Datasets of Conflicting Pedestrian-to-Driver Gestures in Urban Traffic:
Gesture Formality, and Scene Complexity

1. Annotation

We explain the instructions to how the data is annotated including explanation to possible misunderstood annotations. A more in depth reasoning for this framework is argued in the discussion, Sub-section **??**.

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. For now, we only annotate gestures directed towards the ego driver. The remaining subjects can be filtered utilizing the bounding box and the information of the non-existent gesture annotation.

The applied annotation method is referenced as:

> Annotate sequences of upper-body movement with varying detailed descriptors. Mark each new movement with a start and end frame, and an individual caption that can be understood standalone. The caption can refer to previously acknowledged information from the scene. Annotations are limited to gestures explicitly directed toward the ego driver.

1.1.1 Caption Instruction

The instruction considered when captioning the ground truth descriptions of the body movement and interpretation is formulated as follows, inspired by [**?**] (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 initiating and terminating 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. In cases where a single subject makes multiple gestures, use the term '<*skip*>' to indicate gestures not directed towards the ego driver, for further annotation.

1.1.2 Caption Examples

The following examples provide an understanding of expressing the human body, its movements, relationships to other objects, and usage of <skip>. The examples are inspired by [**?**].

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."

3. "They're gesturing a flat hand towards the ego driver. <skip>" *(Remaining information towards other subjects is currently excluded from the annotation.)*

1.1.3 Classification

The classification is only towards the ego driver and ignores gestures not directed towards the ego driver. In cases of multiple gestures from a single subject, we classify the gesture directed towards the ego driver. Be aware of the potential for insufficient utilization of 'Drive', as it has multiple meanings. Instead, we use 'Pass' meaning drive across an intersection, 'Left' and 'Right' meaning turn, and 'Advance' meaning drive wherever. 'Idle' is not being used, since we only focus on direct gestures, but it is still included for clarity. The gesture classification classes are as follows, inspired by [**?**]:

#.	Gesture	Description
0.	Idle	No gestures
1.	Transition	Initial or ascending gesture
2.	Stop	Stopping in any manner
3.	Advance	Drive forward in any manner
4.	Return	Backup by reverse or turn the vehicle
5.	Accelerate	Increase current speed
6.	Decelerate	Decrease current speed
7.	Left	Turn to the left lane
8.	Right	Turn to the right lane
9.	Hail	Hail for a ride
10.	Attention	Seeking awareness
11.	Pointing	Pointing in any manner
12.	Other	Non-navigation gesture
13.	*Unlisted*	*Unlisted navigation gesture*
14.	*Unclear*	*Unknown or unclear*

Table 1. Gesture classes used to annotate directions towards the ego driver. It is not sufficient to utilize 'Drive' as a word, as it is too broad. 'Pass' means drive across, 'Left' and 'Right' mean turn, and 'Advance' means drive wherever. The aspect of *Pointing* is optimal in 3D space for understanding specific locations. For now, it is only seen as a class classification, not a particular location. 'U-turn' could have its own class, but the AV should understand that it should go back, in any manner it figures out to be the safest and fastest. The term 'Unlisted' is not supposed to be used, but it can be used in cases of forgotten navigation gestures. Instead of leaving the annotation blank, which could confuse the process, this term can be used to reanalyze. Similarly, 'Unclear' is used to avoid incorrect annotation by guessing. These samples can either be re-annotated, excluded, or handled in other ways.

1.2. Format

The annotation includes bounding boxes and IDs for each pedestrian, if there are multiple pedestrians in the scene. To simplify the concept for now, the gesture class only describes the gesture towards the ego driver. 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.

video_name	camera_name	pedestrian_id	start_frame	end_frame	gt_id	body_desc	interpret_desc
str	*str*	*int*	*int*	*int*	*int*	*str*	*str*
video_04	front	24	41	56	2	*"..flat hand.."*	*"..stop.."*
video_04	front	24	57	63	12	*"..nods head.."*	*"..approve.."*
video_04	back	71	45	56	10	*"..points.."*	*"..go there.."*
video_04	back	52	48	46	13	*"..spins.."*	*"..unknown.."*

Table 2. Annotation format example including multiple pedestrians. The annotations contain the features: Name of video (`video_name`) , Name of camera (`camera_name`) , Pedestrian ID (`pedestrian_id`) , Start frame at movement (`start_frame`) , End frame at movement (`end_frame`) , Gesture class ID (`gt_id`) , Body movement description (`body_desc`) , Interpretation description (`interpret_desc`).