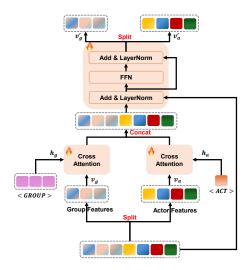

Figure 3: Overview of MDAF module. In this module, the group features and actor features are separately aligned with the textual features of <GROUP> tokens and <ACT> token via two separate cross-attention layers, then concatenated for effective feature fusion.

<ACT> tokens, also serving as keys and values, in a separate cross-attention block. The enhanced group and actor features are concatenated and fed into a feed-forward network (FFN) with residual connections for feature fusion. The mentioned process is represented as:

$$v'_a, v'_g = MDAF(v_a, v_g, h_a, h_g) \qquad (5)$$

where v'_a and v'_g denote the actor and group features after being processed by the MDAF module, respectively. They are fed into the subsequent N-layer grouping transformers which decode and refine the fused multimodal features, to perform final grouping and activity recognition tasks.

In the MDAF module, the two cross-attention blocks transfer the knowledge learned from the pre-trained MLLM into visual features, while the FFN serves as a feature fusion module that implicitly models interactions between group and actor features.

Multi-label Classification Loss

To supervise the learning of activity representations, we employ a multi-label classification loss. This loss function is specifically designed for the GAD scene where each input frame can be associated with multiple activity classes simultaneously. In our framework, this loss is applied to the hidden embedding of the <ACT> token, which are intended to capture the semantic information of co-occurring activities within a scene. For the Café dataset, which comprises 6 activity categories, we treat "Outlier" as an additional activity category to identify individual outlier activities. As shown in Figure 2, we employ a feed-forward network (FFN) as a regression head to project the <ACT> token embedding from a high-dimensional space into a 7-dimensional output space, corresponding to the 6 predefined activity categories and an additional "Outlier" class. This output space represents the

probability distribution over the possible activities. For instance, in the video frames shown in Figure 2, the model assigns a probability of 0.8 to "Studying", 0.8 to "Outlier", and 0.1 to "Eating". The loss $\mathcal{L}_{act}$ computes the discrepancy between the model's predicted logits for each activity class and a multi-hot encoded ground-truth label, where each element indicates the presence (1) or absence (0) of a specific activity. This loss is formulated as the binary cross-entropy across all n classes:

$$\mathcal{L}_{act} = -\frac{1}{n}\sum_{i=1}^{n}[y_i \cdot \log(\sigma(z_i)) + (1-y_i) \cdot \log(1-\sigma(z_i))] \quad (6)$$

where $y_i \in \{0,1\}$ is the ground-truth label for class i, z_i is the raw logit output by the model, and $\sigma(\cdot)$ is the sigmoid function. By minimizing $\mathcal{L}_{act}$, the model is explicitly trained to accurately predict the complete set of activity classes present in the visual input. This process ensures that the <ACT> token embedding can learn discriminative semantic representations, which are essential for downstream tasks such as group reasoning and activity recognition.

Moreover, we use additional feed-forward networks (FFNs) as regression heads for group activity classification and identifying group members. The overall loss function for our network is given as:

$$\mathcal{L} = \mathcal{L}_{ind} + \lambda_g \mathcal{L}_{group} + \lambda_m \mathcal{L}_{mem} + \lambda_c \mathcal{L}_{con} + \lambda_a \mathcal{L}_{act} \quad (7)$$

where $\mathcal{L}_{ind}$ is the individual action loss, $\mathcal{L}_{group}$ denotes the group activity classification loss, $\mathcal{L}_{mem}$ is the membership loss, and $\mathcal{L}_{con}$ is the group consistency loss. The loss functions described above are all derived from (Kim et al. 2024).

Experiments

Experimental Setting

Datasets   We evaluate our model on two public datasets: Café (Kim et al. 2024) and JRDB-Act (Ehsanpour et al. 2022). Café is a latest benchmark for group activity detection, designed with real-world coffee shop scenarios to address practical challenges like complex interactions. The dataset comprises over 4 hours of multi-view video footage captured across six different cafés, with four synchronized cameras recording diverse perspectives. It includes 3.5 million annotated human bounding boxes with track IDs, along with group membership labels and six distinct group activity categories (e.g., eating, fighting). Each video is segmented into 6-second clips, where actors are either assigned to a group performing a shared activity or labeled as outliers (non-group individuals). JRDB-Act is a large-scale benchmark for spatio-temporal action detection, social group identification, and collective activity recognition, captured in real-world environments using a mobile robot platform. It includes 11 pose-based, 3 human-human interaction and 12 human-object interaction action labels.

Evaluation Metrics   Following previous works (Ehsanpour et al. 2022; Kim et al. 2024; Han et al. 2022b), we evaluate our model on the Café dataset using two metrics: Group mAP and Outlier mIoU. Group mAP is calculated by averaging the interpolated average precision (AP) scores across all activity classes, where each AP is computed using the classification score of the true activity class as the detection confidence and Group IoU (Choi et al. 2014) as the localization criterion. Outlier mIoU measures the mean intersection over Union between predicted and ground-truth outliers, evaluating how accurately the model identifies non-group individuals in the scene. Also, we use G1 AP, G2 AP, G3 AP, G4 AP, G5$^+$ AP, and mAP as evaluation metrics on JRDB-Act dataset. G1 AP to G5$^+$ AP are metrics that report the average precision for groups containing exactly 1 to 5 or more members respectively, assessing a model's performance in detecting activities across different group sizes. The mAP is the mean AP across G1 AP to G5$^+$ AP.

Implementation Details   For the MLLM, we use LLaVA-Phi-3-V (Rasheed et al. 2024) as the base model based on Phi-3 (Research 2024) with 3.8B parameters. During training, the weights of LLaVA are frozen and we leverage LoRA (Hu et al. 2022) to perform efficient fine-tuning. The number of <GROUP> tokens is equal to the predefined number K of learnable group tokens in GAD module. Additionally, we adopt an pretrained ResNet-18 (He et al. 2016) as the vision backbone to extract frame-level features from the input video and use RoIAlign (He et al. 2017) with a 5×5 crop size to extract actor features based on the given actor bounding boxes. We completely freeze the vision backbone and RoIAlign. Our network employs two cascaded N-layer grouping transformers with 4 attention heads. The N is set to 3. The MDAF module is integrated between these two grouping transformers. The input frame number T is set to 5 and 2 for Café and JRDB-Act, respectively. We train our model for 20 epochs, using 4 NVIDIA 32G V100 GPUs with a distributed training script based on DeepSpeed (Rasley et al. 2020). We use the Adam (Kingma and Ba 2014) optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = $ 1e-8. The learning rate is initially set to 1e-5 with linear warmup to 1e-4 for 5 epochs, and linearly decayed for remaining epochs. The batch size per device is set to 4. The weights of the group activity classification loss λ_g and the membership loss λ_m are set to 2.0 and 5.0, respectively, and those of the group consistency loss λ_c and the multi-label classification loss λ_a are set to 2.0 and 2.0, respectively.

Comparison with State-of-the-art Methods

Results on Café   We compare our results with recent state-of-the-art (SOTA) methods on the Café dataset. For a fair comparison, we use ImageNet pretrained ResNet-18 as the vision backbone, and apply distance mask for all the methods including ours. Following previous work (Kim et al. 2024), our method is evaluated under two different dataset splits: *split by view* and *split by place*, with the number of group tokens set to 4 and 12, respectively. As shown in Table 1, our method achieves the best performance under both dataset splits, as measured by Group mAP$_{0.5}$ and Outlier mIoU. Under the *split-by-view* configuration with group tokens set to 12, our method significantly outperforms GAD-Former (Kim et al. 2024) by achieving a Group mAP$_{0.5}$ of 44.95. The significant performance gap (+7.42) demonstrates our approach's superior capability in group activity

Method	Venue	# Token (# Cluster)	Split by view			Split by place		
			Group mAP$_{1.0}$	Group mAP$_{0.5}$	Outlier mIoU	Group mAP$_{1.0}$	Group mAP$_{0.5}$	Outlier mIoU
ARG (Wu et al. 2019)	CVPR'19	4	11.03	34.50	56.61	6.87	28.44	46.72
		6	1.27	27.69	60.41	2.59	22.33	51.00
Joint (Ehsanpour et al. 2020)	ECCV'20	4	13.86	34.68	53.67	6.69	27.76	49.50
		6	5.94	33.14	60.63	5.11	24.55	56.94
JRDB-base (Ehsanpour et al. 2022)	CVPR'22	4	15.43	34.81	60.43	9.42	25.75	48.00
		6	6.77	35.22	63.85	6.37	26.23	51.53
HGC (Tamura, Vishwakarma, and Vennelakanti 2022)	ECCV'22	12	5.18	23.02	57.23	3.50	17.92	57.42
		50	6.55	26.29	56.84	3.47	18.46	52.56
GADFormer (Kim et al. 2024)	ECCV'24	4	16.02	40.22	64.06	8.97	27.33	62.35
		12	**18.84**	37.53	67.64	**10.85**	30.90	63.84
Ours	-	4	15.93	40.91	**69.34**	9.58	30.15	63.01
		12	16.64	**44.95**	68.98	9.51	**32.34**	**64.52**

Table 1: Quantitative comparison results with state-of-the-art methods on Café. All methods use the same ResNet-18 as the backbone. The second column indicates the number of tokens for transformer-based methods and the number of clusters for clustering-based methods. The subscripts of Group mAP denote Group IoU thresholds. The best results are highlighted in **bold**.

Method	Venue	Backbone	G1 AP	G2 AP	G3 AP	G4 AP	G5^{+} AP	mAP
SHGD (Li et al. 2022)	ECCV'22	Unipose (Artacho and Savakis 2020)	3.1	25.0	17.5	45.6	25.2	23.3
Joint (Ehsanpour et al. 2020)	ECCV'20	I3D (Carreira and Zisserman 2017)	8.0	29.3	37.5	65.4	**67.0**	41.4
PAR (Han et al. 2022b)	ECCV'22	Inception-v3 (Szegedy et al. 2016)	52.0	59.2	46.7	46.6	31.1	47.1
JRDB-base (Ehsanpour et al. 2022)	CVPR'22	I3D (Carreira and Zisserman 2017)	**81.4**	**64.8**	49.1	63.2	37.2	59.2
GADFormer (Kim et al. 2024)	ECCV'24	ResNet-18 (He et al. 2016)	70.1	56.3	50.4	71.7	50.8	59.8
Ours	-	ResNet-18 (He et al. 2016)	71.3	58.1	**52.0**	**73.0**	52.9	**61.4**

Table 2: Quantitative comparison results with state-of-the-art methods on JRDB-Act validation set.

Component				Group mAP$_{1.0}$	Group mAP$_{0.5}$	Outlier mIoU
Base	GROUP	ACT	$\mathcal{L}_{act}$			
✓				15.00	37.26	66.23
✓	✓			15.46	39.96	67.39
✓		✓		15.32	38.21	66.70
✓		✓	✓	15.65	39.01	67.15
✓	✓	✓		15.90	41.98	68.06
✓	✓	✓	✓	**16.64**	**44.95**	**68.98**

Table 3: Performance contribution of each component in our method. "Base" refers to using only the cascaded 6-layer grouping transformers. "GROUP" indicates that the <GROUP> embeddings are used to enhance the group features. "ACT" indicates that the <ACT> embedding is utilized to enhance the actor features.

Method	Group mAP$_{1.0}$	Group mAP$_{0.5}$	Outlier mIoU
CON1	14.28	38.34	67.77
CON2	14.62	42.25	66.15
SP1	15.51	40.21	68.46
SP2	**16.64**	**44.95**	**68.98**

Table 4: Ablation on different designs of the MDAF module. The designs of CON1, CON2, SP1 and SP2 are elaborated in the 'Impact of Different MDAF Designs' section.

detection. Moreover, we have achieved an Outlier mIoU of 69.34 with 4 group tokens, surpassing that of GADFormer by 5.28, which demonstrates the exceptional reasoning capabilities in recognizing the individual outliers.

Results on JRDB-Act Table 2 compares our results with those of SOTA methods on the JRDB-Act dataset. The number of group tokens K is set to 12 for JRDB-Act. Our method achieves 61.4 mAP with the same ResNet-18 backbone as GADFormer (Kim et al. 2024), outperforming all other methods. This demonstrates that our method can more effectively detect group activities across varying group sizes. Furthermore, these results highlight the generalization capability of our method.

Ablation Study

To verify the effectiveness of the proposed modules, we conducted ablation experiments using the "*split by view*" dataset partitioning strategy on the Café dataset. The number of group tokens K is set to 12. Due to page limitations, we have

discussed the impact of the placement of MDAF within the network and the impact of the number of <ACT> tokens in the supplementary material.

Effect of Each Component in Our Method Table 3 details the contribution of each component in our methods towards the overall performance. As shown, the introduction of <GROUP> embeddings, <ACT> embedding and the loss $\mathcal{L}_{act}$ individually contributes to improved performance across all three evaluation metrics. The best performance is achieved when all three components are integrated simultaneously, with the Group mAP$_{0.5}$ reaching 44.95.

Impact of Different MDAF Designs We analyze the impacts to performance for four different designs of MDAF module. CON1 denotes the concatenation of <ACT> and <GROUP> embeddings to enhance the concatenated visual features of groups and actors. CON2 denotes the concatenation of <ACT> and <GROUP> embeddings to separately enhance the visual features of groups and actors. SP1 employs the <ACT> embedding to enhance group features and <GROUP> embeddings for actor features, while SP2 implements the inverse strategy by applying the <ACT> embedding to actors and <GROUP> embeddings to groups. Table 4 shows the comparison results between the four designs, with the design SP2 achieving the best performance

There are about 10 people inside the image. Please analyze the image to identify individual human activities. Then, group the people based on whether they are performing the same activity or interacting with each other. Output the recognized activities of all individuals and their corresponding grouped results.

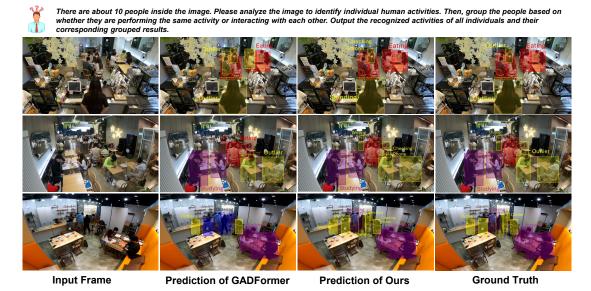| **Input Frame**	**Prediction of GADFormer**	**Prediction of Ours**	**Ground Truth**

Figure 4: Qualitative comparison on Café test-set. Our method achieves more accurate group localization and per-group activity classification through language-instructed reasoning capabilities.

There are about 10 people inside the image. Please analyze the image to identify individual human activities. Then, group the people based on whether they are performing the same activity or interacting with each other. Output the recognized activities of all individuals and their corresponding grouped results.

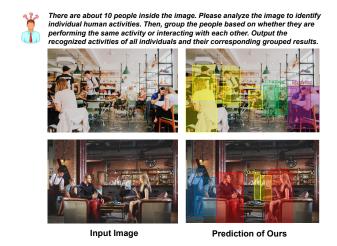| **Input Image**	**Prediction of Ours**

Figure 5: Visual results of our method on custom images.

across three metrics. The superior performance of SP2 over SP1 indicates that infusing high-level textual information from <ACT> embedding into actor features proves more beneficial for GAD, while enriching group features with <GROUP> embeddings contributes to learning more effective group representations. In contrast, the strategy of concatenating two embeddings for joint enhancement may introduce additional redundant information, leading to the suboptimal performance. Therefore, we have selected SP2 as the implementation strategy for MDAF.

Qualitative Analysis

In this section, we present visual results to validate the effectiveness of our proposed method.

Qualitative comparison on Café dataset. As shown in Figure 4, we provide a qualitative comparison between

our proposed method and GADFormer (Kim et al. 2024). Given a prompt containing the approximate number of people in the image, our method achieves more accurate group localization and per-group activity classification through language-instructed reasoning capabilities. Furthermore, our method can identify and predict activities for outliers who fall outside the predefined activity categories, demonstrating the flexibility of reasoning capabilities in GAD task.

Visualization results of our method on custom images. We evaluate the performance of our model on custom images to assess its generalization capability. Figure 5 demonstrates that our method can effectively achieve group localization and classify activities for each group in scenarios involving social group activities, while simultaneously identifying outliers through language-instructed reasoning.

Conclusion

In this paper, we propose LIR-GAD, a novel approach to language-instructed reasoning for group activity detection (GAD) via MLLM. Our method introduces an activity-level <ACT> token and multiple cluster-specific <GROUP> tokens, leveraging the MLLM's pretrained commonsense knowledge to effectively capture the semantic information of collective activities and learn distinct representational features of different groups, respectively. Our proposed MDAF module integrates the MLLM's hidden embeddings corresponding to the designed tokens with visual features, thereby significantly improving the performance of GAD. Also, our proposed multi-label classification loss enabling the model to achieve more accurate prediction of all concurrent activities in a scene, including those of outliers performing activities outside predefined categories. Both quantitative and qualitative experiments demonstrate the superior performance of our method in group activity detection.

References

Artacho, B.; and Savakis, A. 2020. Unipose: Unified human pose estimation in single images and videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7035–7044.

Bai, Z.; He, T.; Mei, H.; Wang, P.; Gao, Z.; Chen, J.; Zhang, Z.; and Shou, M. Z. 2024. One token to seg them all: Language instructed reasoning segmentation in videos. *Advances in Neural Information Processing Systems*, 37: 6833–6859.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Choi, W.; Chao, Y.-W.; Pantofaru, C.; and Savarese, S. 2014. Discovering groups of people in images. In *European conference on computer vision*, 417–433. Springer.

Ehsanpour, M.; Abedin, A.; Saleh, F.; Shi, J.; Reid, I.; and Rezatofighi, H. 2020. Joint learning of social groups, individuals action and sub-group activities in videos. In *European Conference on Computer Vision*, 177–195. Springer.

Ehsanpour, M.; Saleh, F.; Savarese, S.; Reid, I.; and Rezatofighi, H. 2022. Jrdb-act: A large-scale dataset for spatio-temporal action, social group and activity detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20983–20992.

Gong, S.; Zhuge, Y.; Zhang, L.; Yang, Z.; Zhang, P.; and Lu, H. 2025. The devil is in temporal token: High quality video reasoning segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 29183–29192.

Han, M.; Zhang, D. J.; Wang, Y.; Yan, R.; Yao, L.; Chang, X.; and Qiao, Y. 2022a. Dual-AI: Dual-path actor interaction learning for group activity recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2990–2999.

Han, R.; Yan, H.; Li, J.; Wang, S.; Feng, W.; and Wang, S. 2022b. Panoramic human activity recognition. In *European Conference on Computer Vision*, 244–261. Springer.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Huang, Y.; Xie, L.; Wang, X.; Yuan, Z.; Cun, X.; Ge, Y.; Zhou, J.; Dong, C.; Huang, R.; Zhang, R.; et al. 2024. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8362–8371.

Kim, D.; Lee, J.; Cho, M.; and Kwak, S. 2022. Detector-free weakly supervised group activity recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20083–20093.

Kim, D.; Song, Y.; Cho, M.; and Kwak, S. 2024. Towards more practical group activity detection: A new benchmark and model. In *European Conference on Computer Vision*, 240–258. Springer.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9579–9589.

Li, J.; Han, R.; Yan, H.; Qian, Z.; Feng, W.; and Wang, S. 2022. Self-supervised social relation representation for human group detection. In *European Conference on Computer Vision*, 142–159. Springer.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, S.; Cao, Q.; Liu, L.; Yang, K.; Liu, S.; Hou, J.; and Yi, S. 2021. Groupformer: Group activity recognition with clustered spatial-temporal transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13668–13677.

Li, Z.; Xu, Q.; Zhang, D.; Song, H.; Cai, Y.; Qi, Q.; Zhou, R.; Pan, J.; Li, Z.; Vu, V. T.; et al. 2024. Groundinggpt: Language enhanced multi-modal grounding model. *arXiv preprint arXiv:2401.06071*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Ng, A.; Jordan, M.; and Weiss, Y. 2001. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14.

Peng, Z.; Wang, W.; Dong, L.; Hao, Y.; Huang, S.; Ma, S.; and Wei, F. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2024. LLaVA++: extending visual capabilities with LLaMA-3 and Phi-3 (2024). *URL https://github. com/mbzuai-oryx/LLaVA-pp*.

Rasley, J.; Rajbhandari, S.; Ruwase, O.; and He, Y. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 3505–3506.

Research, M. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. Technical report, Microsoft. Accessed: 2024-07-20.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Tamura, M.; Vishwakarma, R.; and Vennelakanti, R. 2022. Hunting group clues with transformers for social group activity recognition. In *European Conference on Computer Vision*, 19–35. Springer.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wu, J.; Wang, L.; Wang, L.; Guo, J.; and Wu, G. 2019. Learning actor relation graphs for group activity recognition. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 9964–9974.

Xu, Y.; Kong, J.; Wang, J.; Pan, X.; Lin, B.; and Liu, Q. 2025. Insightedit: Towards better instruction following for image editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2694–2703.

Yuan, H.; and Ni, D. 2021. Learning visual context for group activity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 3261–3269.

Yuan, H.; Ni, D.; and Wang, M. 2021. Spatio-temporal dynamic inference network for group activity recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7476–7485.

Zelnik-Manor, L.; and Perona, P. 2004. Self-tuning spectral clustering. *Advances in neural information processing systems*, 17.

Zhang, Y.; Liu, W.; Xu, D.; Zhou, Z.; and Wang, Z. 2024. Bicausal: Group activity recognition via bidirectional causality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1450–1459.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.