1. Annotation

1.1. Guide/framework

The selected method description depends on the specific situation, and it is up to the annotator to interpret. Considering the degrees of each feature, they must explain the scene with an accurate amount of detail to convey the necessary information, so that an AI system or a human can understand the intended meaning. An interpretation is added to ensure that the gesture is captured. However, the gesture should be possible to interpret solely from the description of body movements.

The applied annotation method is referenced as:

> Annotate sequences of movement with varying detailed descriptors, temporally aligned with gesture onset frames and incrementally updated to reflect new motion events. The transcription leverages prior contextual knowledge. Limit gestures explicitly directed toward the ego driver.

1.1.1 Caption Instruction

The instruction considered when captioning the ground truth descriptions is formulated as follows (examples can be found in Sub-section 1.1.2):

> Transcribe the pedestrians' upper-body posture and expressive gestures, specifying the intended recipient *(e.g., signaling the ego driver to stop, requesting another driver to pull over)*. Segment the annotation using start- and end-frames at the initiation and termination of a new meaning movement. For each relevant sub-part *(e.g., arm, finger, head)* describe its position, distance, speed, and direction relative to themselves *(e.g., at their side, facing 9 o'clock of themselves, at the dog)* and the ego driver *(e.g., towards the ego driver, far 10 o'clock of the ego driver)*. Follow up with an interpretation of the given gesture to understand what is being communicated.

1.1.2 Caption Examples

1. *"The pedestrian is standing close at 11 o'clock of the ego driver with both their torso and head facing the ego driver. Their hands are held flat at their chest, facing the ego driver, while they move back and forth towards the ego driver, gesturing for the ego driver to reverse."*

2. *"... They are facing you, shaking their head, indicating denial of driving permission."*

1.1.3 Classification

The gesture classification classes are:

#.	Gesture	Description
0.	Idle	No movement or gestures
1.	Transition	Initial or ascending gesture, but unclear
2.	Stop	Bring vehicle to a halt in any manner
3.	Advance	Drive forward in any manner
4.	Accelerate	Increase current speed
5.	Decelerate	Decrease current speed
6.	Reverse	Reverse the vehicle
7.	U-turn	Make an u-turn
8.	Pass	Straight pass
9.	Left	Turn to the left lane
10.	Right	Turn to the right lane
11.	Hail	Hail for a ride
12.	Point	Pointing in any manner
13.	Other	Non-navigation gesture (e.g., 'thanks')

Table 1. Gesture classes

It is insufficient to utilize 'Drive' as a word, as it can have multiple meanings. 'Pass' means drive across an intersection, 'Left' and 'Right' means turn, and 'Advance' means drive wherever.

1.2. Format

The annotation includes bounding boxes and IDs for each pedestrian, if there are multiple pedestrians in the scene. The gesture class only describes the gesture towards the ego driver, to simplify the concept for now. A description is delegated per pedestrian separately, still maintaining the overall picture in relation to other subjects. The caption annotations are supplemented with bounding-box data to link each pedestrian ID to its corresponding bounding box in every frame. See example in Tab. 2. The annotations contain the features: Name of video (`video_name`) , Name of camera (`camera_name`) , Start frame at movement (`start_frame`) , End frame at movement (`end_frame`) , Pedestrian ID (`pedestrian_id`) , Gesture class ID (`gesture_id`) , Body movement description (`body_desc`) , Interpretation description (`interpret_desc`) .

video	camera	start	end	ped	gest	body	interpret
`str`	`str`	`int`	`int`	`int`	`int`	`str`	`str`
video_04	front	41	56	0	2	*"..flat hand.."*	*"..stop.."*
video_04	front	57	63	0	0	*"..nothing.."*	*"..nothing.."*
video_04	front	45	56	1	1	*"..points.."*	*"..go there.."*

Table 2. Annotation format example including multiple pedestrians.