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

Supplementary Material

1. Dataset

1.1. Intra-Group Level Dynamic Interactions

As mentioned in the paper, we presented a protocol to our trained annotators for labeling each interaction. The protocol for annotating each interaction label is outlined below:

- *Walking Together:* Individuals walking in the same direction with close proximity, either alongside each other or with one person behind the other.
- *Walking Toward Each Other:* Individuals walking towards each other, or one person standing while the other approaches.
- *Standing Together:* Both individuals standing closely at the same time.
- *Moving Together:* One person walking while the other engages in alternative modes such as skating, cycling, or riding a scooter.
- *Sitting Together:* Both individuals sitting simultaneously.
- *Going Upstairs Together:* Both individuals ascending stairs together at the same time.
- *Cycling Together:* Both individuals cycling either alongside each other or in tandem.
- *Going Downstairs Together:* Both individuals descending stairs together at the same time.
- *Bending Together:* Both individuals bending simultaneously.
- *Pointing at Something Together:* Both individuals pointing at something together simultaneously.
- *Skating Together:* Both individuals skating concurrently.
- *Scootering Together:* Both individuals riding scooters together simultaneously.
- *Conversation:* One individual listening while the other talks, or vice versa.
- *Looking Into Something Together:* Both individuals looking into something together simultaneously.
- *Looking at the Robot:* Both individuals directing their gaze at the robot simultaneously.
- *Looking at Something Together:* Both individuals looking at something together at the same time.
- *Eating Something Together:* Both individuals consuming food simultaneously.
- *Interaction with Door Together:* Both individuals interacting with a door together at the same time.
- *Waving Hand:* One or both individuals perform a waving hand gesture, indicating a greeting as specified in individual action labels.
- *Shaking Hand:* The individuals shake hands with each other, expressing a greeting as indicated in individual action labels.
- *Hugging Each Other:* The individuals embrace each other, conveying a greeting as specified in individual action labels.
- *Running Together:* Both individuals running together at the same time.
- *Holding Something Together:* Both individuals holding something together at the same time.

These annotations provide a detailed understanding of various dyadic interactions, offering valuable insights.

1.2. Social Group Level Context

Engagement of Body Position with the Content and Salient Scene Content. In this section, we discussed the association between body positions and content (BPC) and the identification of salient scene content in proximity to a group (SSC). The category of annotations for both BPC and SSC are detailed below:

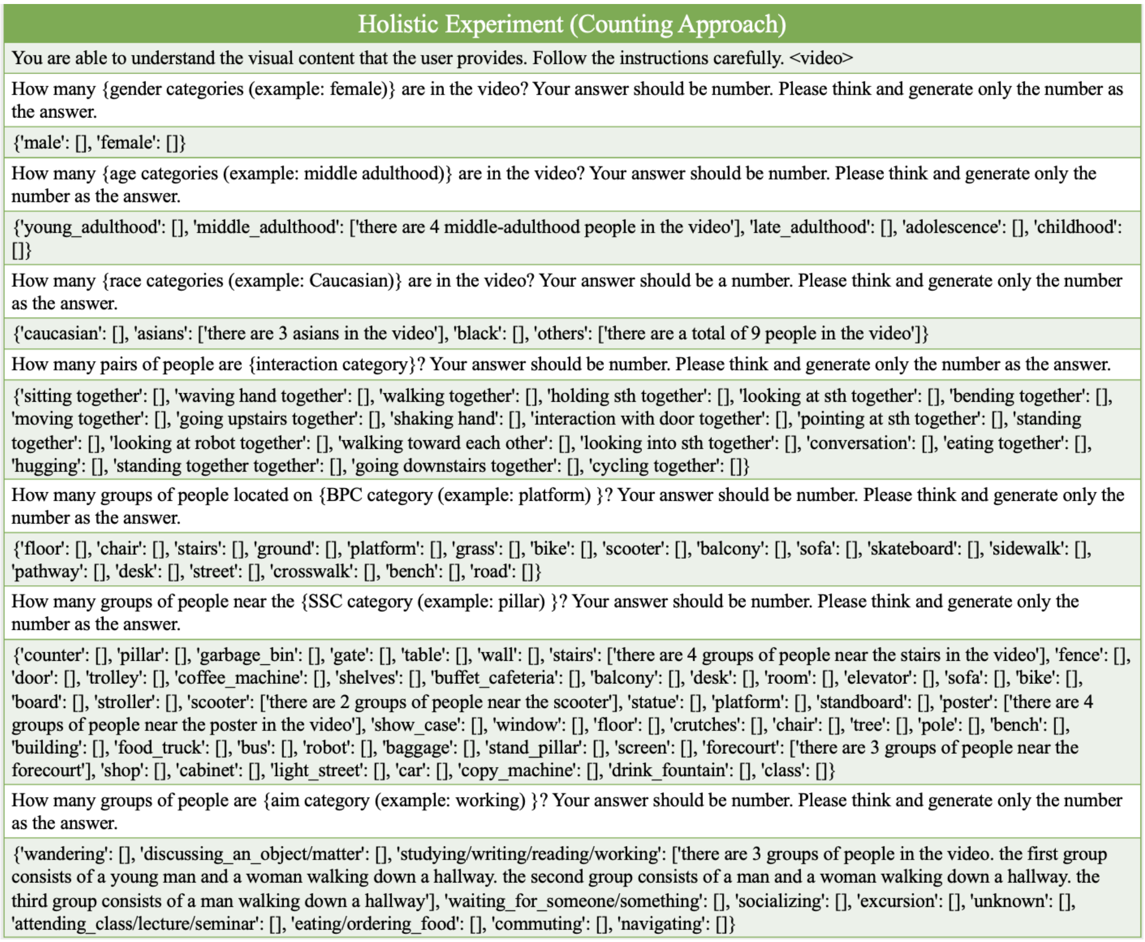
BPC Annotations: *floor*, *ground*, *chair*, *sidewalk*, *bike*, *stairs*, *platform*, *sofa*, *grass*, *street*, *crosswalk*, *road*, *scooter*, *skateboard*, *pathway*, *desk*, *balcony*, *bench*.

SSC Annotations: *gate*, *table*, *counter*, *door*,

Guided Perception Experiment
You are able to understand the visual content that the user provides. Follow the instructions carefully.
What is the gender of the person in the centre of the video? Your answer should be one of {male or female}. Please think and generate only one word as the answer.
'the person in the center of the video is male'
What is the age of the person in the centre of the video? Your answer should be one of {young adulthood, middle adulthood, late adulthood, adolescence or childhood}. Please think and generate only one word as the answer.
'the person in the center of the video is young, so the answer is young adult.'
What is the race of the person in the centre of the video? Your answer should be one of {Caucasian, Asian, Black or others}. Please think and generate only one word as the answer.
"i see a young man walking through a dark room at the beginning of the video. he is wearing a blue shirt and jeans. there is a person standing behind him, but i can't see their face. the room has a white wall, a white shelf, and a white piece of paper. the man in the center is tall, and he's carrying a white object in his hand. he's wearing a blue shirt with a white stripe."
What are the interactions between the people in the video? Your answer should be one or multiple of the following: {interactions category}. Please think and list all possible answers.
'there is no specific interaction between the people in the video. however, the people are generally going about their daily routines, such as walking down the hall or sitting in the living room. \n\nthey are also shown interacting with objects such as a clock on a wall, a sign outside a building, or a phone on a desk. \n\nthere is no physical interaction or conversation between the people in the video, and they do not appear to be together at any point.'
Where are the locations of most of the individuals in the group in the video? Your answer should be one or multiple of the following: {BPC category}. Please think and generate only one word as the answer.
'first, the video shows a hallway with people walking down it. one person is walking down the hallway, and there is a person and a woman walking in the hallway. next, people are walking through a corridor in an abandoned building. a person is walking down the corridor, and a young man is walking down the corridor in the darkness. then, there is a scene of a young man and a woman walking down a hallway in an abandoned building. they walk past a window, and a man walks past a window on the right. finally, the video shows a hallway in a dark building with people walking down it. there is a group of people walking down the hallway, and a young woman is walking down the hallway. overall, the video shows various scenes of people walking down hallways and corridors in abandoned buildings.'
What are the objects situated close to the group in the video? Your answer should be one or multiple of the following: {SSC category}. Please think and list all possible answers.
'in this video, the object of the person is a trolley and a bag. the person is seen walking down a long hallway in an old building. the hallway is dark and appears to be lit only by overhead lights. there are several people walking on the hallway floor. the walls are made of wood and there are several large windows that allow natural light to enter. the floors are made of concrete. a man with a backpack is seen walking down the hallway with a woman walking with a purse and wearing a white dress shirt. other people are seen walking down the hallway with bags on their shoulders. one person is seen walking and holding a briefcase. the sound of the video includes footsteps and people talking. the video is silent in terms of speech but there is background noise.'
What is the venue of the groups of people in the video? Your answer should be one of the following: {venue}. Please think and generate only one word as the answer."
'the venue of the groups of people in the video is a corridor.'
What are the aims and purposes of the group of people in the video? Your answer should be one or multiple of the following: {aim category} Your answer should be one or multiple of the following:
'the groups of people in the video appear to be wandering, studying, waiting, socializing, and commuting. there is no clear intention stated, but they seem to be enjoying the atmosphere of the hallway, walking down the corridors, and possibly attending class or lectures. some are holding books and others have backpacks on their shoulders. one group is seen walking down a dark hallway, while another one is in a dimly lit hall with a woman walking towards a building.'

Figure 1. Example of Guided Perception experiment: Questions are shown in white background, raw model answers in green.

pillar, *shelves*, *wall*, *standboard*, *poster*, *desk*, *food-truck*, *bike*, *chair*, *stairs*, *fence*, *show-case*, *room*, *board*, *cabinet*, *garbage-bin*, *stroller*, *elevator*, *buffet-cafeteria*, *trolley*, *forecourt*, *scooter*, *bus*, *robot*, *platform*, *window*, *tree*, *pole*, *crutches*, *stand-pillar*, *screen*, *car*, *copy-machine*, *class*, *coffee-machine*, *balcony*, *sofa*, *statue*, *floor*, *bench*, *building*, *baggage*, *shop*, *light-street*, *drink-*



Holistic Experiment (Counting Approach)
You are able to understand the visual content that the user provides. Follow the instructions carefully. <video>
How many {gender categories (example: female)} are in the video? Your answer should be number. Please think and generate only the number as the answer.
{'male': [], 'female': []}
How many {age categories (example: middle adulthood)} are in the video? Your answer should be number. Please think and generate only the number as the answer.
{'young_adulthood': [], 'middle_adulthood': ['there are 4 middle-adulthood people in the video'], 'late_adulthood': [], 'adolescence': [], 'childhood': []}
How many {race categories (example: Caucasian)} are in the video? Your answer should be a number. Please think and generate only the number as the answer.
{'caucasian': [], 'asians': ['there are 3 asians in the video'], 'black': [], 'others': ['there are a total of 9 people in the video']}
How many pairs of people are {interaction category}? Your answer should be number. Please think and generate only the number as the answer.
{'sitting together': [], 'waving hand together': [], 'walking together': [], 'holding sth together': [], 'looking at sth together': [], 'bending together': [], 'moving together': [], 'going upstairs together': [], 'shaking hand': [], 'interaction with door together': [], 'pointing at sth together': [], 'standing together': [], 'looking at robot together': [], 'walking toward each other': [], 'looking into sth together': [], 'conversation': [], 'eating together': [], 'hugging': [], 'standing together together': [], 'going downstairs together': [], 'cycling together': []}
How many groups of people located on {BPC category (example: platform) }? Your answer should be number. Please think and generate only the number as the answer.
{'floor': [], 'chair': [], 'stairs': [], 'ground': [], 'platform': [], 'grass': [], 'bike': [], 'scooter': [], 'balcony': [], 'sofa': [], 'skateboard': [], 'sidewalk': [], 'pathway': [], 'desk': [], 'street': [], 'crosswalk': [], 'bench': [], 'road': []}
How many groups of people near the {SSC category (example: pillar) }? Your answer should be number. Please think and generate only the number as the answer.
{'counter': [], 'pillar': [], 'garbage_bin': [], 'gate': [], 'table': [], 'wall': [], 'stairs': ['there are 4 groups of people near the stairs in the video'], 'fence': [], 'door': [], 'trolley': [], 'coffee_machine': [], 'shelves': [], 'buffet_cafeteria': [], 'balcony': [], 'desk': [], 'room': [], 'elevator': [], 'sofa': [], 'bike': [], 'board': [], 'stroller': [], 'scooter': ['there are 2 groups of people near the scooter'], 'statue': [], 'platform': [], 'standboard': [], 'poster': ['there are 4 groups of people near the poster in the video'], 'show_case': [], 'window': [], 'floor': [], 'crutches': [], 'chair': [], 'tree': [], 'pole': [], 'bench': [], 'building': [], 'food_truck': [], 'bus': [], 'robot': [], 'baggage': [], 'stand_pillar': [], 'screen': [], 'forecourt': ['there are 3 groups of people near the forecourt'], 'shop': [], 'cabinet': [], 'light_street': [], 'car': [], 'copy_machine': [], 'drink_fountain': [], 'class': []}
How many groups of people are {aim category (example: working) }? Your answer should be number. Please think and generate only the number as the answer.
{'wandering': [], 'discussing_an_object/matter': [], 'studying/writing/reading/working': ['there are 3 groups of people in the video. the first group consists of a young man and a woman walking down a hallway. the second group consists of a man and a woman walking down a hallway. the third group consists of a man walking down a hallway'], 'waiting_for_someone/something': [], 'socializing': [], 'excursion': [], 'unknown': [], 'attending_class/lecture/seminar': [], 'eating/ordering_food': [], 'commuting': [], 'navigating': []}

Figure 2. Example of Holistic Counting Approach experiment: Questions are shown in white background, raw model answers in green.

fountain.

2. Experiments

In the paper, we only presented F1 scores for all data. In this section, we offer accuracy values for the training, validation, and test datasets separately for both Guided Perception and Holistic approaches, Table 1, 2, 3, 4, 5, 6, 7, 8, 9 and 10. Additionally, as previously mentioned, we employed a Five Ensemble strategy, denoted in tables as (5 Ens).

2.1. Prompting

In the Experiment section of the paper, we initially presented the prompt schematic. In this section, we provided the raw response of Video-LLaMA. Figures 1, 2, and 3 illustrate the guided perception experiment, holistic experiment (counting approach), and holistic experiment (binary approach), respectively.

Holistic Experiment (Binary Approach)
You are able to understand the visual content that the user provides. Follow the instructions carefully. <video>
Do you see {gender categories (example: female)} in the video? Your answer should be yes or no. Please think and generate only the word as the answer.
{'male': ['yes, there is a male in the video'], 'female': ['yes, i see a female in the video']}
Do you see {age categories (example: middle adulthood)} in the video? Your answer should be yes or no. Please think and generate only the word as the answer.
{'young_adulthood': ['yes, i see young adulthood in the video'], 'middle_adulthood': ['yes, i see middle_adulthood in the video'], 'late_adulthood': ['yes, i see late adulthood in the video'], 'adolescence': ['yes, i see adolescence in the video'], 'childhood': ['yes, i see childhood in the video']}
Do you see {race categories (example: Caucasian)} in the video? Your answer should be yes or no. Please think and generate only the word as the answer.
{'caucasian': ['yes, i see a caucasian man in the video'], 'asians': ['yes, i see an asian man in the video'], 'black': ['yes, i see black in the video'], 'others': ['yes, there are other people in the video. i see a man, a woman, and a person in the video']}
Do you see any pair of people are {interaction category (example: standing together)}? Your answer should be yes or no. Please think and generate only the word as the answer.
{'sitting together': ['yes, i see a pair of people sitting together. they are a young man and a woman'], 'waving hand together': ['yes, i see a pair of people waving their hands together'], 'walking together': ['yes, i see a pair of people walking together in the video'], 'holding sth together': ['yes, i see a pair of people holding a white plastic bag together'], 'looking at sth together': ['yes, i see a pair of people looking at something together. they are walking down a hallway in a building'], 'bending together': ['yes, i see a pair of people bending together in the video'], 'moving together': ['yes, i see a pair of people walking together in the video. they are walking down a hallway in a school building'], 'going upstairs together': ['yes, i see a pair of people going upstairs together. they are walking in an office building'], 'shaking hand': ['yes, there is a pair of people shaking hands in the video'], 'interaction with door together': ['yes, i see a pair of people interacting with a door together in the video'], 'pointing at sth together': ['yes, i see a pair of people pointing at something together'], 'standing together': ['yes, there is a pair of people standing together in the video. they are a man and a woman walking down a hallway in an old building'], 'looking at robot together': ['yes, i see a pair of people looking at a robot together in one of the frames'], 'walking toward each other': ['yes, i see a pair of people walking toward each other. they are wearing different clothes and are walking on a floor'], 'looking into sth together': ['yes, there is a pair of people looking into something together. they are a man and a woman standing in the hallway of a building. they appear to be looking at something on their phone'], 'conversation': ['yes, i see a pair of people in a conversation in the video. they are standing on a street, and they are having a conversation'], 'eating together': ['yes, i see a pair of people eating together. they are a man and a woman, and they are eating together in a restaurant'], 'hugging': ['yes, i see a pair of people hugging in the video'], 'standing together together': ['yes, i see a pair of people standing together in the video. they are a young man and a young woman walking down a hallway in an old building'], 'going downstairs together': ['yes, there is a pair of people going downstairs together in the video. they are a young woman and a young man'], 'cycling together': ['yes, there is a pair of people cycling together in the video. they are cycling down the hallway of a building']}
Do you see any group located on {BPC category (example: floor} }? Your answer should be yes or no. Please think and generate only the word as the answer.
{'floor': ['yes, there is a group of people on the floor'], 'chair': ['yes, there is a group of people located on a chair in the video'], 'stairs': ['yes, i see a group of people walking down the stairs in the video'], 'ground': ['yes, i see a group of people standing on the ground'], 'platform': ['yes, i see a group of people on the platform'], 'grass': ['yes, there is a group of people on the grass in the second frame'], 'bike': ['yes, i see a group of people on bikes in the video. they are located on a sidewalk in front of a building'], 'scooter': ['yes, i see a group of people riding scooters in the video'], 'balcony': ['yes, i see a group of people on a balcony'], 'sofa': ['yes, i see a group of people sitting on a sofa in the video'], 'skateboard': ['yes, there is a group of people on a skateboard in the video'], 'sidewalk': ['yes, i see a group of people on the sidewalk'], 'pathway': ['yes, i see a group of people on a pathway in the video'], 'desk': ['yes, i see a group of people located on a desk'], 'street': ['yes, i see a group of people standing on the street'], 'crosswalk': ['yes, there is a group of people located on the crosswalk. they are walking towards the camera and are wearing different clothes'], 'bench': ['yes, i see a group of people sitting on a bench in the video'], 'road': ['yes, there is a group of people walking on the road. they are wearing different types of clothing and carrying different types of bags. one of them is wearing a backpack, and another one is carrying a purse. there is also a man walking on the road, and a man standing on the road']}, 'answers': {'floor': 1, 'chair': 1, 'stairs': 1, 'ground': 1, 'platform': 1, 'grass': 1, 'bike': 1, 'scooter': 1, 'balcony': 1, 'sofa': 1, 'skateboard': 1, 'sidewalk': 1, 'pathway': 1, 'desk': 1, 'street': 1, 'crosswalk': 1, 'bench': 1, 'road': 1}, 'label': None}
Do you see any group near the {SSC category (example: pillar)}? Your answer should be yes or no. Please think and generate only the word as the answer.
{'counter': ['yes, there is a group of people near the counter'], 'pillar': ['yes, i see a group of people near the pillar. they are wearing various clothes, including a brown shirt, a black shirt, and a white shirt'], 'garbage_bin': ['yes, there is a group of people near the garbage bin in the second frame'], 'gate': ['yes, i see a group of people standing near the gate'], 'table': ['yes, there is a group of people near the table. they are sitting on the floor and on chairs'], 'wall': ['yes, i see a group of people near the wall. they are standing near the wall, and there is a white bag on the ground'], 'stairs': ['yes, there is a group of people near the stairs'], 'fence': ['yes, i see a group of people near the fence'], 'door': ['yes, i see a group of people near the door'], 'trolley': ['yes, i see a group of people near the trolley. they are standing in a hallway, and there are people in the background'], 'coffee_machine': ['yes, i see a group of people near the coffee machine'], 'shelves': ['yes, i see a group of people near the shelves in the video'], 'buffet_cafeteria': ['yes, i see a group of people near the buffet/cafeteria'], 'balcony': ['yes, i see a group of people near the balcony. they are walking down the hallway towards the door'], 'desk': ['yes, there is a group of people near the desk'], 'room': ['yes, i see a group of people near the room'], 'elevator': ['yes, i see a group of people near the elevator'], 'sofa': ['yes, i see a group of people near the sofa'], 'bike': ['yes, there is a group of people near the bike in the video'], 'board': ['yes, i see a group of people near the board'], 'stroller': ['yes, i see a group of people walking near a stroller'], 'scooter': ['yes, i see a group of people near the scooter'], 'statue': ['yes, i see a group of people standing near a statue in the video'], 'platform': ['yes, i see a group of people near the platform'], 'standboard': ['yes, i see a group of people near the standboard'], 'poster': ['yes, there is a group of people near the poster'], 'show_case': ['yes, i see a group of people near the showcase'], 'window': ['yes, i see a group of people near the window. they are standing in the hallway of a building'], 'floor': ['yes, i see a group of people near the floor'], 'crutches': ['yes, i see a group of people walking near the crutches'], 'chair': ['yes, there is a group of people near the chair'], 'tree': ['yes, i see a group of people near the tree'], 'pole': ['yes, there is a group of people near the pole. they are standing on the sidewalk and walking towards the entrance of a building'], 'bench': ['yes, there is a group of people sitting on the bench'], 'building': ['yes, i see a group of people near the building. they are walking down the hallway'], 'food_truck': ['yes, i see a group of people near the food truck'], 'bus': ['yes, i see a group of people near the bus'], 'robot': ['yes, i see a group of people near the robot'], 'baggage': ['yes, i see a group of people near the baggage in the video'], 'stand_pillar': ['yes, i see a group of people near the stand pillar'], 'screen': ['yes, there is a group of people near the screen'], 'forecourt': ['yes, i see a group of people near the forecourt'], 'shop': ['yes, i see a group of people standing near a shop in the video'], 'cabinet': ['yes, i see a group of people near the cabinet'], 'light_street': ['yes, there is a group of people near the light street'], 'car': ['yes, there is a group of people near the car'], 'copy_machine': ['yes, there is a group of people near the copy machine'], 'drink_fountain': ['yes, there is a group of people near the drink fountain'], 'class': ['yes, i see a group of people near the class']}
Do you see any group are {aim category (example: socializing)} }? Your answer should be yes or no. Please think and generate only the word as the answer.
{'wandering': ['yes, there are people walking in a group in the video'], 'discussing_an_object/matter': ['yes, i see a group of people discussing something in the video. they are standing in the hallway of a building and appear to be having a conversation'], 'studying/writing/reading/working': ['yes, there is a group of people studying in the library'], 'waiting_for_someone/something': ['yes, i see a group of people waiting for someone or something in the video. they are standing in a hallway, and it seems that they are waiting for a train'], 'socializing': ['yes, i see a group of people socializing in the video. they are standing in a hallway and appear to be talking to each other'], 'excursion': ['yes, i see a group of people on an excursion. they are walking down a hallway in a building'], 'unknown': ['yes, i can see a group of people walking down a hallway in the video. they are walking towards a door and appear to be wearing different colored clothing. there is a man in a brown jacket, a woman in a brown jacket, and a woman wearing a black jacket. the group is walking towards the door, and the camera captures a clear view of the hallway. the walls are painted white, and there are windows on the left side of the building. the door is made of wood and has a glass window. the video captures the scene at different times, and the people are walking in different directions. there are also other objects in the video, such as a white and black table, a white and black chair, and a white and black pillow'], 'attending_class/lecture/seminar': ['yes, there is a group of people attending a lecture or seminar in the video'], 'eating/ordering_food': ['yes, there is a group of people eating food in the restaurant'], 'commuting': ['yes, i see a group of people commuting in the video. they are walking down a hallway in an old building, and there are multiple people in the group'], 'navigating': ['yes, i see a group of people navigating through the building']}

Figure 3. Example of Holistic Binary Approach experiment: Questions are shown in white background, raw model answers in green.

Multi-modal LLM	**Individual Level**			**Intra-Group Level**	**Social Group Level**			
	Gender	Age	Race	Interactions	BPC	SSC	Venue	Purpose
Video-LLaMA (LLaMA-2 13B) [1] (5 Ens)	0.8675	0.8874	0.5661	0.8495	0.7128	0.9010	0.1965	0.7741
Video-LLaMA (LLaMA-2 13B) [1]	0.8188	0.5916	0.4261	0.8505	0.4940	0.8359	0.1780	0.7532
Video-LLaMA (LLaMA-2 7B) [1] (5 Ens)	0.7909	0.9489	0.5461	0.8901	0.5434	0.5434	0.1719	0.8528
Video-LLaMA (LLaMA-2 7B) [1]	0.7549	0.7703	0.4815	0.8804	0.4910	0.7855	0.2088	0.8159
Valley (LLaMA-1 13B) [2] (5 Ens)	0.7630	0.9535	0.1984	0.5376	0.1769	0.9628	0.0644	0.8336
Valley (LLaMA-1 13B) [2]	0.7630	0.9396	0.1015	0.6261	0.0614	0.9414	0.0818	0.8229
Valley (LLaMA-2 7B) [2] (5 Ens)	0.4239	0.8712	0.1123	0.6725	0.4400	0.9321	0.0102	0.8934
Valley (LLaMA-2 7B) [2]	0.1881	0.5614	0.0646	0.7318	0.2248	0.8934	0.0286	0.8769
OTTER (LLaMA-1 7B) [3] (5 Ens)	0.2473	0.3561	0.0153	0.9420	0.2068	0.9346	0.0429	0.8816
OTTER (LLaMA-1 7B) [3]	0.2473	0.3561	0.0153	0.9420	0.2068	0.9346	0.0429	0.8816
MiniGPT-4 (LLaMA-2 7B) [4] (5 Ens)	0.8687	0.7412	0.5215	0.7665	0.8058	0.8467	0.3190	0.8832
MiniGPT-4 (LLaMA-2 7B) [4]	0.7735	0.5591	0.3923	0.7999	0.6979	0.7230	0.2852	0.8370
InstructBLIP (Vicuna-V1 13B) [5] (5 Ens)	0.8385	0.6380	0.5261	0.8609	0.7533	0.9667	0.0460	0.1025
InstructBLIP (Vicuna-V1 13B) [5]	0.8362	0.2552	0.5292	0.8571	0.7563	0.9608	0.0460	0.1089
InstructBLIP (Vicuna-V1 7B) [5] (5 Ens)	0.8606	0.0348	0.5446	0.0821	0.28782	0.9124	0.1441	0.1251
InstructBLIP (Vicuna-V1 7B) [5]	0.8466	0.0336	0.5215	0.1626	0.2631	0.87819	0.1688	0.1513

Table 1. **Guided Perception**: Comparing popular multi-modal LLMs across the JRDB-Social at three levels in accuracy for the train set. BPC = Engagement of Body Position's connection with the Content, SSC = Salient Scene Content, (5 Ens) = Five Ensemble Strategy.

Multi-modal LLM	**Individual Level**			**Intra-Group Level**	**Social Group Level**			
	Gender	Age	Race	Interactions	BPC	SSC	Venue	Purpose
Video-LLaMA (LLaMA-2 13B) [1] (5 Ens)	0.7795	0.7603	0.6220	0.6236	0.5707	0.9288	0.5630	0.7661
Video-LLaMA (LLaMA-2 13B) [1]	0.7731	0.5047	0.4418	0.6729	0.3863	0.8467	0.4761	0.7427
Video-LLaMA (LLaMA-2 7B) [1] (5 Ens)	0.7150	0.8498	0.5465	0.8610	0.3636	0.8936	0.5154	0.8479
Video-LLaMA (LLaMA-2 7B) [1]	0.6932	0.7188	0.5058	0.7297	0.3888	0.8164	0.5126	0.8082
Valley (LLaMA-1 13B) [2](5 Ens)	0.7667	0.5271	0.5523	0.8819	0.5732	0.9653	0.4817	0.1102
Valley (LLaMA-1 13B) [2]	0.6932	0.8498	0.1046	0.5450	0.0277	0.9020	0.4425	0.8304
Valley (LLaMA-2 7B) [2](5 Ens)	0.4632	0.8434	0.1337	0.6029	0.2070	0.9143	0.4789	0.8924
Valley (LLaMA-2 7B) [2]	0.2619	0.5878	0.0639	0.6854	0.1035	0.8658	0.4677	0.8696
OTTER (LLaMA-1 7B) [3](5 Ens)	0.2268	0.2044	0.0058	0.9288	0.1691	0.9143	0.0084	0.8880
OTTER (LLaMA-1 7B) [3]	0.2268	0.2044	0.0058	0.9286	0.1691	0.9143	0.0084	0.8880
MiniGPT-4 (LLaMA-2 7B) [4] (5 Ens)	0.8019	0.6709	0.5058	0.8510	0.6363	0.8876	0.5742	0.8763
MiniGPT-4 (LLaMA-2 7B) [4]	0.7252	0.5527	0.4302	0.8505	0.5984	0.7931	0.3445	0.8313
InstructBLIP (Vicuna-V1 13B) [5] (5 Ens)	0.7667	0.5271	0.5523	0.8819	0.5732	0.9653	0.4817	0.1102
InstructBLIP (Vicuna-V1 13B) [5]	0.7444	0.2076	0.5465	0.8971	0.5732	0.9653	0.4845	0.1102
InstructBLIP (Vicuna-V1 7B) [5](5 Ens)	0.7667	0.1054	0.5755	0.0821	0.4444	0.9473	0.5042	0.1578
InstructBLIP (Vicuna-V1 7B) [5]	0.7635	0.6709	0.5290	0.1270	0.3813	0.8814	0.5042	0.1780

Table 2. **Guided Perception**: Comparing popular multi-modal LLMs across the JRDB-Social at three levels in accuracy for the validation set. BPC = Engagement of Body Position's connection with the Content, SSC = Salient Scene Content, (5 Ens) = Five Ensemble Strategy.

Multi-modal LLM	**Individual Level**			**Intra-Group Level**	**Social Group Level**			
	Gender	Age	Race	Interactions	BPC	SSC	Venue	Purpose
Video-LLaMA (LLaMA-2 13B) [1] (5 Ens)	0.7609	0.6896	0.4215	0.8361	0.5094	0.6199	0.5619	0.7795
Video-LLaMA (LLaMA-2 13B) [1]	0.7413	0.4674	0.3296	0.8425	0.3541	0.8731	0.5004	0.7520
Video-LLaMA (LLaMA-2 7B) [1] (5 Ens)	0.6723	0.7154	0.4035	0.8597	0.4495	0.9048	0.4035	0.8524
Video-LLaMA (LLaMA-2 7B) [1]	0.6485	0.5566	0.3475	0.8557	0.3996	0.8349	0.4796	0.8095
Valley (LLaMA-1 13B) [2](5 Ens)	0.6110	0.7288	0.1535	0.4947	0.1165	0.9700	0.4001	0.8441
Valley (LLaMA-1 13B) [2]	0.6306	0.7596	0.0862	0.5913	0.0688	0.9446	0.3972	0.8263
Valley (LLaMA-2 7B) [2](5 Ens)	0.6111	0.7288	0.1536	0.4947	0.1165	0.9701	0.4001	0.8441
Valley (LLaMA-2 7B) [2]	0.1566	0.4648	0.0381	0.6866	0.1842	0.8743	0.3887	0.8693
OTTER (LLaMA-1 7B) [3](5 Ens)	0.2221	0.2480	0.0168	0.9274	0.2253	0.9473	0.0702	0.8836
OTTER (LLaMA-1 7B) [3]	0.2221	0.2480	0.0168	0.9280	0.2253	0.9477	0.0630	0.8836
MiniGPT-4 (LLaMA-2 7B) [4] (5 Ens)	0.7353	0.5879	0.4148	0.7536	0.5694	0.7416	0.4989	0.8735
MiniGPT-4 (LLaMA-2 7B) [4]	0.6783	0.4398	0.3363	0.7885	0.5361	0.6676	0.3644	0.8214
InstructBLIP (Vicuna-V1 13B) [5] (5 Ens)	0.6817	0.4942	0.3924	0.8182	0.4340	0.9751	0.4803	0.1076
InstructBLIP (Vicuna-V1 13B) [5]	0.6928	0.1971	0.3991	0.8375	0.4351	0.9655	0.4860	0.1164
InstructBLIP (Vicuna-V1 7B) [5](5 Ens)	0.7106	0.2168	0.4013	0.0924	0.3674	0.8995	0.5634	0.1327
InstructBLIP (Vicuna-V1 7B) [5]	0.7089	0.2061	0.3857	0.1447	0.3019	0.8435	0.5555	0.1571

Table 3. **Guided Perception**: Comparing popular multi-modal LLMs across the JRDB-Social at three levels in accuracy for the test set. BPC = Engagement of Body Position's connection with the Content, SSC = Salient Scene Content, (5 Ens) = Five Ensemble Strategy.

Multi-modal LLM	**Individual Level**			**Intra-Group Level**	**Social Group Level**			
	Gender	Age	Race	Interactions	BPC	SSC	Venue	Purpose
Video-LLaMA (LLaMA-2 13B) [1] (5 Ens)	0.2634	0.1575	0.2780	0.0711	0.0531	0.0257	0.3158	0.1173
Video-LLaMA (LLaMA-2 13B) [1]	0.1196	0.1556	0.2151	0.0517	0.0603	0.0270	0.3579	0.0969
Video-LLaMA (LLaMA-2 7B) [1] (5 Ens)	0.1799	0.1380	0.2404	0.0662	0.0444	0.0212	0.2632	0.0943
Video-LLaMA (LLaMA-2 7B) [1]	0.0545	0.0612	0.1344	0.0280	0.0259	0.0166	0.2632	0.0283
Valley (LLaMA-1 13B) [2](5 Ens)	0.0004	0.0067	0.0177	0.0017	0.0000	0.0000	0.1790	0.0004
Valley (LLaMA-1 13B) [2]	0.0000	0.0060	0.0066	0.0017	0.0000	0.0000	0.1790	0.0000
Valley (LLaMA-2 7B) [2](5 Ens)	0.0251	0.0353	0.0150	0.0061	0.0000	0.0000	0.0842	0.0091
Valley (LLaMA-2 7B) [2]	0.0009	0.0034	0.0025	0.0038	0.0043	0.0000	0.0526	0.0013
OTTER (LLaMA-1 7B) [3](5 Ens)	0.0000	0.0002	0.0003	0.0000	0.0000	0.0016	0.1474	0.0000
OTTER (LLaMA-1 7B) [3]	0.0000	0.0002	0.0003	0.0000	0.0000	0.0016	0.1474	0.0000
MiniGPT-4 (LLaMA-2 7B) [4] (5 Ens)	0.1921	0.1310	0.2202	0.0691	0.0484	0.0320	0.4211	0.1265
MiniGPT-4 (LLaMA-2 7B) [4]	0.2428	0.1255	0.2096	0.0297	0.1241	0.0410	0.2632	0.1241
InstructBLIP (Vicuna-V1 13B) [5] (5 Ens)	0.0665	0.0921	0.0425	0.0455	0.0132	0.0309	0.0632	0.0631
InstructBLIP (Vicuna-V1 13B) [5]	0.1205	0.1041	0.0927	0.0592	0.0237	0.0275	0.0632	0.0663
InstructBLIP (Vicuna-V1 7B) [5](5 Ens)	0.0907	0.1260	0.0640	0.0678	0.0262	0.0351	0.0947	0.0565
InstructBLIP (Vicuna-V1 7B) [5]	0.1399	0.1256	0.1228	0.0824	0.0082	0.0258	0.1368	0.0417

Table 4. **Holistic Counting Approach**: Comparing popular multi-modal LLMs across the JRDB-Social at three levels in accuracy for the train set. BPC = Engagement of Body Position's connection with the Content, SSC = Salient Scene Content, (5 Ens) = Five Ensemble Strategy.

Multi-modal LLM	**Individual Level**			**Intra-Group Level**	**Social Group Level**			
	Gender	Age	Race	Interactions	BPC	SSC	Venue	Purpose
Video-LLaMA (LLaMA-2 13B) [1] (5 Ens)	0.3083	0.2189	0.2668	0.0636	0.0755	0.0355	0.3429	0.1343
Video-LLaMA (LLaMA-2 13B) [1]	0.1186	0.2224	0.2006	0.0403	0.0802	0.0419	0.3143	0.0899
Video-LLaMA (LLaMA-2 7B) [1] (5 Ens)	0.1458	0.1814	0.2281	0.0401	0.0835	0.0305	0.2286	0.1254
Video-LLaMA (LLaMA-2 7B) [1]	0.0235	0.1129	0.1125	0.0132	0.0462	0.0217	0.2571	0.0486
Valley (LLaMA-1 13B) [2](5 Ens)	0.0000	0.0202	0.0213	0.0045	0.0000	0.0000	0.2571	0.0000
Valley (LLaMA-1 13B) [2]	0.0000	0.0109	0.0001	0.0000	0.0000	0.0000	0.1429	0.0000
Valley (LLaMA-2 7B) [2](5 Ens)	0.0355	0.0762	0.0012	0.0087	0.0000	0.0044	0.1429	0.0046
Valley (LLaMA-2 7B) [2]	0.0000	0.0107	0.0000	0.0000	0.0069	0.0000	0.1143	0.0074
OTTER (LLaMA-1 7B) [3](5 Ens)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
OTTER (LLaMA-1 7B) [3]	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
MiniGPT-4 (LLaMA-2 7B) [4] (5 Ens)	0.1946	0.1785	0.2317	0.0621	0.0722	0.0376	0.3714	0.1703
MiniGPT-4 (LLaMA-2 7B) [4]	0.2051	0.1744	0.2033	0.0606	0.0931	0.0446	0.4571	0.1601
InstructBLIP (Vicuna-V1 13B) [5] (5 Ens)	0.0592	0.0678	0.0193	0.0333	0.0079	0.0246	0.2000	0.0554
InstructBLIP (Vicuna-V1 13B) [5]	0.0949	0.0966	0.0335	0.0586	0.0069	0.0288	0.2000	0.0685
InstructBLIP (Vicuna-V1 7B) [5](5 Ens)	0.0713	0.1136	0.0311	0.0236	0.0285	0.0491	0.3429	0.0776
InstructBLIP (Vicuna-V1 7B) [5]	0.1060	0.1566	0.0882	0.0519	0.0275	0.0381	0.3429	0.0600

Table 5. **Holistic Counting Approach**: Comparing popular multi-modal LLMs across the JRDB-Social at three levels in accuracy for the validation set. BPC = Engagement of Body Position's connection with the Content, SSC = Salient Scene Content, (5 Ens) = Five Ensemble Strategy.

Multi-modal LLM	**Individual Level**			**Intra-Group Level**	**Social Group Level**			
	Gender	Age	Race	Interactions	BPC	SSC	Venue	Purpose
Video-LLaMA (LLaMA-2 13B) [1] (5 Ens)	0.2263	0.2239	0.2897	0.0894	0.0669	0.0093	0.4444	0.1581
Video-LLaMA (LLaMA-2 13B) [1]	0.1142	0.2058	0.2109	0.0619	0.0822	0.0107	0.4000	0.1097
Video-LLaMA (LLaMA-2 7B) [1] (5 Ens)	0.1539	0.2019	0.2282	0.0680	0.0449	0.0065	0.3185	0.1122
Video-LLaMA (LLaMA-2 7B) [1]	0.0000	0.0076	0.0040	0.0036	0.0000	0.0000	0.3111	0.0078
Valley (LLaMA-1 13B) [2] (5 Ens)	0.0011	0.0229	0.0223	0.0000	0.0000	0.0000	0.2000	0.0007
Valley (LLaMA-1 13B) [2]	0.0011	0.0081	0.0038	0.0000	0.0010	0.0000	0.2222	0.0000
Valley (LLaMA-2 7B) [2] (5 Ens)	0.0059	0.0397	0.0077	0.0099	0.0038	0.0000	0.1818	0.0244
Valley (LLaMA-2 7B) [2]	0.0000	0.0076	0.0040	0.0036	0.0000	0.0000	0.0637	0.0078
OTTER (LLaMA-1 7B) [3] (5 Ens)	0.0020	0.0000	0.0001	0.0000	0.0000	0.0000	0.0296	0.0000
OTTER (LLaMA-1 7B) [3]	0.0020	0.0000	0.0001	0.0000	0.0000	0.0000	0.0296	0.0000
MiniGPT-4 (LLaMA-2 7B) [4] (5 Ens)	0.1271	0.1842	0.1846	0.0878	0.0754	0.0138	0.4148	0.1689
MiniGPT-4 (LLaMA-2 7B) [4]	0.1927	0.1723	0.2022	0.0899	0.0615	0.0140	0.2741	0.1742
InstructBLIP (Vicuna-V1 13B) [5] (5 Ens)	0.0444	0.0743	0.0132	0.0413	0.0145	0.0168	0.2667	0.0635
InstructBLIP (Vicuna-V1 13B) [5]	0.0913	0.1226	0.0662	0.0744	0.0173	0.0162	0.2593	0.0885
InstructBLIP (Vicuna-V1 7B) [5] (5 Ens)	0.0554	0.1053	0.0376	0.0755	0.0241	0.0194	0.3778	0.0757
InstructBLIP (Vicuna-V1 7B) [5]	0.1062	0.1242	0.0798	0.0943	0.0158	0.0153	0.3852	0.0446

Table 6. **Holistic Counting Approach**: Comparing popular multi-modal LLMs across the JRDB-Social at three levels in accuracy for the test set. BPC = Engagement of Body Position's connection with the Content, SSC = Salient Scene Content, (5 Ens) = Five Ensemble Strategy.

Multi-modal LLM	**Individual Level**			**Intra-Group Level**	**Social Group Level**			
	Gender	Age	Race	Interactions	BPC	SSC	Venue	Purpose
Video-LLaMA (LLaMA-2 13B) [1] (5 Ens)	0.9350	0.3320	0.5800	0.2604	0.1466	0.0755	0.1583	0.3272
Video-LLaMA (LLaMA-2 13B) [1]	0.9300	0.3380	0.5800	0.3466	0.1983	0.1283	0.1583	0.3490
Video-LLaMA (LLaMA-2 7B) [1] (5 Ens)	0.9350	0.3320	0.5800	0.1757	0.2572	0.0504	0.1583	0.2572
Video-LLaMA (LLaMA-2 7B) [1]	0.9350	0.3320	0.5800	0.1938	0.1127	0.0512	0.1583	0.2654
Valley (LLaMA-1 13B) [2] (5 Ens)	0.8600	0.4940	0.5775	0.3576	0.6055	0.4459	0.1716	0.6090
Valley (LLaMA-1 13B) [2]	0.7700	0.5320	0.5825	0.3461	0.5666	0.4581	0.1666	0.6136
Valley (LLaMA-2 7B) [2] (5 Ens)	0.5050	0.6120	0.4100	0.2947	0.2239	0.6742	0.6333	0.5718
Valley (LLaMA-2 7B) [2]	0.4900	0.6220	0.4175	0.3076	0.2483	0.6453	0.6033	0.5690
OTTER (LLaMA-1 7B) [3](5 Ens)	0.8700	0.5280	0.5875	0.1852	0.1200	0.0624	0.3233	0.2990
OTTER (LLaMA-1 7B) [3]	0.8700	0.5280	0.5875	0.1852	0.1200	0.0624	0.3233	0.2990
MiniGPT-4 (LLaMA-2 7B) [4] (5 Ens)	0.7700	0.7240	0.4675	0.7328	0.8139	0.6377	0.7317	0.6845
MiniGPT-4 (LLaMA-2 7B) [4]	0.7250	0.6600	0.5325	0.6861	0.7639	0.5885	0.6683	0.6200
InstructBLIP (Vicuna-V1 13B) [5] (5 Ens)	0.7600	0.4820	0.6225	0.6400	0.7938	0.6083	0.2900	0.5009
InstructBLIP (Vicuna-V1 13B) [5]	0.7650	0.4780	0.6325	0.6309	0.7866	0.5810	0.2983	0.4963
InstructBLIP (Vicuna-V1 7B) [5] (5 Ens)	0.8000	0.4480	0.6125	0.5409	0.6066	0.2314	0.3583	0.4563
InstructBLIP (Vicuna-V1 7B) [5]	0.7714	0.5085	0.5785	0.5823	0.6095	0.3253	0.4809	0.5168

Table 7. **Holistic Experiment (Binary Approach)**: Comparing popular multi-modal LLMs across the JRDB-Social at three levels in accuracy for the train set. BPC = Engagement of Body Position's connection with the Content, SSC = Salient Scene Content, (5 Ens) = Five Ensemble Strategy.

Multi-modal LLM	**Individual Level**			**Intra-Group Level**	**Social Group Level**			
	Gender	Age	Race	Interactions	BPC	SSC	Venue	Purpose
Video-LLaMA (LLaMA-2 13B) [1] (5 Ens)	0.9714	0.3771	0.5571	0.3007	0.2111	0.1155	0.1667	0.3688
Video-LLaMA (LLaMA-2 13B) [1]	0.9714	0.3886	0.5500	0.3633	0.2698	0.1878	0.1667	0.4520
Video-LLaMA (LLaMA-2 7B) [1] (5 Ens)	0.9714	0.3771	0.5571	0.1918	0.1587	0.0623	0.1666	0.3168
Video-LLaMA (LLaMA-2 7B) [1]	0.9714	0.3771	0.5571	0.1986	0.1587	0.0629	0.1666	0.3194
Valley (LLaMA-1 13B) [2] (5 Ens)	0.8857	0.5029	0.5643	0.3415	0.5683	0.4752	0.1762	0.6286
Valley (LLaMA-1 13B) [2]	0.7143	0.5086	0.5643	0.3633	0.5730	0.4647	0.2000	0.6182
Valley (LLaMA-2 7B) [2] (5 Ens)	0.5571	0.6514	0.3643	0.3102	0.2556	0.6974	0.6619	0.6494
Valley (LLaMA-2 7B) [2]	0.5571	0.5543	0.3714	0.3306	0.2762	0.6706	0.6048	0.6208
OTTER (LLaMA-1 7B) [3] (5 Ens)	0.9714	0.5714	0.5500	0.1973	0.1635	0.0659	0.3095	0.3117
OTTER (LLaMA-1 7B) [3]	0.9714	0.5714	0.5500	0.1973	0.1635	0.0659	0.3095	0.3117
MiniGPT-4 (LLaMA-2 7B) [4] (5 Ens)	0.7000	0.7429	0.5929	0.7265	0.7683	0.6583	0.7429	0.6909
MiniGPT-4 (LLaMA-2 7B) [4]	0.7143	0.7029	0.5643	0.6925	0.7349	0.6525	0.6857	0.6779
InstructBLIP (Vicuna-V1 13B) [5] (5 Ens)	0.7714	0.5429	0.6071	0.6599	0.7714	0.6886	0.3095	0.5455
InstructBLIP (Vicuna-V1 13B) [5]	0.7000	0.5714	0.6214	0.6776	0.7746	0.6606	0.3238	0.5584
InstructBLIP (Vicuna-V1 7B) [5] (5 Ens)	0.8143	0.5086	0.5714	0.5605	0.5952	0.2321	0.4476	0.4987
InstructBLIP (Vicuna-V1 7B) [5]	0.7714	0.5086	0.5786	0.5823	0.6095	0.3254	0.4810	0.5169

Table 8. **Holistic Experiment (Binary Approach)**: Comparing popular multi-modal LLMs across the JRDB-Social at three levels in accuracy for the validation set. BPC = Engagement of Body Position's connection with the Content, SSC = Salient Scene Content, (5 Ens) = Five Ensemble Strategy.

Multi-modal LLM	**Individual Level**			**Intra-Group Level**	**Social Group Level**			
	Gender	Age	Race	Interactions	BPC	SSC	Venue	Purpose
Video-LLaMA (LLaMA-2 13B) [1] (5 Ens)	0.9815	0.4459	0.6185	0.3083	0.1621	0.0584	0.1667	0.3953
Video-LLaMA (LLaMA-2 13B) [1]	0.9815	0.4533	0.6204	0.3746	0.2202	0.1306	0.1691	0.4249
Video-LLaMA (LLaMA-2 7B) [1] (5 Ens)	0.9815	0.4444	0.6222	0.2120	0.1210	0.0230	0.1667	0.3354
Video-LLaMA (LLaMA-2 7B) [1]	0.9778	0.4444	0.6222	0.2321	0.1222	0.0239	0.1667	0.3428
Valley (LLaMA-1 13B) [2] (5 Ens)	0.8704	0.5170	0.6259	0.3951	0.6235	0.5051	0.1815	0.6182
Valley (LLaMA-1 13B) [2]	0.8000	0.5170	0.6241	0.3898	0.6160	0.5046	0.1901	0.6108
Valley (LLaMA-2 7B) [2] (5 Ens)	0.5630	0.5630	0.3704	0.2974	0.2465	0.7167	0.6593	0.5960
Valley (LLaMA-2 7B) [2]	0.5667	0.5541	0.3852	0.3256	0.2757	0.6955	0.6642	0.5865
OTTER (LLaMA-1 7B) [3] (5 Ens)	0.9481	0.5600	0.6241	0.2388	0.1280	0.0358	0.2963	0.3852
OTTER (LLaMA-1 7B) [3]	0.9481	0.9481	0.6241	0.2388	0.1280	0.0358	0.2963	0.3852
MiniGPT-4 (LLaMA-2 7B) [4] (5 Ens)	0.6777	0.6533	0.4870	0.7135	0.8349	0.6799	0.7851	0.6707
MiniGPT-4 (LLaMA-2 7B) [4]	0.6852	0.6519	0.4796	0.6832	0.7720	0.6313	0.7037	0.6498
InstructBLIP (Vicuna-V1 13B) [5] (5 Ens)	0.7111	0.5896	0.6815	0.6554	0.8128	0.6812	0.3210	0.5872
InstructBLIP (Vicuna-V1 13B) [5]	0.7407	0.5807	0.6648	0.6236	0.7844	0.6221	0.3185	0.5798
InstructBLIP (Vicuna-V1 7B) [5] (5 Ens)	0.8074	0.5481	0.6500	0.5580	0.6099	0.2435	0.4407	0.5428
InstructBLIP (Vicuna-V1 7B) [5]	0.7851	0.5303	0.6388	0.5492	0.6115	0.2662	0.4395	0.5494

Table 9. **Holistic Experiment (Binary Approach)**: Comparing popular multi-modal LLMs across the JRDB-Social at three levels in accuracy for the test set. BPC = Engagement of Body Position's connection with the Content, SSC = Salient Scene Content, (5 Ens) = Five Ensemble Strategy.

		Accuracy					**F1 Score**				
Resize	**Pad Scale**	**Venue**	**Purpose**	**BPC**	**SSC**	**Avg.**	**Venue**	**Purpose**	**BPC**	**SSC**	**Avg.**
Frame-level	1	0.1148	0.8199	0.5227	0.7322	0.5474	0.0686	0.2357	0.1818	0.0671	0.1383
	1.2	0.0868	0.8211	0.5657	0.7452	0.5547	0.0553	0.2234	0.1997	0.0677	0.1365
	1.4	0.0952	0.8173	0.5657	0.7195	0.5494	0.0664	0.2058	0.1761	0.0603	0.1271
	1.6	0.0812	0.8290	0.5884	0.7801	0.5697	0.0420	0.2272	0.2027	0.0757	0.1369
	1.8	0.0700	0.8196	0.6086	0.7455	0.5609	0.0335	0.2258	0.2003	0.0729	0.1331
	2	0.0868	0.8214	0.6162	0.7731	0.5744	0.0456	0.2236	0.1693	0.0706	0.1273
	2.5	0.0896	0.8313	0.5985	0.7931	0.5781	0.0472	0.2850	0.1803	0.0685	0.1452
	3	0.0952	0.8296	0.6086	0.7650	0.5746	0.0677	0.2399	0.1828	0.0721	0.1406
Fixed Black Mask	1	0.0840	0.7954	0.5631	0.6958	0.5346	0.0631	0.1972	0.1285	0.0573	0.1115
	1.2	0.0840	0.8179	0.5152	0.6610	0.5195	0.0434	0.2318	0.1007	0.0558	0.1079
	1.4	0.0616	0.8144	0.5606	0.6702	0.5267	0.0304	0.1890	0.1890	0.0640	0.1181
	1.6	0.0700	0.8232	0.5707	0.6787	0.5356	0.0399	0.2390	0.1666	0.0573	0.1257
	1.8	0.0924	0.8243	0.5960	0.6913	0.551	0.0506	0.2324	0.1797	0.0645	0.1318
	2	0.0952	0.8091	0.5682	0.6726	0.5363	0.0727	0.2363	0.1889	0.0651	0.1408
	2.5	0.0728	0.8270	0.5909	0.6358	0.5316	0.0408	0.2169	0.1545	0.0528	0.1162
	3	0.0560	0.8068	0.6035	0.6725	0.5347	0.0300	0.2159	0.1712	0.0610	0.1195
Fixed W/O Mask	1	0.0728	0.8223	0.5606	0.7251	0.5452	0.0399	0.2323	0.1486	0.0655	0.1216
	1.2	0.0924	0.8179	0.6212	0.7443	0.5690	0.0465	0.2164	0.1568	0.0684	0.1220
	1.4	0.0756	0.8270	0.6010	0.7939	0.5744	0.0417	0.2673	0.1894	0.0782	0.1442
	1.6	0.1036	0.8351	0.5884	0.7140	0.5603	0.0594	0.2316	0.1619	0.0685	0.1304
	1.8	0.0952	0.8141	0.6263	0.7036	0.5598	0.0485	0.2282	0.1677	0.0618	0.1265
	2	0.1008	0.8199	0.6010	0.7181	0.5600	0.0502	0.2123	0.1684	0.0679	0.1247
	2.5	0.0980	0.8346	0.5859	0.7048	0.5558	0.0676	0.2687	0.2067	0.0635	0.1516
	3	0.1036	0.8188	0.6187	0.7458	0.5819	0.0511	0.2211	0.2427	0.0626	0.1444

Table 10. Exploring diverse cropping scales at the group level: The left side of the table presents results in Accuracy, while the right side illustrates results in F1 score..

References

[1] H. Zhang, X. Li, L. Bing, Video-llama: An instruction-tuned audio-visual language model for video understanding, arXiv preprint arXiv:2306.02858 (2023). 5, 6, 7, 8, 9

[2] R. Luo, Z. Zhao, M. Yang, J. Dong, M. Qiu, P. Lu, T. Wang, Z. Wei, Valley: Video assistant with large language model enhanced ability, arXiv preprint arXiv:2306.07207 (2023). 5, 6, 7, 8, 9

[3] B. Li, Y. Zhang, L. Chen, J. Wang, J. Yang, Z. Liu, Otter: A multi-modal model with in-context instruction tuning, arXiv preprint arXiv:2305.03726 (2023). 5, 6, 7, 8, 9

[4] D. Zhu, J. Chen, X. Shen, X. Li, M. Elhoseiny, Minigpt-4: Enhancing vision-language understanding with advanced large language models, arXiv preprint arXiv:2304.10592 (2023). 5, 6, 7, 8, 9

[5] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, S. Hoi, Instructblip: Towards general-purpose vision-language models with instruction tuning (2023). `arXiv:2305.06500`. 5, 6, 7, 8, 9