JRDB-Social: A Multifaceted Robotic Dataset for Understanding of Context and Dynamics of Human Interactions Within Social Groups

Simindokht Jahangard, Zhixi Cai, Shiki Wen, Hamid Rezatofighi
Monash University
{simindokht.jahangard,zhixi.cai,hamid.rezatofighi}@monash.edu, swen0021@student.monash.edu
https://jrdb.erc.monash.edu/dataset/social

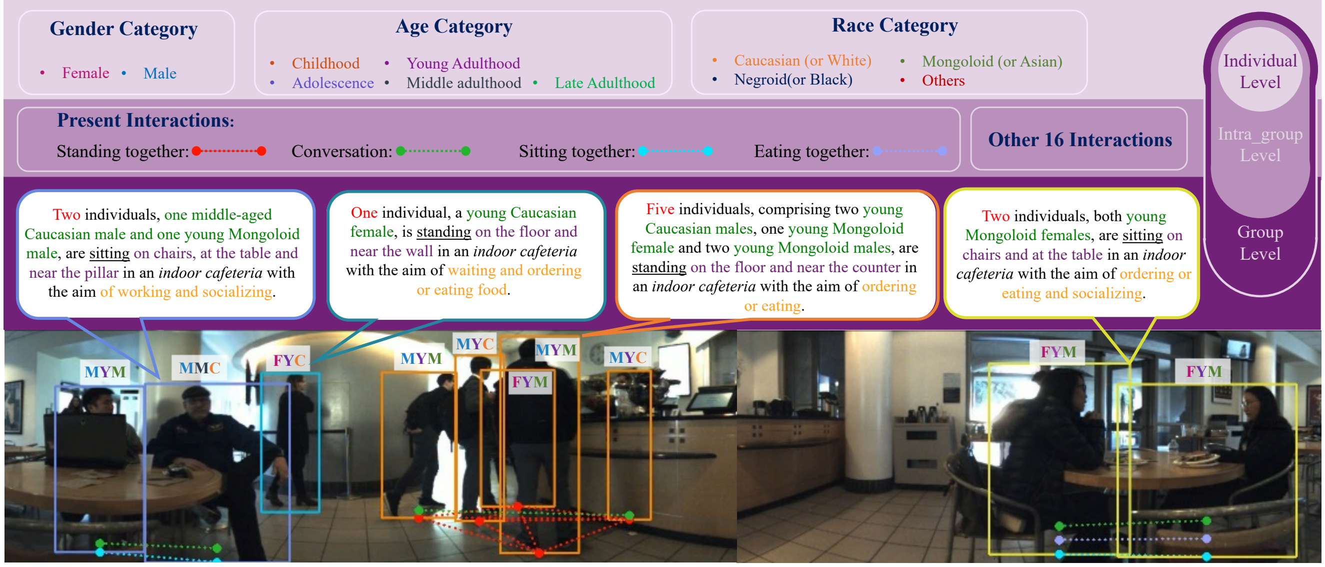
Gender Category
• Female • Male
Age Category
• Childhood • Young Adulthood
• Adolescence • Middle adulthood • Late Adulthood
Race Category
• Caucasian (or White) • Mongoloid (or Asian)
• Negroid(or Black) • Others
Individual Level
Intra_group Level
Group Level
Present Interactions:
Standing together:
Conversation:
Sitting together:
Eating together:
Other 16 Interactions
Two individuals, one middle-aged Caucasian male and one young Mongoloid male, are sitting on chairs, at the table and near the pillar in an indoor cafeteria with the aim of working and socializing.
One individual, a young Caucasian female, is standing on the floor and near the wall in an indoor cafeteria with the aim of waiting and ordering or eating food.
Five individuals, comprising two young Caucasian males, one young Mongoloid female and two young Mongoloid males, are standing on the floor and near the counter in an indoor cafeteria with the aim of ordering or eating.
Two individuals, both young Mongoloid females, are sitting on chairs and at the table in an indoor cafeteria with the aim of ordering or eating and socializing.
MYM
MMC
FYC
MYM
MYC
MYM
FYM
MYC
FYM
FYM

Figure 1. Some highlighted instances from the JRDB-Social dataset featuring detailed annotations across three levels: **Individual Level)** Representing specific attributes like age, gender, and race are shown through color-coded abbreviations. For example, 'MMC' represents Male, Middle Adulthood, Caucasian. **Intra-group Level)** This level focuses on group dynamics and interactions between each pair at the frame level, represented by dashed lines. **Group Level)** Each social group [1] is represented by the same colour and accompanied by textual descriptions that detail the number of members, their specific attributes, their body position's connection with the content, the presence of salient scene content near the group, the *venue*, and the group's aim or purpose.

Abstract

Understanding human social behaviour is crucial in computer vision and robotics. Micro-level observations like individual actions fall short, necessitating a comprehensive approach that considers individual behaviour, intra-group dynamics, and social group levels for a thorough understanding. To address dataset limitations, this paper introduces JRDB-Social, an extension of JRDB [2]. Designed to fill gaps in human understanding across diverse indoor and outdoor social contexts, JRDB-Social provides annotations at three levels: individual attributes, intra-group interactions, and social group context. This dataset aims to enhance our grasp of human social dynamics for robotic applications. Utilizing the recent cutting-edge multi-modal large language models, we evaluated our benchmark to explore their capacity to decipher social human behaviour.

1. Introduction

Human social behaviour understanding finds numerous applications in computer vision and robotics. Simply observing the micro-level information like the actions of an individual is inadequate for a comprehensive understanding of human behaviour because humans are inherently social beings and require analysis within a broader social context. Therefore, a comprehensive and multi-layered approach is required to perceive human social behaviour thoroughly. For example, in security and surveillance systems, integrating individual-level data, identifying social groups, and taking context into account significantly enhance the overall capacity to better understand crowd behaviors [3]. Additionally, this integration fosters more natural and intuitive experiences in human-robot interaction like telerobots [4], coworker robots [5] and social robots [6].

In recent years, significant progress has been made in vision-based understanding of human behaviour and activity, furnishing datasets at different levels. Some datasets focused on individual-level information such as human attributes and atomic actions [7–14]. Conversely, other datasets primarily concentrate on human-human interactions [9–14]. On a higher level, certain datasets provide information regarding human groups and video captioning, describing various events occurring in videos [15–22]. While serving as valuable resources for the research community, these datasets mainly consider one aspect of this multi-level hierarchy in understanding human behaviour and activity and fall short in adequately capturing and reflecting the complexity of dynamics and context inherent in human social behaviors within crowded scenes. To bridge this gap, we introduce JRDB-Social an extension of the JRDB dataset [2]. JRDB features a social manipulator robot with stereo RGB 360° cameras, dual LiDAR sensors for 3D point clouds, audio, GPS, and over 1.8 million annotations in the form of 2D bounding boxes and 3D oriented cuboids. The JRDB dataset has already contained very useful annotations such as human atomic actions and social grouping [17] and human body pose annotations [23]. Our proposed annotations serve as a perfect complement to enrich this popular dataset. JRDB-Social is structured at three distinct levels including: **individual level**, **intra-group level** and the **social group level**. Firstly, at the individual level, we provide annotations for gender, age, and race. Secondly, at the intra-group level, we capture fine-grained, dynamic multi-interactions between (20 categories) each pair within a subgroup at the frame level. Lastly, at the social group level, we incorporate text captions that describe information about the context including the connection between the group's body position and the content, the presence of salient scene content situated in close proximity to the group, the specific location or venue, and the group's aim and purpose, thus offering a holistic contextual overview. This benchmark facilitates exploration into how demographic factors influence social behaviour, allowing for examination of differences in interactions based on gender or race. Venue annotations provide contextual information for interactions, recognizing that behaviours and social dynamics in settings like cafeterias may differ from those in formal environments such as classrooms. Understanding the purpose of a group can illuminate the motivation behind the interaction, whether the group gathers for leisure, work, or education. Ultimately, this benchmark seeks to narrow gaps in comprehending human behavior within social settings, furnishing valuable insights to enrich our understanding of social dynamics.

With the surge in popularity and significant advancements in large language models (LLMs) and vision-language models (VLMs) [24–27], which claim proficiency in visual understanding and reasoning, we explore and assess their capabilities using our dataset. We applied these models to our dataset to evaluate their effectiveness in perceiving and reasoning about human social behavior in crowded environments. Our evaluation focuses on examining and discussing the strengths and limitations of current methodologies in understanding human social and contextual interaction dynamics.

In sum, the key contributions of this work are as follows:

- Providing JRDB-Social benchmark on dynamic human-human interactions at the frame level, revealing multi-label annotations between each pair within a group.
- We offer individual attribute annotation and descriptions of social groups. These descriptions elaborate on the relationship between the group's body position and the content, the presence of salient scene context near the group, the venue location, and the group's aim or purpose.
- We assess the performance of the most recent vision-language models within the framework of JRDB-Social, performing a comprehensive examination to identify the advantages and shortcomings of current approaches.

2. Related works

2.1. Datasets

In the following section, we provide the commonly used public datasets across three distinct levels *i.e.* individual, intra-group, and social group level.

Individual Level. Analysing individual-level human behaviour, which encompasses factors like age, gender, and race, alongside detailed atomic action data, is paramount across diverse domains. The MovieGraph dataset [28] specializes in delineating inferred properties of human-centric situations through intricate, graph-based annotations of social scenarios depicted in movie clips. Also, recently, autonomous vehicle datasets like [29, 30] have been released featuring individual-level annotations comprehending the behaviour of various age groups and genders in traffic scenarios. Conversely, certain datasets, such as [31–34], focus on atomic actions by offering comprehensive data that specifically highlights individual actions within their content. Shifting focus, other datasets delve into emotions [35], providing additional layers of information to understand human behaviour by considering variables such as age, gender, and ethnicity. While valuable, existing datasets lack a perspective from the robot within a social environment and they are not from human crowded environments. JRDB-Social addresses this gap by providing demographic information in real-world data from the robot's perspective.

Intra-Group Level Interactions. Some image-based datasets focus intensely on specific interactions such as [7–10, 36, 37]. Also, some video-based including [11–15, 32] offer a diverse range of interaction scenarios, contributing to understanding of human interaction dynamics in various contexts. The drawback of these datasets lies in their limited number of label categories or the treatment of interaction labels as a subset. Moreover, they often involve interactions between only two or very few individuals, lacking representation of crowd dynamics. JRDB-Social offers frame-level multi-label annotation of human interactions within social groups in crowded scenes.

Social Group Level. A more comprehensive understanding of human behaviour emerges when contextual information is available. In this context, certain datasets provide higher-level information such as [15–17] furnish valuable insights into social group dynamics. On the other hand, some datasets such as [18–22] that primarily focus on video captioning, offer sets of descriptions for multiple events occurring in videos and aim to temporally localize them. However, they often overlook crucial, detailed information—especially pertaining to how individuals interact with each other and their surroundings. JRDB-Social offers comprehensive information by providing group-level details such as the group's body position related to the content, the proximity of salient scene content within the group, the group's objective, and key information about the main environment. This approach enhances human understanding by presenting a more holistic view of the scenario.

2.2. Vision-based Large Language Models

In recent years, Large Language Models (LLMs) [38–41] have made significant strides in achieving multi-modal capabilities. Notable models include Video-LLaMA [24], which enhances LLMs for detailed video comprehension, and NExT-GPT [25], a holistic multi-modal model navigating text, images, videos, and audio seamlessly. Other models like VideoChat [27], Visual ChatGPT [42], VALLEY [43], Otter [44], ViperGPT [45], and MiniGPT-4 [46] contribute to advancements in video understanding, visual processing, and instruction tuning for improved contextual learning. Additionally, efforts such as InstructBLIP [47], M^{3}IT [48], and VisionLLM [49] focus on instruction tuning, multilingual datasets, and vision-centric tasks, collectively propelling AI systems towards greater versatility in language understanding and nuanced video comprehension. While these models excel in understanding and reasoning over videos, their capacity to comprehend human social behaviour and conduct contextual activity analysis remains unexplored. This paper aims to assess their performance on the JRDB-Social dataset.

3. The JRDB-Social Dataset

We developed JRDB-Social to complement the current annotation of JRDB dataset [2] by providing new annotations to better comprehend human activity in a social context. JRDB dataset contains 64 minutes of sensory data, comprising 54 sequences reflecting diverse indoor and outdoor locations within a university campus environment. The JRDB dataset has been captured by a social manipulator robot featuring stereo RGB 360° cameras, dual LiDAR sensors for 3D point clouds, audio, GPS, and boasts over 1.8 million annotations in the form of 2D bounding boxes and 3D oriented cuboids. The JRDB dataset already contains valuable annotations, such as human atomic actions, social grouping [17], and human body pose annotations [23] and JRDB-Social serves as a complementary extension to this dataset, providing a multifaceted perspective at three levels: the individual, intra-group, and the social group level.

Annotation Process. For annotating JRDB-Social, at each level, we designed a toolbox, featuring unique IDs corresponding to existing 2D and 3D bounding box annotations. We adhered to a quality assessment protocol aligned with established benchmarks known for high-quality annotated data, such as previous JRDB benchmarks [2, 17, 23]. In line with these benchmarks, we implemented a standardised data annotation process to ensure consistency with past JRDB annotations [2, 17]; for instance, our interaction annotations align seamlessly with the actions of each individual involved. Also, our annotators, chosen for their expertise in behaviour analysis, adhere to strict guidelines and protocols for standardized annotation. Encountering challenges such as significant distance from the robot, varying lighting conditions, occlusion, and crowded scenes, each label in our dataset is accompanied by difficulty level—categorized as *Easy* (1), *Medium* (2), or *Hard* (3)—reflecting the annotator's confidence. To ensure fairness and consistency, labels undergo a thorough review by two additional individuals, alongside random quality assessments by multiple assessors.

Text Description Structure. We enhance JRDB-Social by including text descriptions for each group to offer contextual understanding. This aligns with the trend of combining natural language understanding with computer vision, benefiting tasks like image captioning. This enhancement also has potential in Human-Robot interaction, helping robots adjust behaviour based on group context, thus improving interactions. We construct our sentences in the colour-coded format, shown in the yellow box below.

3.1. Individual Level Attributes

The JRDB-Social dataset includes individual attributes, as understanding these is crucial for studying diverse social behaviours in groups and deepening insights into human behaviour in social situations especially in social sciences and psychology research. Additionally, in human-robot interactions, awareness of individuals' demographics aids in personalizing the robot's behaviour for more culturally sensitive interactions. Therefore, in addition to the currently available annotations of human body pose and atomic ac-

Text Description Structure:[number of individuals], including attribute of each person involved (e.g., Person 1:[age, gender, race], Person 2:[age, gender, race], Person 3:[age, gender, race], and so on). These individuals engage in activities on [the content relates to group's body position and the presence of a salient scene content nearby] in *[a specific venue location]* with the purpose of [group's goal].

tion in [17, 23], we annotated gender, age, and race in this dataset. Under the gender category, the dataset distinctly classifies individuals into two primary groups: *Male* and *Female*. The age attribute is finely segmented into five distinct groups: *Childhood* (3-12 years), *Adolescence* (13-20 years), *Young Adulthood* (21-40 years), *Middle Adulthood* (41-65 years), and *Late Adulthood* (66 years and above). In terms of racial classification, the dataset adopts Alfred L. Kroeber's classification[1] which is based on physical characteristics. It includes *Caucasian/White* (light skin, varied eye colours), *Negroid/Black* (dark skin, coiled hair), *Mongoloid/Asian* (almond-shaped eyes, black hair, varied skin tones). Figure 3 illustrates attribute distributions within the JRDB-Social dataset excluding impossible ones. As illustrated, male individuals predominate in the gender category. The video, primarily captured in a university environment, predominantly features individuals in the young adulthood category, reflecting distribution of this category in real-life situations. The racial breakdown shows equal representation from Caucasian and Asian populations, with a smaller proportion representing the Black community. Figure 1 shows some samples.

3.2. Intra-Group Level Dynamic Interactions

The concept of multi-label interaction at the frame level provides a detailed understanding of social dynamics within social groups [17], offering detailed insights into simultaneous actions and gestures among individuals. These fine-grained annotations are instrumental in training machine learning models for the recognition of diverse social interactions, especially in social navigation scenarios. Additionally, the frame-level annotations facilitate behavioural studies, allowing researchers to examine in-depth the temporal dynamics of interactions and how individuals engage with each other in specific social settings. In JRDB-Social, we provided multi-label fine-grained interaction annotation at the frame level and categorized it into three distinct groups, each encompassing various dimensions of shared experiences. The first category, shown in purple in Figure 2, focuses on shared physical activities, including behaviours with physical proximity and posture. The second, in dark purple, involves joint engagement with external entities, often centred around interacting with objects together. The third, in light pink, encompasses interpersonal exchanges and gestures as part of social interactions. The distribution of dynamic interaction classes for both training and test sets is depicted in Figure 2. The vertical axis, presented on a log scale, represents the number of frames. Notably, prevalent interactions include walking, standing, sitting together, and engaging in conversations and less frequent activities like pointing at something together and shaking hands well reflect distribution biases in real-world daily scenarios. Additionally, the accompanying pie chart illustrates difficulty levels, with medium difficulty comprising the largest portion at 36.64%, followed by hard at 34.2%, and easy at 29.2%, indicating an even distribution of difficulty. During the annotation process, interactions between individuals within each group are meticulously annotated. We identify the individuals participating, document the frame range of the interaction, and to improve accuracy, integrate the individual actions outlined in [17], aligning them with the corresponding interaction. More details about our protocols are provided in supplementary materials.

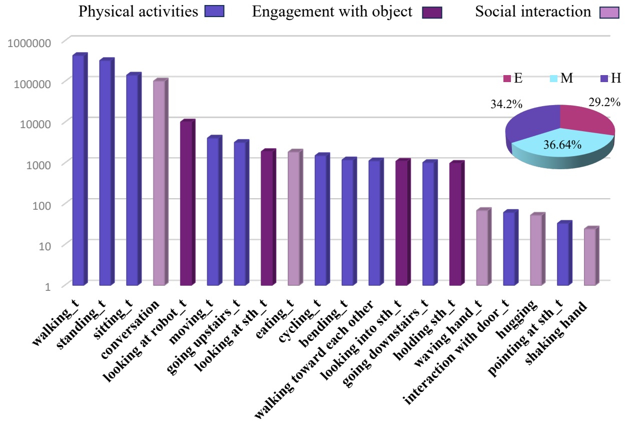

Figure 2. Sorting interaction classes on a log-scale distribution, displaying descending frame numbers for all data. Difficulty levels indicated as E (Easy), M (Medium), and H (Hard).

3.3. Social Group Level Context

These annotations aim to provide a comprehensive understanding of social behaviour at the social group level. By

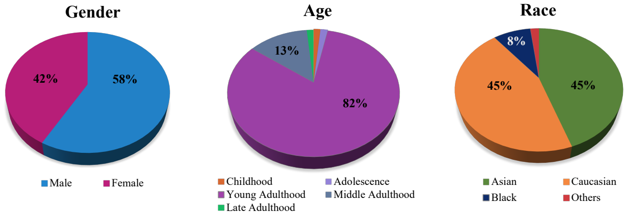

Figure 3. Statistics of individual attributes.

[1] https://en.m.wikipedia.org/wiki/Mongoloid

including information beyond individual attributes and interactions, the dataset becomes richer and more reflective of real-world scenarios involving groups of people and context. It encompasses the details about the group's surrounding environment, their specific venue and aims and purposes. Figure 4 illustrates a word cloud depicting labels for each category. Further statistical details for each category can be found in the supplementary materials.

Engagement of Body Position with the Content and Salient Scene Content. These annotations contribute to a contextual analysis of the physical engagement of the group and add layers of context to the dataset. This involves the examination of body position related to the content (BPC), considering the majority of group members. Additionally, it offers valuable insights into the presence of salient scene elements (SSC) in their surroundings. To elaborate, the BPC encompasses specifics regarding how body position is linked to the content. For instance, sitting on a *chair* or standing on the *platform*. On the other hand, SSC provides information about the presence of dominant scene content near the group. This includes observations like standing near a *pillar* or *counter*. As the location of the group may vary, we annotate this information at the frame level.

The Venue Location. This annotation offers information about the locations where the group participates in activities, helping in modelling and predicting the movement patterns of individuals and groups. This is essential for robots to navigate through diverse environments, adapting their behaviour based on the spatial context. These are classified into indoor spaces like *cafeterias, dining halls, or food courts*, *open spaces or corridors*, *rooms or classrooms*, and *study areas*. Furthermore, it includes outdoor categories such as *open areas or campuses* and *streets*.

Group's Aims and Purposes. These annotations provide information at the social group level about the purpose behind the formation and activities of each group, the dataset becomes a valuable resource for advancing research in social understanding, behavioural analysis, and contextual reasoning. Our categorization provides information from the act of moving through spaces, utilizing corridors in *navigating*, to routine travel in *commuting*, and aimless strolls in *wandering*, the categories capture various facets of human interaction. *Socializing* emphasizes communal connections, while *studying, writing, reading, and working* highlighting focused intellectual activities. *Discussing an object or a matter* centres on engaging conversations around specific topics, and *attending class, lecture, or seminar* underscores educational gatherings. *Ordering and eating food* portrays communal aspects of meal-related activities, and *excursion* adds a recreational dimension to the group's aim. Moreover, *Waiting for someone or something* demonstrates the anticipation and patience associated with awaiting a person or an event. In essence, this categorization offers nuanced insights into the multifaceted dynamics of collective human behaviour in diverse contexts. .

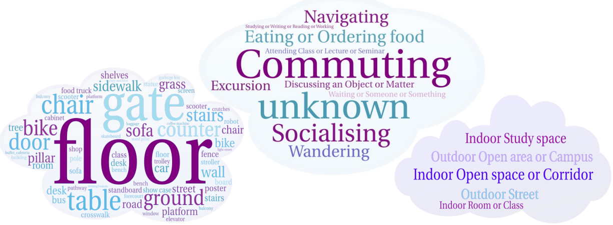

Figure 4. Social group level word cloud in the dataset. Left: location of body posture and objects. Top: group aim. Right: venue locations. Larger words indicate higher frequency.

4. Experiments

In this section, we delve into the recent advancements of large language models, particularly their progress in vision-related aspects and multi-modal capabilities. Our objective is to explore the effectiveness of state-of-the-art multimodal Language Models (LLMs) using the JRDB-Social benchmark. Our focus is to assess their ability to comprehend various complexities of human social behaviour across different difficulty levels and conditions. Specifically, we aim to evaluate their performance at individual, intra-group, and social group levels.

Multi-modal LLMs Selection. For our evaluation, we opted for prominent and well-established multi-modal models that have exhibited promising results in recent studies. This selection includes video-based models like Video-LLaMA [24], VALLEY [43], and Otter [44]. Additionally, our analysis incorporates image-based models, such as MINIGPT-4 [46] and InstructBLIP [47]. This diverse set ensures a comprehensive examination of the current state-of-the-art in multi-modal language models.

Metric and Evaluation. For evaluating these models based on textual descriptions in JRDB-Social, common metrics like BLEU [50], ROUGE [51], and METEOR [52] are often used to measure overall sentence similarity. However, these metrics may lack specificity when the focus is on key entities such as gender, age, aims, venues, etc., embedded in the hard-coded sentence structure. For instance, BLEU and ROUGE lack precision by concentrating on n-gram overlap without considering individual term precision, while METEOR, despite incorporating additional linguistic features, is sensitive to parameter choices. To sidestep these limitations arising from these metric limitations, we opt to assess the models by prompting questions to extract named entities, such as coloured words in the text description structure, reflecting crucial elements of meaning. We then reformulate the problem as a single or multi-label classification task. This approach aligns with the unique demands of our task, providing a focused and rigorous evaluation frame-

Multi-modal LLM	Individual Level			Intra-Group Level	Social Group Level				Overall
	Gender	Age	Race	Interactions	BPC	SSC	Venue	Purpose	Average
Video-LLaMA (LLaMA-2 13B) [24]	0.7139	0.3069	**0.2837**	0.3253	0.1639	0.1252	**0.2413**	0.2595	**0.3025**
Video-LLaMA (LLaMA-2 7B) [24]	0.5200	**0.3196**	0.2308	**0.3852**	0.1642	0.1639	0.2147	**0.3003**	0.2874
Valley (LLaMA-1 13B) [43]	0.3991	0.2041	0.1253	0.1674	0.0364	0.0456	0.0904	0.2632	0.1603
Valley (LLaMA-2 7B) [43]	0.4658	0.1731	0.0905	0.2035	0.1115	0.0559	0.0695	0.2515	0.1400
OTTER (LLaMA-1 7B) [44]	0.1959	0.1131	0.0115	0.2761	0.0799	0.1242	0.0420	0.0411	0.1105
MiniGPT-4 (LLaMA-2 7B) [46]	**0.7391**	0.2204	0.2604	0.2068	**0.1970**	0.0978	0.2574	0.2736	0.2816
InstructBLIP (Vicuna-V1 13B) [47]	0.5860	0.2482	0.1875	0.0665	0.0639	**0.2354**	0.1636	0.1841	0.2169
InstructBLIP (Vicuna-V1 7B) [47]	0.6444	0.0697	0.2587	0.0937	0.1337	0.1174	0.2020	0.1880	0.2135

Table 1. **Guided Perception** Experiment: Comparing popular multi-modal LLMs across the JRDB-Social in F1 score for all sets. Optimal results in bold, second best underlined. BPC = Engagement of Body Position's connection with the Content, SSC = Salient Scene Content.

Multi-modal LLM	Individual Level			Intra-Group Level	Social Group Level				Overall
	Gender	Age	Race	Interactions	BPC	SSC	Venue	Purpose	Average
Video-LLaMA (LLaMA-2 13B) [24]	**0.3338**	**0.2543**	**0.3507**	0.2786	0.0795	0.0238	0.2471	0.1792	**0.1965**
Video-LLaMA (LLaMA-2 7B) [24]	0.2177	0.2256	0.2984	**0.2970**	0.0637	0.0195	**0.2705**	0.1375	0.1645
Valley (LLaMA-2 7B) [43]	0.0215	0.0579	0.0122	0.0104	0.0025	0.0008	0.0449	0.0211	0.0217
MiniGPT-4 (LLaMA-2 7B) [46]	0.2344	0.2109	0.2619	0.0994	**0.0829**	0.0282	0.2432	**0.1861**	0.1684
InstructBLIP (Vicuna-V1 13B) [47]	0.0856	0.0856	0.0346	0.0542	0.0172	0.0267	0.1119	0.0778	0.0643
InstructBLIP (Vicuna-V1 7B) [47]	0.1111	0.1478	0.0686	0.0841	0.0314	**0.0338**	0.1457	0.0821	0.0881

Table 2. **Holistic (Counting) Experiment**: Comparing popular multi-modal LLMs across the JRDB-Social in F1 score for all sets. Optimal results in bold, second best underlined. BPC = Engagement of Body Position's connection with the Content, SSC = Salient Scene Content.

work that addresses the shortcomings of more generic textual metrics. Also, for evaluating interaction labels, we apply the same metrics. To assess the selected models' performance, we use accuracy and F1 score as metrics. While accuracy measures overall correctness, the F1 score provides a balanced evaluation of precision and recall, particularly valuable in imbalanced scenarios, as observed in the JRDB-Social dataset. While the F1 score results for entire data are outlined here, more comprehensive details and accuracy results are provided in the supplementary materials.

Experimental Setup and Implementation Details. Generally, we conducted two separate experiments, **Guided Perception** and **Holistic**, to investigate how multi-modal LLMs perform under different difficulty conditions and levels of guidance. In the guided perception experiment, we use ground truth bounding boxes to direct the model's focus to specific video regions, providing clear cues for analyzing areas of interest. In the holistic study, the model is exposed to the entire video without external aids like bounding boxes. This methodology allows the model to conduct a thorough analysis of the video, relying solely on its inherent information, mimicking real-world scenarios where detailed annotations might be lacking. Figure 6 shows this study on three levels, and more detail is provided in section 4.1 and section 4.2.

To enhance both reliability and performance, we implemented a *Five Ensemble Strategy*. In this strategy, each model undergoes five iterations, and the final output is derived through the utilization of an aggregation strategy. Further details regarding its implementation for both video-based and image-based models can be found in the supplementary materials. Additionally, in our guided perception experiment for social group analysis, we explored different *cropping scales* to identify the most effective cropping region. Unlike individual or intra-group levels, the model needed to account for a broader context beyond mere bounding boxes. This approach ensured the model's capability to encompass diverse contextual information and maintain robustness across different scenarios, adeptly adapting to scenes featuring both small and large groups. Figure 5 displays the various scales using different methods that utilize MiniGPT-4 (model LLaMA-2 7B). Frame-level processing involves cropping videos based on bounding boxes for each frame and resizing them uniformly to 512 × 512 pixels or 256 × 256 pixels. In the fixed black mask method, videos are cropped with non-object areas masked in black. The object's centre point is retained without resizing across frames. The fixed without mask method is akin to the fixed black mask method, but it maintains the full context without using black masking on non-object areas. Considering the overall F1 average, it was observed that the Frame-level method, with an F1 average of 0.1452 and a scaling factor of 2.5, outperformed both the Fixed-Black Mask and Fixed W/O Mask methods. Consequently, the Frame-level method is selected. More details have been provided in the supplementary material.

4.1. Guided Perception

In this experiment, we employ ground truth bounding boxes to crop regions of interest. The objective of this approach is to aid the model in localization, directing its attention to specific regions and evaluating its capability to detect the

Multi-modal LLM	Individual Level			Intra-Group Level	Social Group Level				Overall
	Gender	Age	Race	Interactions	BPC	SSC	Venue	Purpose	Average
Video-LLaMA (LLaMA-2 13B) [24]	**0.9800**	0.5657	0.7458	0.2786	0.2326	0.0771	0.2814	0.4788	0.4622
Video-LLaMA (LLaMA-2 7B) [24]	0.9800	0.5633	0.7482	0.3213	0.2177	0.0730	0.2810	0.4663	0.4564
Valley (LLaMA-2 7B) [43]	0.6958	0.4415	0.1629	0.2602	0.2298	0.0637	0.3350	0.4415	0.3288
OTTER (LLaMA-1 7B) [44]	0.8194	0.5290	0.5687	0.3053	0.2796	0.0913	**0.4309**	0.3198	0.4542
MiniGPT-4 (LLaMA-2 7B) [46]	0.8194	0.5290	0.5687	0.2796	0.0913	0.0282	**0.4309**	0.3198	0.4180
InstructBLIP (Vicuna-V1 13B) [47]	0.8493	**0.6223**	**0.7663**	**0.3318**	**0.4045**	**0.1026**	0.3197	**0.4797**	0.4846
InstructBLIP (Vicuna-V1 7B) [47]	0.7621	0.5408	0.6450	0.3081	0.2947	0.0711	0.3313	0.4326	**0.4848**

Table 3. **Holistic (Binary) Experiment**: Comparing popular multi-modal LLMs across the JRDB-Social in F1 score for all sets. Optimal results in bold, second best underlined. BPC = Engagement of Body Position's connection with the Content, SSC = Salient Scene Content.

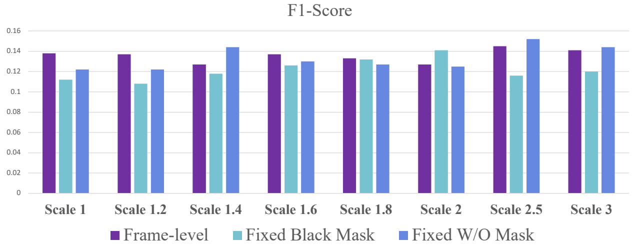

Figure 5. Exploring diverse cropping scales with MiniGPT-4 at the group level in F1 score.

category at three distinct levels. For example, on an individual level, we query the models for the type of gender, age, and race within specific areas of interest for each person. At the intra-group level, we isolate each pair within a group for the entire duration of their interaction. The model's role in this context is to observe the interaction type and discern singular or multiple interactions taking place among these pairs. At the social group level, the model is presented with each isolated group throughout its entirety. Its task involves recognizing the engagement of body position's connection to the content (BPC), identifying the proximity of significant scene context (SSC), determining the venue where the group is active, and comprehending the group's purpose. These prompts and processes are illustrated in Figure 6.

Based on the results presented in Table 1, the analysis reveals a consistently reliable performance in predicting individual attributes such as gender and age. However, the detection of race proves to be more intricate, primarily due to the subjective and complex nature of this attribute. Notably, Video-LLaMA and MiniGPT-4 stand out as the top-performing models, attributed to the quality of the data on which they were trained and their design framework. These models exhibit promising results, particularly in tasks related to gender and age prediction. Nevertheless, even these models, experience a decline in performance as the evaluation progresses from the individual to intra-group and social group levels. This observed pattern signifies a significant challenge for the models in comprehending higher-level social contexts. The intricacies associated with attributes like body position's connection with the content (BPC) and salient scene context (SSC) contribute to the limitations faced by these models, underscoring the ongoing need for advancements to enhance their understanding of diverse and complex social dynamics beyond individual attributes. In this experiment, we explored the model's capabilities by focusing on a limited region. However, the vital question remains: can the model effectively capture intricate details when presented with the entire scene? To answer this query, we delve into a holistic experiment, explained below.

4.2. Holistic

In this experiment, the model receives the complete video. The objective is to evaluate how well multi-modal LLMs capture fine-grained information without any cropping or additional assistance and explore their performance across three distinct levels. To this end, we employed two approaches: the counting approach and the binary approach.

Counting Approach. In this approach, our central objective is to evaluate the model's ability to identify detailed information and quantify occurrences throughout an entire video. For example, on an individual level, we analyzed the count of females or individuals in young adulthood. At the intra-group level, we inquired about the frequency of diverse interactions between pairs. One instance is the exploration of the number of pairs sitting together in a video, and this investigative process was reiterated for all interaction labels. Similarly, this methodology is replicated at the social group level. Figure 6 visually depicts this approach. As indicated in Table 2, the experiment demonstrates a significant decline in performance compared to the guided perception experiment, Table 1. Valley (LLaMA-1 13B) and OTTER (LLaMA-1 7B) models are excluded from Table 1 due to poor performance. This suggests that capturing information at this level of detail poses a substantial challenge for these models. Moreover, the difficulty increases when transitioning to a higher-level social group context, similar to the guided perception experiment. This observation prompted us to simplify the task by assessing whether the model can perceive fine-grained information or not. To explore this, we conducted a binary approach experiment.

Binary Approach. As previously mentioned, this approach aims to evaluate the models' ability to capture intricate de-

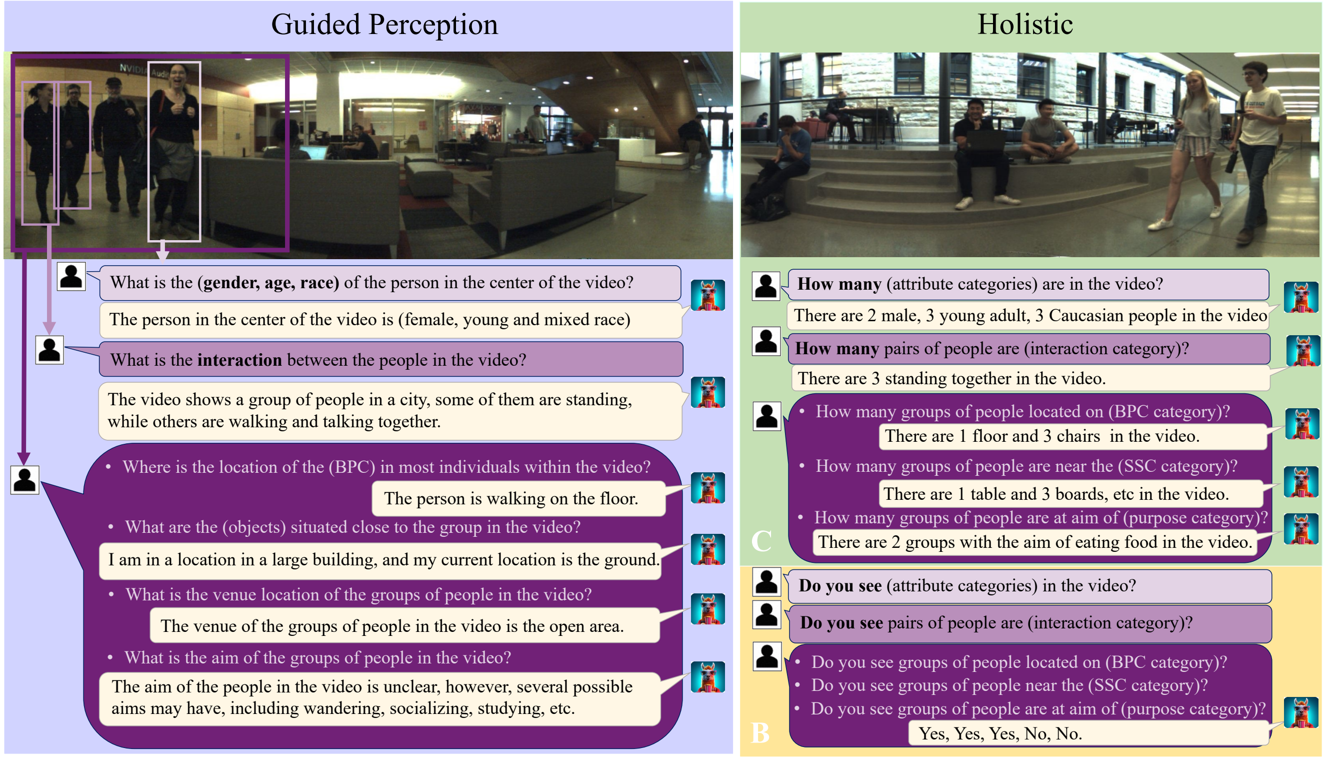
Guided Perception
Holistic
What is the (gender, age, race) of the person in the center of the video?
The person in the center of the video is (female, young and mixed race)
What is the interaction between the people in the video?
The video shows a group of people in a city, some of them are standing, while others are walking and talking together.
Where is the location of the (BPC) in most individuals within the video?
The person is walking on the floor.
What are the (objects) situated close to the group in the video?
I am in a location in a large building, and my current location is the ground.
What is the venue location of the groups of people in the video?
The venue of the groups of people in the video is the open area.
What is the aim of the groups of people in the video?
The aim of the people in the video is unclear, however, several possible aims may have, including wandering, socializing, studying, etc.
How many (attribute categories) are in the video?
There are 2 male, 3 young adult, 3 Caucasian people in the video
How many pairs of people are (interaction category)?
There are 3 standing together in the video.
How many groups of people located on (BPC category)?
There are 1 floor and 3 chairs in the video.
How many groups of people are near the (SSC category)?
There are 1 table and 3 boards, etc in the video.
How many groups of people are at aim of (purpose category)?
There are 2 groups with the aim of eating food in the video.
C
Do you see (attribute categories) in the video?
Do you see pairs of people are (interaction category)?
Do you see groups of people located on (BPC category)?
Do you see groups of people near the (SSC category)?
Do you see groups of people are at aim of (purpose category)?
Yes, Yes, Yes, No, No.
B

Figure 6. Illustrating the Guided Perception experiment is depicted through cropped regions delineated by bounding boxes on the left image. The colours—light pink, dark pink, and purple—signify the individual, intra-group, and social group levels, respectively, as detailed in Figure 1's colour legend. Holistic experiments are denoted by a green background for the Counting approach (C) and a yellow background for the Binary approach (B).

tails without specifying their type and quantity. The purpose of this assessment is to examine their level of understanding achieved by simplifying the task. Similar to the counting approach, the entire video is presented to the model, and the query is simplified to a binary response, either *Yes* or *No*. This process is visually depicted in Figure 6. Examining the outcomes of this experiment in Table 3, despite not altering the input of the model (entire video), we observed enhanced performance compared to the counting approach, Table 2. The improvement may stem from the hallucination problem [24, 44] present in the text encoder of these models, as the video encoder in this aspect works similarly to the counting approach. However, even with the simplicity of this approach, the models still encounter challenges in capturing information at the intra-group and social context levels. This implies that social group contexts pose challenges for multi-modal LLMs, and improvements in various aspects, such as training on more challenging datasets that offer finer-grained information, and developing a more effective framework, are required.

5. Conclusion

This paper introduces JRDB-Social, a comprehensive robotic dataset designed to investigate human social behaviour within varied contexts. Annotations within the dataset operate across three levels: individual, intra-group, and group, providing detailed attributes, interactions, and contextual descriptions. Leveraging recent advancements in VLMs, the dataset was assessed to gauge their proficiency in understanding human social behaviour in crowded environments. However, findings suggest that VLMs encounter challenges in meaningful visual perception and reasoning on this dataset, particularly in tasks involving complex social interactions. The observed weaknesses may stem from design choices or differences in training data. Thus, there is a need for further advancements in these models to enhance their capability to capture nuanced social understanding within diverse contexts.

Acknowledgments. The work has received partial funding from The Australian Research Council Discovery Project ARC DP2020102427. Additionally, it is based on research partially sponsored by the DARPA Assured Neuro Symbolic Learning and Reasoning (ANSR) program under award number FA8750-23-2-1016 and the DARPA Computational Cultural Understanding (CCU) program under agreement number HR001122C0029.

References

[1] Mahsa Ehsanpour, Alireza Abedin, Fatemeh Saleh, Javen Shi, Ian Reid, and Hamid Rezatofighi. Joint learning of social groups, individuals action and sub-group activities in videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 177–195. Springer, 2020. 1

[2] Roberto Martin-Martin, Mihir Patel, Hamid Rezatofighi, Abhijeet Shenoi, JunYoung Gwak, Eric Frankel, Amir Sadeghian, and Silvio Savarese. Jrdb: A dataset and benchmark of egocentric robot visual perception of humans in built environments. *IEEE transactions on pattern analysis and machine intelligence*, 2021. 1, 2, 3

[3] Shaogang Gong, Chen Change Loy, and Tao Xiang. Security and surveillance. *Visual analysis of humans: Looking at people*, pages 455–472, 2011. 1

[4] Hrishav Bakul Barua, Chayan Sarkar, Achanna Anil Kumar, Arpan Pal, et al. I can attend a meeting too! towards a human-like telepresence avatar robot to attend meeting on your behalf. *arXiv preprint arXiv:2006.15647*, 2020. 1

[5] Marc Hanheide, Denise Hebesberger, and Tomáš Krajník. The when, where, and how: An adaptive robotic infoterminal for care home residents. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 341–349, 2017. 2

[6] Deirdre E Logan, Cynthia Breazeal, Matthew S Goodwin, Sooyeon Jeong, Brianna O'Connell, Duncan Smith-Freedman, James Heathers, and Peter Weinstock. Social robots for hospitalized children. *Pediatrics*, 144(1), 2019. 2

[7] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5534–5542, 2016. 2

[8] Gokhan Tanisik, Cemil Zalluhoglu, and Nazli Ikizler-Cinbis. Facial descriptors for human interaction recognition in still images. *Pattern Recognition Letters*, 73:44–51, 2016.

[9] Gokhan Tanisik, Cemil Zalluhoglu, and Nazli Ikizler-Cinbis. Multi-stream pose convolutional neural networks for human interaction recognition in images. *Signal Processing: Image Communication*, 95:116265, 2021. 2

[10] Matteo Ruggero Ronchi and Pietro Perona. Describing common human visual actions in images. *arXiv preprint arXiv:1506.02203*, 2015. 2

[11] Michael S Ryoo and JK Aggarwal. Ut-interaction dataset, icpr contest on semantic description of human activities (sdha). In *IEEE International Conference on Pattern Recognition Workshops*, volume 2, page 4, 2010. 2

[12] Dong-Gyu Lee and Seong-Whan Lee. Human interaction recognition framework based on interacting body part attention. *Pattern Recognition*, 128:108645, 2022.

[13] Alonso Patron-Perez, Marcin Marszalek, Andrew Zisserman, and Ian Reid. High five: Recognising human interactions in tv shows. In *BMVC*, volume 1, page 33, 2010.

[14] Marcin Marszalek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2929–2936. IEEE, 2009. 2

[15] Xavier Alameda-Pineda, Jacopo Staiano, Ramanathan Subramanian, Ligia Batrinca, Elisa Ricci, Bruno Lepri, Oswald Lanz, and Nicu Sebe. Salsa: A novel dataset for multimodal group behavior analysis. *IEEE transactions on pattern analysis and machine intelligence*, 38(8):1707–1720, 2015. 2, 3

[16] Xueyang Wang, Xiya Zhang, Yinheng Zhu, Yuchen Guo, Xiaoyun Yuan, Liuyu Xiang, Zerun Wang, Guiguang Ding, David Brady, Qionghai Dai, et al. Panda: A gigapixel-level human-centric video dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3268–3278, 2020.

[17] Mahsa Ehsanpour, Fatemeh Saleh, Silvio Savarese, Ian Reid, and Hamid Rezatofighi. 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*, pages 20983–20992, 2022. 2, 3, 4

[18] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 3

[19] Gabriel Huang, Bo Pang, Zhenhai Zhu, Clara Rivera, and Radu Soricut. Multimodal pretraining for dense video captioning. In *AACL-IJCNLP 2020*, 2020.

[20] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[21] Ganchao Tan, Daqing Liu, Meng Wang, and Zheng-Jun Zha. Learning to discretely compose reasoning module networks for video captioning. *arXiv preprint arXiv:2007.09049*, 2020.

[22] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591, 2019. 2, 3

[23] Edward Vendrow, Duy Tho Le, Jianfei Cai, and Hamid Rezatofighi. Jrdb-pose: A large-scale dataset for multi-person pose estimation and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4811–4820, 2023. 2, 3, 4

[24] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 2, 3, 5, 6, 7, 8, 13, 14, 15, 16, 17, 18

[25] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023. 3

[26] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

[27] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao.

Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 2, 3

[28] Paul Vicol, Makarand Tapaswi, Lluis Castrejon, and Sanja Fidler. Moviegraphs: Towards understanding human-centric situations from videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8581–8590, 2018. 2

[29] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 206–213, 2017. 2

[30] Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John K Tsotsos. Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6262–6271, 2019. 2

[31] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6047–6056, 2018. 2

[32] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 28–35. IEEE, 2012. 2

[33] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

[34] AJ Piergiovanni and Michael S. Ryoo. Learning shared multimodal embeddings with unpaired data. *arXiv preprint arXiv:1806.08251*, 2018. 2

[35] Yu Luo, Jianbo Ye, Reginald B Adams, Jia Li, Michelle G Newman, and James Z Wang. Arbee: Towards automated recognition of bodily expression of emotion in the wild. *International journal of computer vision*, 128:1–25, 2020. 2

[36] Yuqing Cui, Apoorv Khandelwal, Yoav Artzi, Noah Snavely, and Hadar Averbuch-Elor. Who's waldo? linking people across text and images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1374–1384, 2021. 2

[37] Astrid Orcesi, Romaric Audigier, Fritz Poka Toukam, and Bertrand Luvison. Detecting human-to-human-or-object (h 2 o) interactions with diabolo. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE, 2021. 2

[38] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 3

[39] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[40] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

[41] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 3

[42] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023. 3

[43] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*, 2023. 3, 5, 6, 7, 13, 14, 15, 16, 17, 18

[44] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 3, 5, 6, 7, 8, 13, 14, 15, 16, 17, 18

[45] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*, 2023. 3

[46] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3, 5, 6, 7, 13, 14, 15, 16, 17, 18

[47] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 3, 5, 6, 7, 13, 14, 15, 16, 17, 18

[48] Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. M^{3}it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*, 2023. 3

[49] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*, 2023. 3

[50] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 5

[51] Lin Chin-Yew. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out, 2004*, 2004. 5

[52] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 5