Rank2Tell: A Multimodal Driving Dataset for Joint Importance Ranking and Reasoning: Supplementary Materials

Enna Sachdeva*[1], Nakul Agarwal*[1], Suhas Chundi[2], Sean Roelofs[2], Jiachen Li[2], Mykel Kochenderfer[2], Chiho Choi[1,†], and Behzad Dariush[1]

*equal contribution

[1]Honda Research Institute USA [2]Stanford University

1. Rank2Tell Dataset

1.1. Example Scenarios

We show several frames from different scenarios of our introduced dataset-Rank2Tell, in Fig. 1, 2, and 3 with intentions of going straight, turning left, or turning right at the intersection. These frames are extracted from the scenarios, while the ego vehicle is entering the intersection. As shown, the ego vehicle's intention (straight, right or left) and speed (in mph) are displayed on the frames. Moreover, Fig. 1, 2, and 3 show the annotators responses to various close and open-ended visual questions in the form of 4W+1H (What, Which, Where, Why, How). Further, we take the mode of the various importance levels from different annotators as the final importance level. An agent with $High$, $Medium$, and Low as the final importance level is shown in red, $yellow$ and $green$ color.

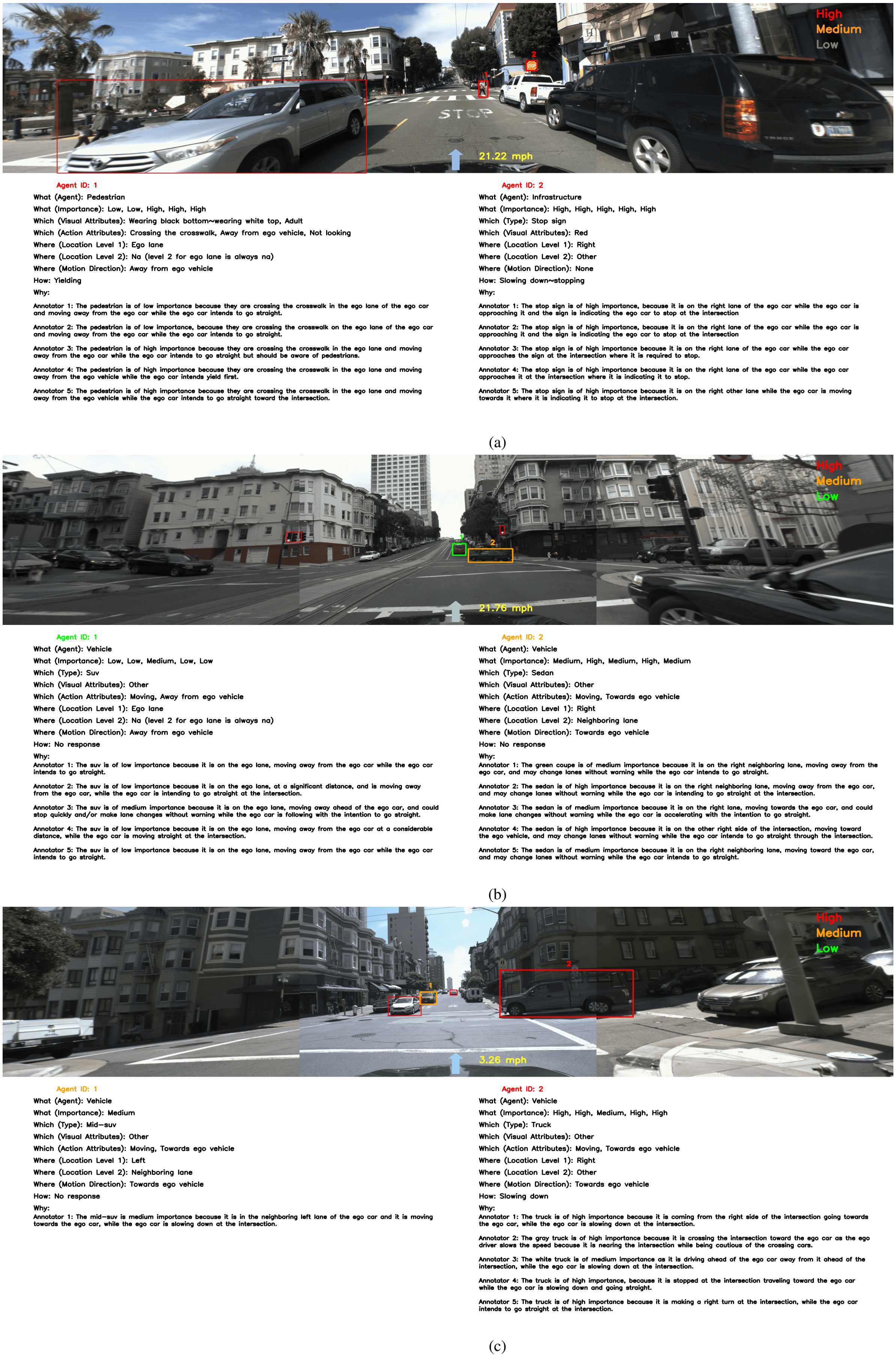
The video corresponding to these scenarios is in the attached supplementary video. The video provides a comprehensive overview of the different scenarios and the corresponding annotations across time frames, offering a valuable insights in understanding the development of importance levels in complex driving situations.

In Fig. 1a, we observe that the ego vehicle's intention is to go straight. All five annotators deem the pedestrian crossing the cross-walk and the stop sign to be important, with the majority of them assigning a high importance level. The visual attributes for the pedestrian, such as color of the dress (top, bottom) and age, are captured in the *Which* category. Similarly, the color of the stop sign is annotated as the visual attribute for infrastructure. The *Where (action attributes)* section shows the pedestrian's action in the current frame. Furthermore, the *Where (location level-1, location level-2, motion direction)* category captures the pedestrian's location and motion direction according to the annotation schema. The *How* section records the ego vehicle's response to the important agent. As the pedestrian crosses the cross-walk, the ego vehicle decides to yield. Since the annotators can view the previous four seconds of the video, they can understand the yielding behavior based on the change in speed. Additionally, the *Why* category shows the diverse reasons provided by the annotators based on their selected importance level for the pedestrian. The annotations effectively capture the responses of *3W + 1H (What, Which, Where, How)* to generate a caption for the response to the *Why*. These annotations demonstrate the usefulness of these structured visual question answering methods in perceiving risks and reasoning about them.

Moreover, in Fig. 1b, we see that the vehicle exiting the intersection from the right side is considered of medium importance by the majority of annotators, as it could influence the ego's behavior if it decides to make a lane change. Similarly, in Fig. 1c, the vehicle exiting the intersection from the right side is considered of high importance by the majority of annotators, as the front end of the vehicle is dangerously close to the ego vehicle, and the latter is slowing down to avoid a collision.

In Fig. 2a, where ego intends to turn right at the intersection, the annotators assigned $High$ and $Medium$ importance levels to the pedestrians crossing the cross-walk on the right of the ego-vehicle. Specifically, for the pedestrian with ID 1 is marked as important by 4 out of 5 annotators, while that with ID 2 is marked as important by only 1 annotator. The difference in perceived importance levels comes from various factors such as annotator's driving experience, age, gender, etc. The captions provided by annotators for agent with ID 1, show consistency in reasoning. They all capture the location, motion direction and the intention of the ego vehicle to reason about the importance of an agent. In Fig. 2b, the majority of annotators mark all pedestrians as $High$ important. This could be because of the close proximity of the pedestrians, as compared to the pedestrians in Fig. 2a. Further, Fig 2c shows that 3 and 5 annotators marked the truck and the crossing pedestrian of $High$ importance, because they directly influence the ego's decision.

[†]Now at Samsung Semiconductor Inc.



Agent ID: 1

What (Agent): Pedestrian

What (Importance): Low, Low, High, High, High

Which (Visual Attributes): Wearing black bottom~wearing white top, Adult

Which (Action Attributes): Crossing the crosswalk, Away from ego vehicle, Not looking

Where (Location Level 1): Ego lane

Where (Location Level 2): Na (level 2 for ego lane is always na)

Where (Motion Direction): Away from ego vehicle

How: Yielding

Why:

Annotator 1: The pedestrian is of low importance because they are crossing the crosswalk in the ego lane of the ego car and moving away from the ego car while the ego car intends to go straight.

Annotator 2: The pedestrian is of low importance, because they are crossing the crosswalk on the ego lane of the ego car and moving away from the ego car while the ego car intends to go straight.

Annotator 3: The pedestrian is of high importance because they are crossing the crosswalk in the ego lane and moving away from the ego car while the ego car intends to go straight but should be aware of pedestrians.

Annotator 4: The pedestrian is of high importance because they are crossing the crosswalk in the ego lane and moving away from the ego vehicle while the ego car intends yield first.

Annotator 5: The pedestrian is of high importance because they are crossing the crosswalk in the ego lane and moving away from the ego vehicle while the ego car intends to go straight toward the intersection.

Agent ID: 2

What (Agent): Infrastructure

What (Importance): High, High, High, High, High

Which (Type): Stop sign

Which (Visual Attributes): Red

Where (Location Level 1): Right

Where (Location Level 2): Other

Where (Motion Direction): None

How: Slowing down~stopping

Why:

Annotator 1: The stop sign is of high importance, because it is on the right lane of the ego car while the ego car is approaching it and the sign is indicating the ego car to stop at the intersection

Annotator 2: The stop sign is of high importance, because it is on the right lane of the ego car while the ego car is approaching it and the sign is indicating the ego car to stop at the intersection

Annotator 3: The stop sign is of high importance because it is on the right lane of the ego car while the ego car approaches the sign at the intersection where it is required to stop.

Annotator 4: The stop sign is of high importance because it is on the right lane of the ego car while the ego car approaches it at the intersection where it is indicating it to stop.

Annotator 5: The stop sign is of high importance because it is on the right other lane while the ego car is moving towards it where it is indicating it to stop at the intersection.

(a)

Agent ID: 1

What (Agent): Vehicle

What (Importance): Low, Low, Medium, Low, Low

Which (Type): Suv

Which (Visual Attributes): Other

Which (Action Attributes): Moving, Away from ego vehicle

Where (Location Level 1): Ego lane

Where (Location Level 2): Na (level 2 for ego lane is always na)

Where (Motion Direction): Away from ego vehicle

How: No response

Why:

Annotator 1: The suv is of low importance because it is on the ego lane, moving away from the ego car while the ego car intends to go straight.

Annotator 2: The suv is of low importance because it is on the ego lane, at a significant distance, and is moving away from the ego car, while the ego car is intending to go straight at the intersection.

Annotator 3: The suv is of medium importance because it is on the ego lane, moving away ahead of the ego car, and could stop quickly and/or make lane changes without warning while the ego car is following with the intention to go straight.

Annotator 4: The suv is of low importance because it is on the ego lane, moving away from the ego car at a considerable distance, while the ego car is moving straight at the intersection.

Annotator 5: The suv is of low importance because it is on the ego lane, moving away from the ego car while the ego car intends to go straight.

Agent ID: 2

What (Agent): Vehicle

What (Importance): Medium, High, Medium, High, Medium

Which (Type): Sedan

Which (Visual Attributes): Other

Which (Action Attributes): Moving, Towards ego vehicle

Where (Location Level 1): Right

Where (Location Level 2): Neighboring lane

Where (Motion Direction): Towards ego vehicle

How: No response

Why:

Annotator 1: The green coupe is of medium importance because it is on the right neighboring lane, moving away from the ego car, and may change lanes without warning while the ego car intends to go straight.

Annotator 2: The sedan is of high importance because it is on the right neighboring lane, moving away from the ego car, and may change lanes without warning while the ego car is intending to go straight at the intersection.

Annotator 3: The sedan is of medium importance because it is on the right lane, moving towards the ego car, and could make lane changes without warning while the ego car is accelerating with the intention to go straight.

Annotator 4: The sedan is of high importance because it is on the other right side of the intersection, moving toward the ego vehicle, and may change lanes without warning while the ego car intends to go straight through the intersection.

Annotator 5: The sedan is of medium importance because it is on the right neighboring lane, moving toward the ego car, and may change lanes without warning while the ego car intends to go straight.

(b)

Agent ID: 1

What (Agent): Vehicle

What (Importance): Medium

Which (Type): Mid—suv

Which (Visual Attributes): Other

Which (Action Attributes): Moving, Towards ego vehicle

Where (Location Level 1): Left

Where (Location Level 2): Neighboring lane

Where (Motion Direction): Towards ego vehicle

How: No response

Why:

Annotator 1: The mid—suv is medium importance because it is in the neighboring left lane of the ego car and it is moving towards the ego car, while the ego car is slowing down at the intersection.

Agent ID: 2

What (Agent): Vehicle

What (Importance): High, High, Medium, High, High

Which (Type): Truck

Which (Visual Attributes): Other

Which (Action Attributes): Moving, Towards ego vehicle

Where (Location Level 1): Right

Where (Location Level 2): Other

Where (Motion Direction): Towards ego vehicle

How: Slowing down

Why:

Annotator 1: The truck is of high importance because it is coming from the right side of the intersection going towards the ego car, while the ego car is slowing down at the intersection.

Annotator 2: The gray truck is of high importance because it is crossing the intersection toward the ego car as the ego driver slows the speed because it is nearing the intersection while being cautious of the crossing cars.

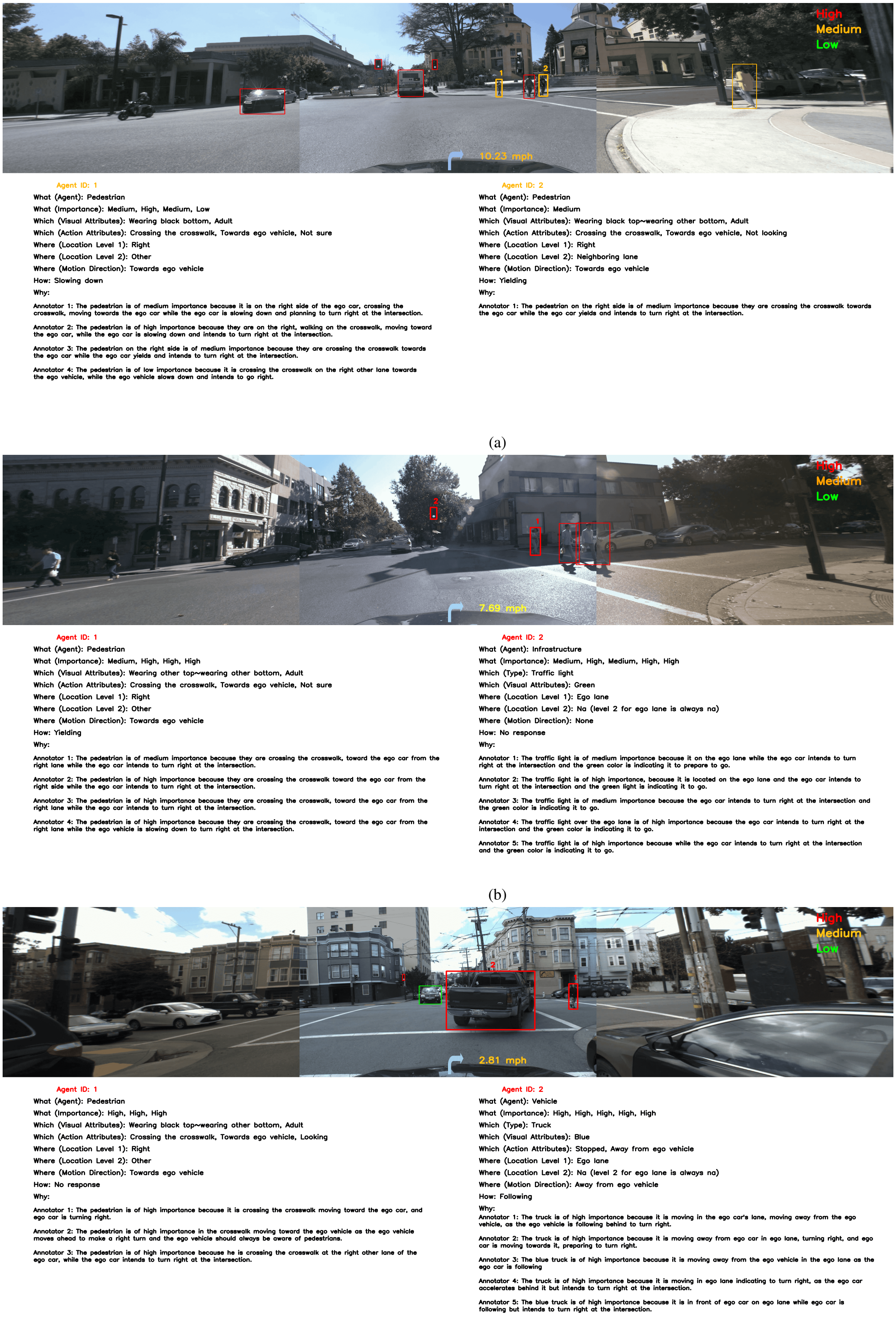
Annotator 3: The white truck is of medium importance as it is driving ahead of the ego car away from it ahead of the intersection, while the ego car is slowing down at the intersection.

Annotator 4: The truck is of high importance, because it is stopped at the intersection traveling toward the ego car while the ego car is slowing down and going straight.

Annotator 5: The truck is of high importance because it is making a right turn at the intersection, while the ego car intends to go straight at the intersection.

(c)

Figure 1. **Example scenarios of Rank2Tell with ego intention to go straight at the intersection**



Agent ID: 1

What (Agent): Pedestrian
What (Importance): Medium, High, Medium, Low
Which (Visual Attributes): Wearing black bottom, Adult
Which (Action Attributes): Crossing the crosswalk, Towards ego vehicle, Not sure
Where (Location Level 1): Right
Where (Location Level 2): Other
Where (Motion Direction): Towards ego vehicle
How: Slowing down
Why:

Annotator 1: The pedestrian is of medium importance because it is on the right side of the ego car, crossing the crosswalk, moving towards the ego car while the ego car is slowing down and planning to turn right at the intersection.

Annotator 2: The pedestrian is of high importance because they are on the right, walking on the crosswalk, moving toward the ego car, while the ego car is slowing down and intends to turn right at the intersection.

Annotator 3: The pedestrian on the right side is of medium importance because they are crossing the crosswalk towards the ego car while the ego car yields and intends to turn right at the intersection.

Annotator 4: The pedestrian is of low importance because it is crossing the crosswalk on the right other lane towards the ego vehicle, while the ego vehicle slows down and intends to go right.

Agent ID: 2

What (Agent): Pedestrian
What (Importance): Medium
Which (Visual Attributes): Wearing black top~wearing other bottom, Adult
Which (Action Attributes): Crossing the crosswalk, Towards ego vehicle, Not looking
Where (Location Level 1): Right
Where (Location Level 2): Neighboring lane
Where (Motion Direction): Towards ego vehicle
How: Yielding
Why:

Annotator 1: The pedestrian on the right side is of medium importance because they are crossing the crosswalk towards the ego car while the ego car yields and intends to turn right at the intersection.

(a)

Agent ID: 1

What (Agent): Pedestrian
What (Importance): Medium, High, High, High
Which (Visual Attributes): Wearing other top~wearing other bottom, Adult
Which (Action Attributes): Crossing the crosswalk, Towards ego vehicle, Not sure
Where (Location Level 1): Right
Where (Location Level 2): Other
Where (Motion Direction): Towards ego vehicle
How: Yielding
Why:

Annotator 1: The pedestrian is of medium importance because they are crossing the crosswalk, toward the ego car from the right lane while the ego car intends to turn right at the intersection.

Annotator 2: The pedestrian is of high importance because they are crossing the crosswalk toward the ego car from the right side while the ego car intends to turn right at the intersection.

Annotator 3: The pedestrian is of high importance because they are crossing the crosswalk, toward the ego car from the right lane while the ego car intends to turn right at the intersection.

Annotator 4: The pedestrian is of high importance because they are crossing the crosswalk, toward the ego car from the right lane while the ego vehicle is slowing down to turn right at the intersection.

Agent ID: 2

What (Agent): Infrastructure
What (Importance): Medium, High, Medium, High, High
Which (Type): Traffic light
Which (Visual Attributes): Green
Where (Location Level 1): Ego lane
Where (Location Level 2): Na (level 2 for ego lane is always na)
Where (Motion Direction): None
How: No response
Why:

Annotator 1: The traffic light is of medium importance because it on the ego lane while the ego car intends to turn right at the intersection and the green color is indicating it to prepare to go.

Annotator 2: The traffic light is of high importance, because it is located on the ego lane and the ego car intends to turn right at the intersection and the green light is indicating it to go.

Annotator 3: The traffic light is of medium importance because the ego car intends to turn right at the intersection and the green color is indicating it to go.

Annotator 4: The traffic light over the ego lane is of high importance because the ego car intends to turn right at the intersection and the green color is indicating it to go.

Annotator 5: The traffic light is of high importance because while the ego car intends to turn right at the intersection and the green color is indicating it to go.

(b)

Agent ID: 1

What (Agent): Pedestrian
What (Importance): High, High, High
Which (Visual Attributes): Wearing black top~wearing other bottom, Adult
Which (Action Attributes): Crossing the crosswalk, Towards ego vehicle, Looking
Where (Location Level 1): Right
Where (Location Level 2): Other
Where (Motion Direction): Towards ego vehicle
How: No response
Why:

Annotator 1: The pedestrian is of high importance because it is crossing the crosswalk moving toward the ego car, and ego car is turning right.

Annotator 2: The pedestrian is of high importance in the crosswalk moving toward the ego vehicle as the ego vehicle moves ahead to make a right turn and the ego vehicle should always be aware of pedestrians.

Annotator 3: The pedestrian is of high importance because he is crossing the crosswalk at the right other lane of the ego car, while the ego car intends to turn right at the intersection.

Agent ID: 2

What (Agent): Vehicle
What (Importance): High, High, High, High, High
Which (Type): Truck
Which (Visual Attributes): Blue
Which (Action Attributes): Stopped, Away from ego vehicle
Where (Location Level 1): Ego lane
Where (Location Level 2): Na (level 2 for ego lane is always na)
Where (Motion Direction): Away from ego vehicle
How: Following
Why:

Annotator 1: The truck is of high importance because it is moving in the ego car's lane, moving away from the ego vehicle, as the ego vehicle is following behind to turn right.

Annotator 2: The truck is of high importance because it is moving away from ego car in ego lane, turning right, and ego car is moving towards it, preparing to turn right.

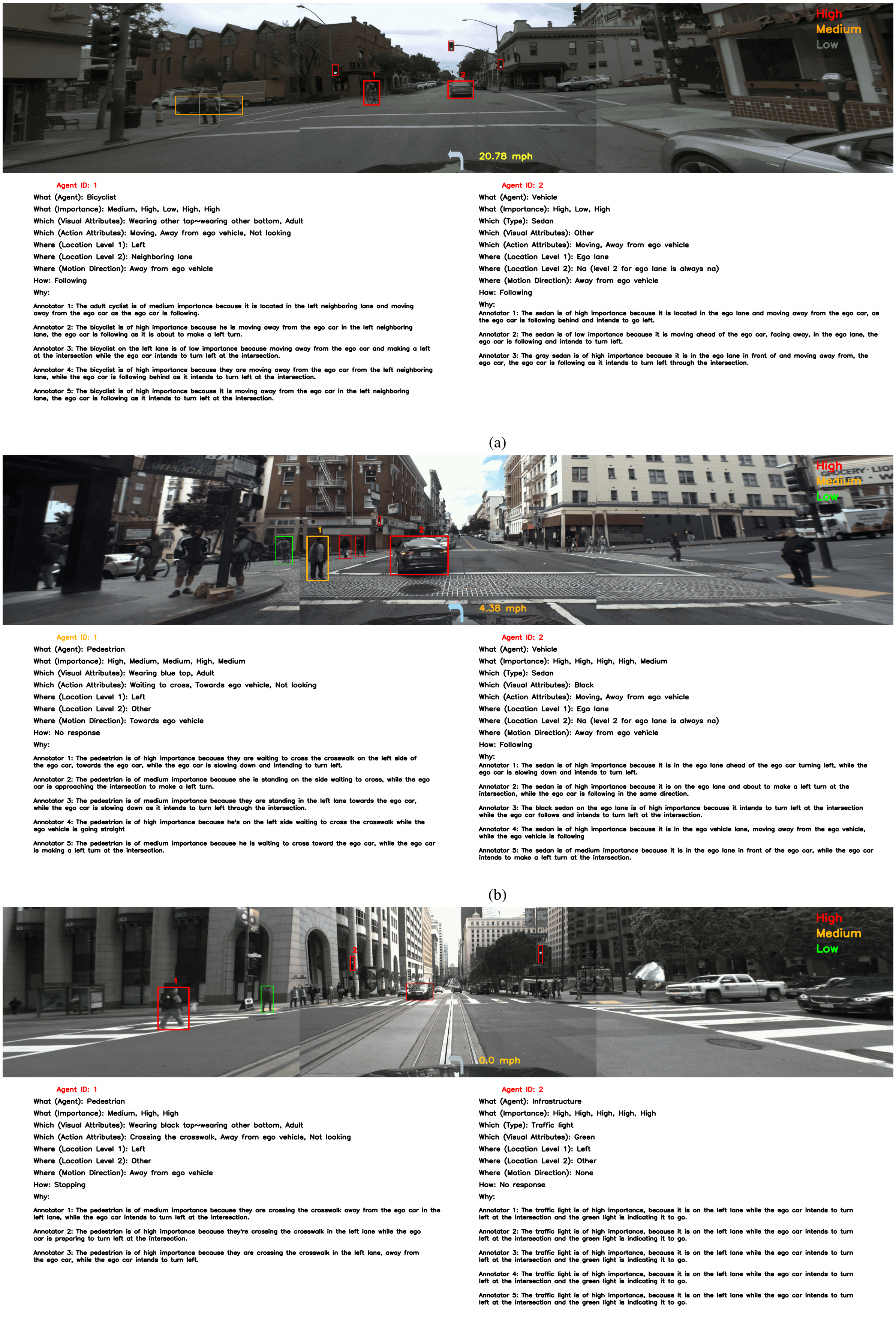
Annotator 3: The blue truck is of high importance because it is moving away from the ego vehicle in the ego lane as the ego car is following

Annotator 4: The truck is of high importance because it is moving in ego lane indicating to turn right, as the ego car accelerates behind it but intends to turn right at the intersection.

Annotator 5: The blue truck is of high importance because it is in front of ego car on ego lane while ego car is following but intends to turn right at the intersection.

(c)

Figure 2. **Example scenarios of Rank2Tell with ego intention to turn right at the intersection**



Agent ID: 1

What (Agent): Bicyclist

What (Importance): Medium, High, Low, High, High

Which (Visual Attributes): Wearing other top~wearing other bottom, Adult

Which (Action Attributes): Moving, Away from ego vehicle, Not looking

Where (Location Level 1): Left

Where (Location Level 2): Neighboring lane

Where (Motion Direction): Away from ego vehicle

How: Following

Why:

Annotator 1: The adult cyclist is of medium importance because it is located in the left neighboring lane and moving away from the ego car as the ego car is following.

Annotator 2: The bicyclist is of high importance because he is moving away from the ego car in the left neighboring lane, the ego car is following as it is about to make a left turn.

Annotator 3: The bicyclist on the left lane is of low importance because moving away from the ego car and making a left at the intersection while the ego car intends to turn left at the intersection.

Annotator 4: The bicyclist is of high importance because they are moving away from the ego car from the left neighboring lane, while the ego car is following behind as it intends to turn left at the intersection.

Annotator 5: The bicyclist is of high importance because it is moving away from the ego car in the left neighboring lane, the ego car is following as it intends to turn left at the intersection.

Agent ID: 2

What (Agent): Vehicle

What (Importance): High, Low, High

Which (Type): Sedan

Which (Visual Attributes): Other

Which (Action Attributes): Moving, Away from ego vehicle

Where (Location Level 1): Ego lane

Where (Location Level 2): Na (level 2 for ego lane is always na)

Where (Motion Direction): Away from ego vehicle

How: Following

Why:

Annotator 1: The sedan is of high importance because it is located in the ego lane and moving away from the ego car, as the ego car is following behind and intends to go left.

Annotator 2: The sedan is of low importance because it is moving ahead of the ego car, facing away, in the ego lane, the ego car is following and intends to turn left.

Annotator 3: The gray sedan is of high importance because it is in the ego lane in front of and moving away from, the ego car, the ego car is following as it intends to turn left through the intersection.

(a)

Agent ID: 1

What (Agent): Pedestrian

What (Importance): High, Medium, Medium, High, Medium

Which (Visual Attributes): Wearing blue top, Adult

Which (Action Attributes): Waiting to cross, Towards ego vehicle, Not looking

Where (Location Level 1): Left

Where (Location Level 2): Other

Where (Motion Direction): Towards ego vehicle

How: No response

Why:

Annotator 1: The pedestrian is of high importance because they are waiting to cross the crosswalk on the left side of the ego car, towards the ego car, while the ego car is slowing down and intending to turn left.

Annotator 2: The pedestrian is of medium importance because she is standing on the side waiting to cross, while the ego car is approaching the intersection to make a left turn.

Annotator 3: The pedestrian is of medium importance because they are standing in the left lane towards the ego car, while the ego car is slowing down as it intends to turn left through the intersection.

Annotator 4: The pedestrian is of high importance because he's on the left side waiting to cross the crosswalk while the ego vehicle is going straight

Annotator 5: The pedestrian is of medium importance because he is waiting to cross toward the ego car, while the ego car is making a left turn at the intersection.

Agent ID: 2

What (Agent): Vehicle

What (Importance): High, High, High, High, Medium

Which (Type): Sedan

Which (Visual Attributes): Black

Which (Action Attributes): Moving, Away from ego vehicle

Where (Location Level 1): Ego lane

Where (Location Level 2): Na (level 2 for ego lane is always na)

Where (Motion Direction): Away from ego vehicle

How: Following

Why:

Annotator 1: The sedan is of high importance because it is in the ego lane ahead of the ego car turning left, while the ego car is slowing down and intends to turn left.

Annotator 2: The sedan is of high importance because it is on the ego lane and about to make a left turn at the intersection, while the ego car is following in the same direction.

Annotator 3: The black sedan on the ego lane is of high importance because it intends to turn left at the intersection while the ego car follows and intends to turn left at the intersection.

Annotator 4: The sedan is of high importance because it is in the ego vehicle lane, moving away from the ego vehicle, while the ego vehicle is following

Annotator 5: The sedan is of medium importance because it is in the ego lane in front of the ego car, while the ego car intends to make a left turn at the intersection.

(b)

Agent ID: 1

What (Agent): Pedestrian

What (Importance): Medium, High, High

Which (Visual Attributes): Wearing black top~wearing other bottom, Adult

Which (Action Attributes): Crossing the crosswalk, Away from ego vehicle, Not looking

Where (Location Level 1): Left

Where (Location Level 2): Other

Where (Motion Direction): Away from ego vehicle

How: Stopping

Why:

Annotator 1: The pedestrian is of medium importance because they are crossing the crosswalk away from the ego car in the left lane, while the ego car intends to turn left at the intersection.

Annotator 2: The pedestrian is of high importance because they're crossing the crosswalk in the left lane while the ego car is preparing to turn left at the intersection.

Annotator 3: The pedestrian is of high importance because they are crossing the crosswalk in the left lane, away from the ego car, while the ego car intends to turn left.

Agent ID: 2

What (Agent): Infrastructure

What (Importance): High, High, High, High, High

Which (Type): Traffic light

Which (Visual Attributes): Green

Where (Location Level 1): Left

Where (Location Level 2): Other

Where (Motion Direction): None

How: No response

Why:

Annotator 1: The traffic light is of high importance, because it is on the left lane while the ego car intends to turn left at the intersection and the green light is indicating it to go.

Annotator 2: The traffic light is of high importance, because it is on the left lane while the ego car intends to turn left at the intersection and the green light is indicating it to go.

Annotator 3: The traffic light is of high importance, because it is on the left lane while the ego car intends to turn left at the intersection and the green light is indicating it to go.

Annotator 4: The traffic light is of high importance, because it is on the left lane while the ego car intends to turn left at the intersection and the green light is indicating it to go.

Annotator 5: The traffic light is of high importance, because it is on the left lane while the ego car intends to turn left at the intersection and the green light is indicating it to go.

(c)

Figure 3. **Example scenarios of Rank2Tell with ego intention to turn left at the intersection**

However, 2 annotators didn't mark the pedestrian to be of any importance. This discrepancy in annotations could be due to experienced drivers tending to consider only objects in close proximity to be of high importance, which is the truck in this case.

In Fig. 3a, the Sedan at the intersection, and the bicycle turning left at the intersection are marked as $High$ importance by 2 and 3 annotators respectively. In Fig. 3b and Fig 3c, the agents towards the left of the intersection are marked as important by various annotators.

These annotations demonstrate the effectiveness of the Rank2Tell dataset in capturing diverse responses of agents and infrastructure with various intentions of ego vehicles. The annotations show how annotators identified important visual attributes, actions, and locations of the agents in the scene, and how they reasoned about the importance of these agents based on their potential impact on the ego vehicle. The annotations also show that different annotators may assign different levels of importance to the same agents, based on their personal experience and perception of risk. This highlights the need for a diverse dataset like Rank2Tell to capture a range of perspectives and improve the robustness of machine learning models for autonomous driving.

These examples show that estimating importance level of various agents in the scene depends on various factors, and we aim to provide comprehensive annotations of those factors to address the problem of important agents classification and natural language explaination.

1.2. Dataset Analysis

1.2.1 Diverse Annotations

The diversity in Rank2Tell is one of its strengths, as it presents broad spectrum of diverse annotations. An agent's importance is annotated with the prior knowledge of ego vehicle's intention (left, right, straight) at the intersection. For instance, if the ego car intends to turn right while an agent is located on the left of the intersection, it is highly probable that none of the anno- tators will deem it as important. In case a more conservative annotator identifies it as important, users have the flexibility to disregard the outliers to calculate the groundtruth impor- tance, guided by various heuristics tailored to their specific use case of the dataset, since we plan to release all 5 anno- tators data with the release of the dataset. It also provides users the flexibility to consolidate 4 levels of importance to 2 levels(important, non-important), based on their use case.

1.2.2 Training, Validation and Testing Data

The Rank2Tell dataset consists of total of 116 recordings, and we split the dataset into thress subsets of 70, 23, 23 scenarios for training, validation and testing data, respectively. We use the same split across all the baselines experiments, and report the results on the validation data.

Table 1a shows the distribution of various agent types and their importance levels in the three data subsets. The table shows the number of scenarios, and total frames in the 3 data splits. Additionally, it shows the distribution of 4 different agents types (vehicles, bicyclists, pedestrians, infrastructure) and their annotated importance levels by different annotators. For instance, it shows that 5283 vehicles are being annotated as of $High$ importance in the training data, while 2604 and 1818 vehicles are being annotated as of $Medium$ and Low importance in the validation and testing data. This comprises 40.1%, 46.7%, and 38.7% of all important vehicles in the training, validation and testing data. Further, vehicles with importance level as $Medium$ consists of 38.7%, 37.5% and 40.5% of important vehicles in training, validation and testing data. And the vehicles with importance level as Low consists of 21.1%, 15.7% and 20.7% of all important vehicles in training, validation and testing data. We observe comparable distribution across other agent types, as shown in Table 1a. This distribution indicates that the data splits consists of comparable number of data samples for different importance levels. We see similar distribution across other agents types- bicyclists, pedestrians and infrastructures.

Table 1b shows the distribution of visual question answering for 3 data splits. The dataset provides 29311 importance levels responses from various annotators across 9305 unique important agents in the training data. It is important to note that the visual and action attributes of the agents consist of multiple sub-attributes associated with features such as color, age, and communicative and visual aspects. Therefore, the numerical values associated with these attributes are significantly higher than the number of unique important objects.

1.2.3 Annotators' data distributions

Figure 4 presents the relationship between different importance levels and annotators' personal information. The dataset employs 43 unique annotators; however, due to privacy concerns, not all annotators share their personal information. Thus, we have complete information for 35 annotators, while 7 annotators only provided their gender and age, and 1 annotator only provided their gender.

For the 116 scenarios in our dataset, the asked questions resulted in a total of 2900 data points, out of which we were able to collect 2585 (89.1% of the total). As shown in Figure 4, there is a strong correlation between the number of years of experience and importance ranking annotations. Additionally, we observe a similar trend with respect to age.

	Agents Importance													
Data	# Scenarios	# Frames	Vehicles			Bicyclists			Pedestrians			Infrastructures		
			High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low
Training	70	2364	5283	5101	2782	249	186	179	2997	1833	1540	8308	499	354
Validation	23	784	2604	2091	881	180	161	87	682	460	640	2556	110	47
Testing	23	685	1818	1905	973	62	145	122	416	437	218	2356	379	60

(a) Agents Importance level distribution for train, validation and test data split

	Visual Question Answering										
Data	All Agents	What		Which			Where			How	Why
	Important and Non-Important	Important Agents	Importance	Types	Visual Attributes	Action Attributes	Location Level-1	Location Level-2	Motion Direction	Ego's Response	Caption
Training	35839	9305	29311	22327	36295	47284	29311	29311	20150	29311	29311
Validation	11970	3285	10499	8289	12709	17782	10499	10499	7786	10499	10499
Testing	9719	2782	8891	7491	10291	13592	8891	8891	6096	8891	8891

(b) Distribution of Annotators' VQA responses for train, validation and test data split

Table 1. Annotators' data distribution

2. Implementation Details

2.1. Groundtruth Importance

Table 2 in the main paper highlights the consistency and quality of importance level across 5 annotators. This is calculated using mode of all 5 annotators importance levels as the final importance of an agent in the scene. This includes importance as well as non-importance levels. Nonetheless, when transitioning to practical use cases of not underestimating an agent's importance, we use groundtruth as the mode of only importance annotations, i.e if 2 of out 5 annotators deem an agent as important, we take the mode of only 2 annotators, and not all 5. In cases of multiple modes, we take the highest importance level as the groundtruth importance. Table 2 shows the consistency using this method for estimating the groundtruth importance. A consistency of 20%, 40%, 60%, 80%, 100% indicates that all five annotators deem an object as important, while only 1, 2, 3, 4, and 5 annotators, respectively, provided the same importance level as the mode. A consistency of 33.33% and 66.66% indicates that three annotators deem an object as important, while only 1 and 2 annotators, respectively, provided the same importance level as the mode. Lastly, the consistency of 50% indicates that if 2 annotators deem an object as important, both provided different importance, and if 4 annotators deem an object as importance, half (2) annotators provided the same importance level, and 75% indicates that 4 annotators deem an object as important, while only 3 provided the same importance level as the mode. It shows that for data instances with High, Low and Non-Important as groundtruth importance, 89.3%, 83.20% and 100% of samples exhibited more than 60% consistency. This suggests that annotators had a high level of agreement on objects marked as High, Low and Non-important. However, for data instances with Medium as the majority importance, only 75.65% of samples had more than 60% consistency. This could be attributed to the ambiguity between medium and high importance, and medium and low importance. Nonetheless, consistency score doesn't demonstrate the quality of the dataset, but shows the agreement/disagreement among annotators with a groundtruth as final importance.

	Agents Importance		
Consistency	Low	Medium	High
20	0	0	0
33.33	2.12	1.03	0.51
40	1.44	5.36	2.69
50	13.21	17.93	7.46
60	3.85	13.72	9.18
66.66	8.00	8.36	4.44
75	3.79	5.31	4.76
80	0.96	7.71	13.73
100	66.60	40.55	57.19
>=60	83.20	75.65	89.3

Table 2. Inter-Annotator consistency of importance annotation based on mode of only importance annotations

2.2. Joint Model

2.3. 2D Deep-Feature extraction

The 2D deep feature extractor uses a sequence of RGB images, depth images, semantic maps, ego vehicle's inertial data, and 2D bounding box of each objects in the scene. The depth images are obtained by projecting the pointcloud to the stitched camera view, and the segmentation maps are obtained using DeepLabv3 on the stitched RGB images. Visual feature extractor adopts ResNet101 pretrained on ImageNet dataset as the backbone. It takes in RGB images to generate object level features and global features of all frames, which are then fed into Sequence Encoder I to obtain image features $v_{2D_{(j,A)}}$ and $v_{global,A}$. Depth feature extractor adopts ResNet18 trained from scratch as a backbone, and takes in concatenated depth and semantic segmentation maps to generate object level and features and global features of all frames, which are fed into Sequence Encoder III to obtain final features, $v_{2D_{(j,DS)}}$ and $v_{global,DS}$. To extract the features of each object from $v_{2D_{(j,A)}}$ and $v_{2D_{(j,DS)}}$, a ROIAlign pooling layer is added before feeding to sequence encoder. The final visual feature of each object, and global context information is obtained by concatenation as $v_{2D_j} = [v_{2D_{(j,A)}}, v_{2D_{(j,DS)}}]$, $v_{global} = [v_{global,A}, v_{global,DS}]$. The ego vehicle features v_{ego} are obtained using Ego State Feature Encoder, and the bounding box features of other agents $v_{2D_{(j,B)}}$, are obtained using Sequence encoder II, as shown in Fig. 5 in the main paper.

2.4. 3D Deep-Feature extraction

We represent the pointcloud data as $\mathcal{P} = (p_i, f_i) \in \mathcal{R}^{N_p \times 6}$, where $p_i \in \mathcal{R}^3, i = 1, 2, ..., N_p$ are the coordinates, and $f_i \in \mathcal{R}^3$ are the additional features, representing radius, confidence and curvature. The point cloud features consist of $< \ x, y, z, radius, confidence, curvature \ >$ data points. Additionally, we utilize the 3D bounding boxes of agents in the scene, obtained from LIDAR annotations, as well as the instance and semantic labels associated with each point cloud. The output of voting module is a set of point clusters $\mathcal{C} \in \mathcal{R}^{M \times 128}$, where M is the upper bound of the number of proposals. We assume that the bounding boxes of objects (i.e., vehicles, infrastructure, pedestrians, and bicyclists) can be obtained using off-the-shelf detection and tracking system in advance. Therefore, we leverage the groundtruth 3D bounding boxes, and groundtruth objects coordinates information for the votenet model to obtain bounding boxes features $v_{3D_{j,P}}$.

The final object level features are obtained by concatenating the corresponding 2D features and 3D features: $v_j = [v_{2D_j}, v_{3D_j}]$

2.5. Relational Feature Extraction

The importance level ranking and reasoning of an object in a scene are influenced by the state and appearance of other objects in the scene. Thus, to capture the mutual influence and relations among objects, we use a graph neural networks that models objects as nodes and their relations as edges in a relational graph module. The module takes in the concatenated object features and extracts both object features v_j and relational features $e_{i,j}$ between objects.

By applying a message passing mechanism over the graph, the nodes can exchange information with their neighbors, allowing them to update their features based on the features of their neighbors. To model the relational (edges) features, the module considers only the K nearest objects surrounding the target object to limit the computation complexity.

$$v \rightarrow e : e_{ij} = f_e^1([v_i, v_j]), \tag{1}$$

$$e \rightarrow \bar{v} : v_i = f_v(\Sigma_{i \neq j} e_{(i,j)}), \tag{2}$$

The final objects features are obtained by concatenating the graph nodes features with the object relations, global features, ego features, and the ego intentions I_E: $o_j = [v_j, \bar{v}_j, v_{global}, v_{ego}, I_E]$

2.6. Importance Classification

The final object features is fed into the classifier to obtain its importance class (high, medium, low or binary). The classifier outputs the logits corresponding to different importance level as $y = [\hat{y_1}, ..., \widehat{y_{n-1}}]$.

2.7. Captioning Decoder

The captioning module takes the object features (o_j) to generate caption with one token at a time, using GRU. Similar to Scan2Cap [1], we add the relational features between object j and its neighbor k to the corresponding o_j to obtain the final attention context feature set $V^r = v_1^r, ..., v_j^r, ..., v_M^r$, where $v_j^r = o_j + \Sigma_{k=1}^M \rceil_{jk}$. The intermediate attention distribution over the context features are defined as:

$$\alpha_t := softmax((\mathcal{V}^r W_v + \mathbb{1}_h h_{t-1}^T W_h) W_a) \mathbb{1}_a \tag{3}$$

The attention module outputs the aggregated context vector $\hat{v_t} = \Sigma_{i=1}^M \mathcal{V}_i^r \odot \alpha_{ti}$ to represent the attended object and corresponding inter-object relation. The language GRUs uses the $\hat{v_t}$ and hidden state h_t^2 to predict the token y_t at current step t.

2.8. Training

The joint model predicts the importance of all agents within the scene - both important and not-important. However, the model solely generates captions of important agents. Subsequent to the Graph Neural Network (GNN) module, the agents deemed non-important are systematically filtered out. This filtration utilizes the ground truth importance level both during training and evaluation phases.

Such an approach ensures a fair comparison with other captioning baselines, which have been modified similarly to exclude non-important agents based on the ground truth importance level.

The training objective combines the caption loss and the importance classification loss in the following manner:

$$\mathcal{L} = \alpha\mathcal{L}_{imp} + \beta\mathcal{L}_{cap} \tag{4}$$

where α, β are the weights for the individual loss terms.

TO enforce the model to reduce the instances of falsely underestimating an agent's importance, we penalize the L_{imp} corresponding to different groundtruth (GT) and predictions (P) for different importance levels of high (H), medium(M), low(L), and not-importance(NI) as follows:

$$\mathcal{L}_{imp} = \Sigma_{i=1}^{N}\mathcal{L}_i \tag{5}$$

$$\mathcal{L}_i = \begin{cases} \lambda_k\mathcal{L}_i, & \text{if } (GT - P) = k > 0. \\ \mathcal{L}_i, & \text{otherwise.} \end{cases} \tag{6}$$

where $\mathcal{L}_i$ is the cross-entropy loss for each object i.

For 4 classes, levels 0, 1 2 and 3 represent $Non - Important$, Low, $Medium$ and $High$ importance, respectively. For binary classes, levels 0 and 1 represent $Non - Important$, and $Important$, respectively.

- Predicted Importance is 1 level lower than groundtruth importance: If groundruth is $High$, prediction is $Medium$, or groundtruth is $Medium$, prediction is Low, or groundtruth is Low, prediction is $Not - Important$ (for 4 classes), or groundtruth is $Important$, prediction is $Not - Important$ (for 2 classes), we weigh the loss by λ_1
- Predicted Importance is 2 level lower than groundtruth importance: If groundruth is $High$, prediction is Low, or groundtruth is $Medium$, prediction is $Not - Important$, we weigh the loss by λ_2
- Predicted Importance is 3 level lower than groundtruth importance: If groundruth is $High$, prediction is $Not - Important$, we weigh the loss by λ_3
- Predicted Importance is higher than groundtruth importance: groundruth is $Not - Important$, prediction is $High$, or $Medium$ or Low importance, or groundruth is Low, prediction is $High$, or $Medium$, groundtruth is $Medium$, prediction is $High$ (for 4 classes), or groundtruth is $Not - Important$, prediction is $Important$ (for 2 classes), we weigh the loss by 1

For joint model, we use the following values of parameters: $\lambda_1 = \lambda_2 = \lambda_3 = 1$, and $\alpha = \beta = 1$

Method	F1 (I)	F1 (NI)	Accuracy
Ours w/o actions	**78.44**	92.97	89.39
Ours w/ actions	78.27	**93.01**	**89.42**

Table 3. Quantitative Evaluation comparing the F1 scores for 2 importance levels of our joint model, with and without augmented action features. I: IMPORTANT, NI: NON-IMPORTANT

Method	C	B-4	M	R
Ours w/o actions	100.15	45.83	36.21	68.56
Ours w/ actions	**107.14**	**47.45**	**36.75**	**69.41**

Table 4. Quantitative Evaluation comparing the performance for captions predictions of our joint model, with and without augmented action features. C: CIDER, B-4: Bleu-4, M: Meteor, R: Rouge

2.9. Importance level classification

For a fair comparison with the joint model, all three baselines [2, 4, 7] use the same feature extractor ResNet101 [3] pre-trained on the ImageNet dataset with Feature Pyramid Networks [5] on top. During training, we adopt stochastic gradient descent with ADAM optimizer to learn the network parameters. The model is trained for 100 epochs using an initial learning rate of 0.0001 and a learning rate decay of 10. All feature layers are jointly updated during training. For consistency, we maintain a fixed input resolution of 5760 × 1200 and use a batch size of 32 for all experiments.

2.10. Captioning Baselines

For Scan2Cap [1] baseline, we use 3D pointcloud data with $< x, y, z, radius, confidence, curvature >$ as the pointcloud features. The point cloud features consist of $< x, y, z, radius, confidence, curvature >$ data points. Additionally, we utilize the 3D bounding boxes of agents in the scene, obtained from LIDAR annotations, as well as the instance and semantic labels associated with each point cloud. For a fair comparison with our proposed model, we assume that the bounding boxes of objects, such as vehicles, pedestrians, and bicyclists, can be obtained using off-the-shelf detection and tracking systems in advance. Consequently, we leverage the ground truth 3D bounding boxes and object instance information to train the votenet module to learn the bounding box features. This is a modified version of the Scan2Cap model that does not perform object detection and instead relies on the ground truth bounding boxes information to learn the bounding box features. We limit the maximum number of objects in a particular frame to 64. However, it is important to note that our current work does not include infrastructure due to the lack of available 3D pointcloud features for traffic lights, stop signs, speed limit signs, etc. To ensure a fair comparison, we do not include infrastructure for the S&T [6] baseline as well. Nev-

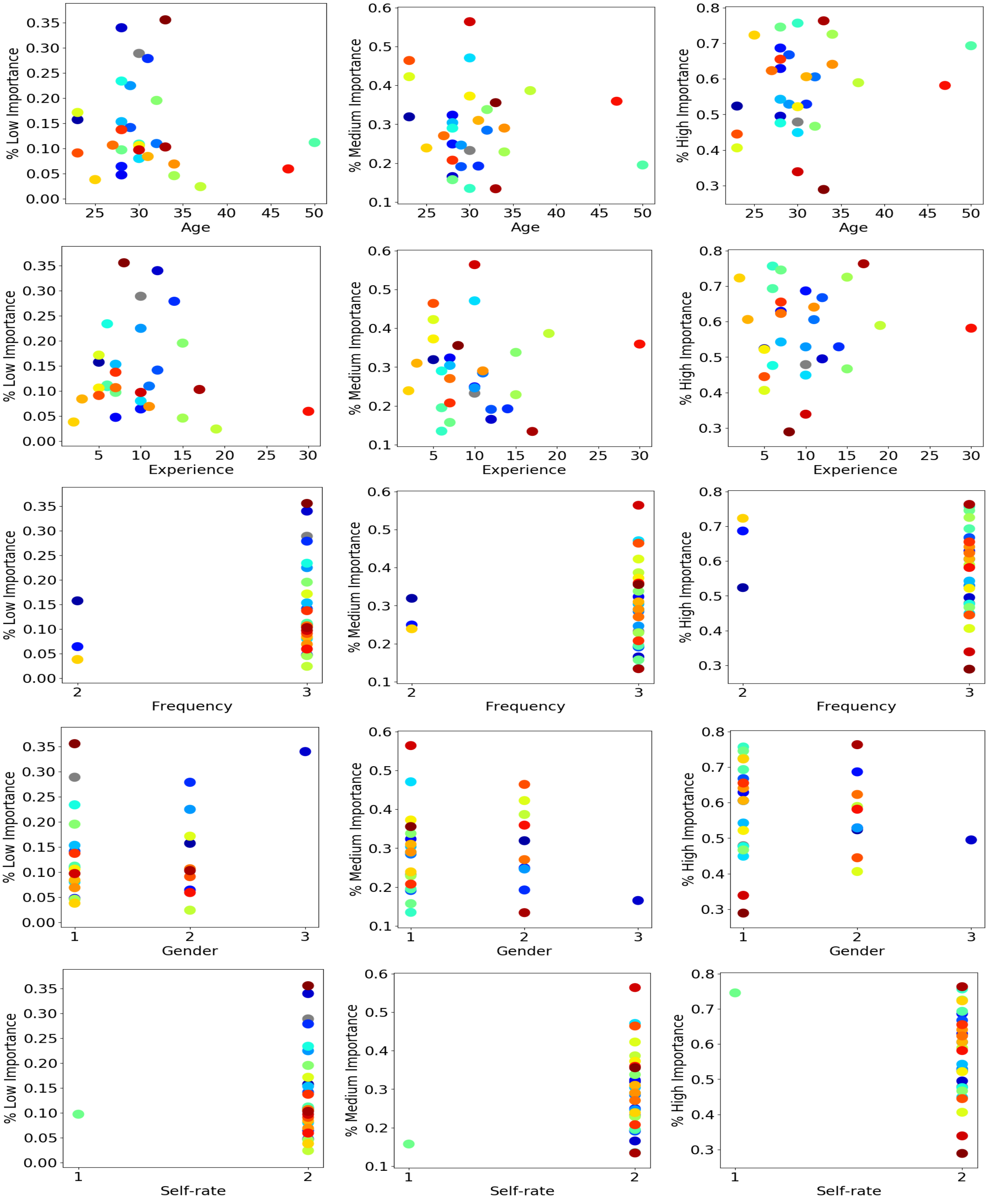


Figure 4. Relationship between importance level (grouped by columns) and annotator personal information (grouped by rows). Each annotator has been assigned a unique color, and is represented in each figure by a dot. From top row: (1) age in years, (2) driving experience in years, (3) frequency of driving, either 1-rarely, less than once a month, 2-occasionally, about once a week, 3-frequently, more than three times a week, (4) gender 1-male, 2-female, 3-not sure (5) rating of driving skill, 1-intermediate, 2-advanced.

ertheless, we plan to incorporate infrastructure in our future extensions of this work.

2.11. Integrating Action Attributes

To demonstrate the usefulness of these annotations, we conduct an ablation study where we integrate action attributes (Which) with object features in a simple way within the joint model. We concatenate one hot vector of action attributes with object features before the GNN, and the results in Table 3 and Table 4 show superior performance of the model as compared to those without action.

References

[1] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3193–3203, 2021. 7, 8

[2] Mingfei Gao, Ashish Tawari, and Sujitha Martin. Goal-oriented object importance estimation in on-road driving videos. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5509–5515. IEEE, 2019. 8

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 8

[4] Jiachen Li, Haiming Gang, Hengbo Ma, Masayoshi Tomizuka, and Chiho Choi. Important object identification with semi-supervised learning for autonomous driving. In *ICRA*, 2022. 8

[5] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 8

[6] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 8

[7] Zehua Zhang, Ashish Tawari, Sujitha Martin, and David Crandall. Interaction graphs for object importance estimation in on-road driving videos. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8920–8927. IEEE, 2020. 8