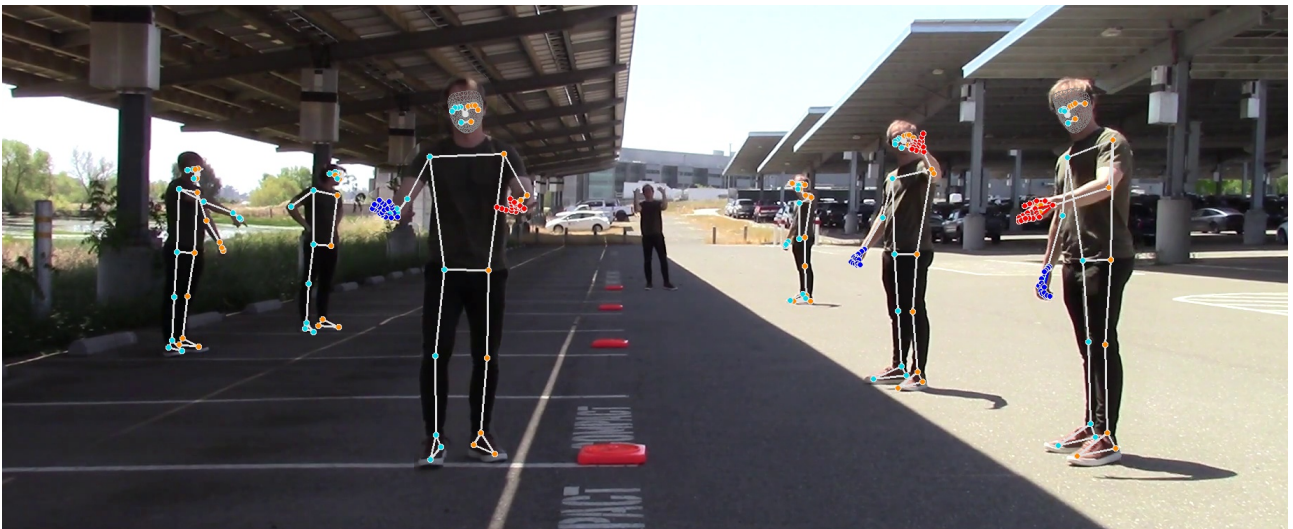

Enhancing Visual-Language Models in Zero-Shot Pedestrian Navigation Gesture Recognition for Conflicting Gesture-Authority Decision-Making in Autonomous Driving



**AALBORG
UNIVERSITY**



**UNIVERSITY OF CALIFORNIA
MERCED**

Master Thesis, 10th semester

Tonko Emil Westerhof Bossen

Student nr.: 20203031

Student mail: tbosse20@student.aau.dk

Computer Engineering - Artificial Intelligence, Vision & Sound

Aalborg University, 2025

Title:

Enhancing Visual-Language Models in Zero-Shot Pedestrian-to-Driver Navigation Gesture Recognition in Conflicting Gesture-Authority Scenarios for Autonomous Driving Decision-Making

Project Period:

3rd February - 4th June 2025

Project Group:

CE10-AVS, gr. 1044

Participants:

Tonko Emil Westerhof Bossen

Supervisors:

Andreas Møgelmoose

Aalborg University

Ross Greer

University of California, Merced

Copies: 1

Page Numbers: 73

Date of Completion:

June 4, 2025

Abstract:

This study aims to address the task of recognizing pedestrian-to-driver navigation gestures in a zero-shot setting, enabling safe decision-making even in conflicting scenarios. Navigation gestures are a daily routine in driving to make it safe for all. Gesture in conflict is more of an edge case, but these situations can also be critical, making gesture recognition and decision-making essential. Recognizing pedestrians' gestures is a significant aspect of the study. This led to the development of enhancement methods *Supplementary Body Description with VLM* and *Pose Projection* and evaluation methods *Classification*, *Natural-language*, and *Reconstruction* of VLMs in this domain. Alongside, three datasets were created with annotations: Acted Traffic Gesture (ATG), Instructive Traffic Gesture In-The-Wild (ITGI), and Acted Conflicting Authorities Navigation Gestures (Act-CANG). Across three VLMs, initial results were poor across all three evaluation domains. VideoLLaMA3, with and without enhancements, achieved F1-scores between 0.02 and 0.06 in classification. These results highlight the current limitations of VLMs in accurately recognizing pedestrian navigation gestures. This underscores the need for further research, either through fine-tuning or alternative approaches.

Preface

Aalborg University, June 4, 2025

This project was written at the Mi³ lab (Machine Intelligence Interaction Imagination) at the University of California, Merced, in the United States of America. Prof. Ross Greer from UC Merced has been the primary supervisor, and Prof. Andreas Møgelmoose from Aalborg University has been the official supervisor. A part of this project led to the publication of a paper at CVPRW (*Conference on Computer Vision and Pattern Recognition Workshop*) DriveX 2025¹. The paper is attached as Appendix A, [VLM Evaluation Paper \[4\]](#), referenced to and elaborated in the report. It is a part of the thesis, but will be kept separate, to stay true to the publication. An ‘[Co-author Statement and Signature for \[4\]](#)’ is attached as Appendix B describing and approving the contributions of all authors.

AI Utilization Along with this project, I have utilized a few tools to assist me. This includes ChatGPT Plus [33], Grammarly Pro [18], and GitHub Copilot Prop [16]. ChatGPT has assisted me in discussing aspects of the project, including terminology and vocabulary, optimizing the formulation of complex sentences, debugging software code, data formatting, and studying to expand my knowledge. This increased the productivity efficiency, helped the project to a higher level of complexity, and formulated it better and more clearly for the reader. Grammarly also assisted with increasing formulation and clarity. GitHub Copilot has helped increase productivity in software development by refining auto-completion and establishing a baseline of code to build upon continuously.



A handwritten signature in black ink, appearing to read 'Tonko'.

Tonko Emil Westerhof Bossen

Student nr.: 20203031

Student mail: tbosse20@student.aau.dk

¹[DriveX-Workshop, Github](#)

Acknowledgments

As an extra note, I would like to thank everyone who was a part of my time writing this project. This wasn't only a thesis but also a journey.

Thank you to,

The University of California, Merced, for allowing me to enroll as a visiting graduate student. The students and staff for making me feel at home. The Mi³ lab and the undergrads for taking me in and making me feel like I was a part of the lab. Sven Kirchner from the Technical University of Munich, for joining me in the lab, even just for a time. Prof. Andreas Møgelmoose from Aalborg University for giving me the opportunity and helping me study abroad. Last but not least, I sincerely thank Prof. Ross Greer for taking me in as his student and making me feel welcomed, challenged, and genuinely supported from day one.

Contents

1	Introduction	1
1.2	Project Summary	3
2	Initial Problem Formulation & Analysis	5
3	Problem Context and Study Scope	9
4	Conflict Data Analysis & Definition	13
5	Background Research	17
5.1	Autonomous Driving	17
5.2	Pedestrian Detection	17
5.3	Ego Driver Action Classification	20
5.4	Video-Language Model (VLM)	20
5.5	Conflict Scenario Data	22
5.6	Signs detection dataset	22
5.7	Large Language Model (LLM)	22
5.8	Trajectory Planning	23
5.9	Findings Summary	23
6	Conflict Decision-Making: Design & Experimentation	24
7	Datasets: Preparation & Creation	29
7.1	Acted Conflicting Authorities & Navigation Gestures (Act-CANG)	29
7.2	Acted Traffic Gestures (ATG)	38
7.3	Instructive Traffic Gestures In-The-Wild (ITGI)	40
8	VLM Enhancing for Navigation Gestures	42
9	Results	47
10	Discussion	49
11	Conclusion	52

12 Future Work	53
A VLM Evaluation Paper [4]	59
B Co-author Statement and Signature for [4]	69
C Review Analysis of the “VLM Evaluation” Paper	70
D Post ‘VLM Evaluation Paper’ Planning	73

1 Introduction

Traffic consists of more than just driving. It includes human interaction through gestures between drivers, pedestrians, and others. This reduces misunderstandings of other people’s anticipated actions, making traffic safer and more efficient for everyone involved [38]. With the rise of autonomous vehicles (AVs), implementing this is becoming increasingly critical for safety and efficient flow. As they can predict others’ behavior [31], it is thought they cannot yet perceive their intended action from gestures, making the AVs uncertain, resulting in a complete stop and waiting for the pedestrian. This solution is unreliable, as it can lead to a situation where everyone is waiting for each other to act. (Pedestrians are referred to as walking people in traffic, but I mainly refer to all humans in a traffic environment, including pedestrians, bikers, vehicle drivers, etc.)

In this project, I research human navigation gestures, focusing on traffic environments to advance AVs’ perception. The gestures in this project are viewed from a second-person perspective, where the driver or vehicle interacts with pedestrians or other subjects outside of the vehicle. First-person gestures inside the vehicles are excluded. I then explore navigation gestures further to examine their capabilities to convey information (e.g., Fig. 1.2) and facilitate decision-making in conflicting scenarios (e.g., Fig. 7.2).



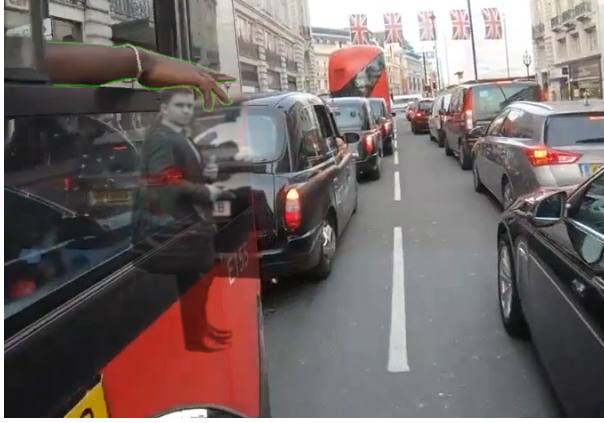
Figure 1.1: Example from the Acted Conflicting Authorities & Navigation Gestures (Act-CANG) dataset. Elaborated in Chap. 7.1. Crop of video_31, showing a supposed police officer (left) and a civilian (right) gesturing the ego driver towards opposite directions. *Can a model learn to make the right call? In terms of law, safety, or culture?* These are the questions I discuss and research in terms of conflict.

1.1 Motivation

We seek to recognize the gesture in a zero-shot setting for two main reasons. The problems are the classic ‘lack of data’, ‘outlier’, and ‘cultural meaning’.

The primary reason is the lack of current data available on navigation gestures from a second perspective. This makes it challenging to train a model well enough. A significant amount of data is required, as most gestures are temporal and can vary in subtlety. Since these gestures are simple for most humans to act out, a pretty large dataset could be made with enough time.

However, the main reason for zero-shot recognition is due to outliers and cultural meaning. This is illustrated in Fig. 1.3. The main point is that an LLM expands the knowledge about the gestures. The sense of gestures such as ‘Stop’, ‘Turn left’, etc., is all ‘common knowledge’. However, the data revealed that some gestures can be challenging for human drivers. These outliers make it challenging to train a model, and can easily be misinterpreted or be important. Some gestures can also turn out to



(a) First frame



(b) Second frame

Figure 1.2: First frame (*a, left*): Bike driver (*POV*) is unaware of occluded pedestrian (*light gray ghost*), but is alerted by the bus driver’s hand (*green outline*). Second frame (*b, right*): The bike driver reacts to the alert and stops. He observes the pedestrian. The pedestrian gestures to the biker to drive with his hand, but the biker nods to signal the pedestrian to proceed.¹ This short clip illustrates the advantages of AVs perceiving gestures, with the two types of liners and outliers.

be quite complex and difficult to categorize. This also concerns the varying meanings of gestures, which do not always have the same meaning in different countries, regions, or areas. This also concerns authority, as specific areas have a higher authority simply due to logos according to the property. A classification model is not capable of understanding these ‘third-party’ features. An LLM with expanded knowledge has a higher abstraction, which enables this knowledge.

An example of this ‘human intuition’ is seen in Fig. 1.2, where the bus driver alerts the biker. The bus driver’s gesture does not resemble a *classic* ‘Stop’ gesture. This is most likely an outlier even in a huge navigation gesture dataset, making it difficult to use for training. By using zero-shot recognition with a high-level abstraction model, it should theoretically be able to interpret the gesture to stop. Perhaps, with or without considering the environment. This is the core goal of this project, which I aim to address and resolve. Hopefully, this can create safer and more natural traffic environments for everyone to enjoy.

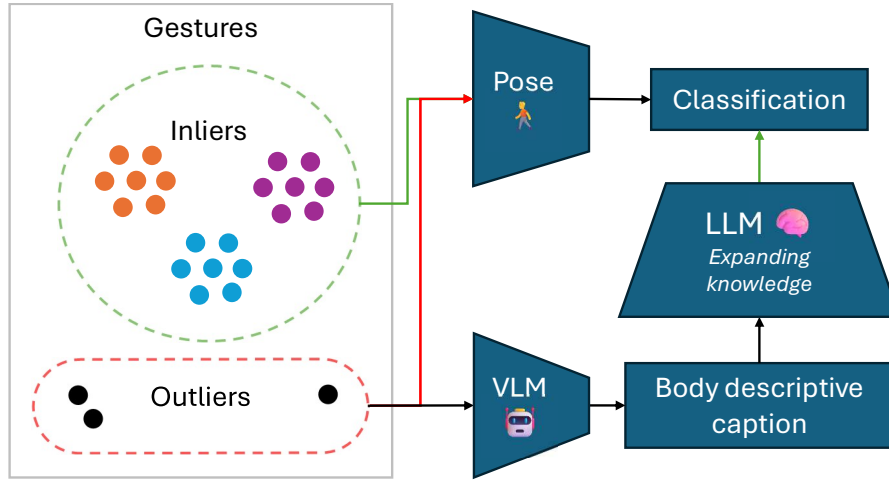


Figure 1.3: Illustrating the argument for using a zero-shot approach instead of a supervised pose classification model. While standard classifiers perform well on in-distribution gestures, they fail to generalize to novel poses. The idea is that the VLM can describe the human body in detail enough for an LLM to classify it due to its expanded knowledge.

1.2 Project Summary

The first chapters include initial research to grasp the complexity of the problem as a whole. This included a range of research areas, data structures, and model architectures. I examined the abstraction levels of multiple types of models for decision-making in conflicting scenarios, which were theoretically explored. As understanding evolved of the complexity that lies in making the optimal decision. First, it was thought to be a linear hierarchy of only a pedestrian authority level. For this, I examined a multi-layer perceptron (MLP). It was then believed that the gesture and the number of pedestrians could also influence the decision. To this end, I explored Transformer architectures with self-attention, enabling the model to learn the priority of each element based on its interactions with other components in the scene. Lastly, I utilize large-language models (LLMs) to increase the abstraction plane to an even higher dimension. This enables the model to perform zero-shot decision-making, as it should be able to conclude knowledge of social culture regarding authority and gestures, as well as traffic guidelines regarding safety and driving rules.

Initially, the primary focus of this project was the decision-making process for conflicting navigation gestures in a zero-shot setting within a traffic scenario. However, a lack of utilization of VLMs for this purpose was found. This led to a shift towards more trainable, fine-tuning, or transferable models. This only lasted for a period, as the attendance at the WACV conference (Winter Conference on Applications of Computer Vision) demonstrated the potential of VLMs, leading to a re-exploration of their capabilities. Especially the paper about CoVLA [2] using VideoLLaMA2 [25, 10] raised the hope of success in zero-shot learning using VLMs. They are using VLMs to decide the trajectory of the ego driver. Explained further in the ‘VLM Evaluation’ paper [4]. A full pipeline was developed inspired by the CoVLA pipeline, explained in Section 6. However, the VLMs did not achieve

notable accuracy in recognizing gestures. Instead of returning to the trainable models, this study of incapability was explored further. This led to focusing on proving this hypothesis, developing enhancement methods, and evaluating them. Time limitations led to the conflicting part of the project staying in the initial and dataset development phase. It still allows for further research with plenty of data and thoughts to extend upon. The discussed possible research areas in this domain are listed in Appendix D. This shifting in focus is the reason for the broad aspect of both zero-shot and few-shot technology discussed throughout the project. The report is only written in chronological order to some extent. This may cause some confusion regarding naming and technology understanding, but this order makes the most sense.

In total, three different datasets were created for this project, in addition to the research on perceiving gestures and decision-making in conflicting scenarios. The datasets all fall under the category of second-person navigation gestures, with two of them also including conflicting scenarios. Each dataset contains the thought behind its creation and annotation.

1. Acted Traffic Gesture (ATG) dataset
VLM evaluation paper [4] & Sec. 7.2, p. 38
2. Instructive Traffic Gesture In-The-Wild (ITGI) dataset
VLM evaluation paper [4] & annotation in Sec. 7.1, p. 29
3. Acted Conflicting Authorities & Navigation Gestures (Act-CANG)
Sec. 7.1, p. 29

The evaluation method was developed before the enhancement. This was due to the developing stop block of the models not working as intended. This led to writing a paper on the matter. Due to the limitations on paper size and time period, the paper only includes the evaluation. It is summarized and extended in Cap. 8 and cited as [4].

The software for the project is divided into multiple repositories on GitHub, since only some parts have been published.

- Conflict Experimentation [tbosse20/mercedthesis](https://github.com/tbosse20/mercedthesis), [Github.com](https://github.com/tbosse20/mercedthesis)
- Conflict Annotation & Guidelines [tbosse20/navigation-gesture](https://github.com/tbosse20/navigation-gesture), [Github.com](https://github.com/tbosse20/navigation-gesture)
- VLM Evaluation [tbosse20/gest_VLM_eval](https://github.com/tbosse20/gest_VLM_eval), [Github.com](https://github.com/tbosse20/gest_VLM_eval)
- VLM Enhancing [tbosse20/gest_VLM_eval/tree/enhance](https://github.com/tbosse20/gest_VLM_eval/tree/enhance), [Github.com](https://github.com/tbosse20/gest_VLM_eval/tree/enhance)

2 Initial Problem Formulation & Analysis

This section serves as a project proposal, providing an overview of the project's scope, its potential division over time, and a plan for its progression. I initially examined this problem comprehensively, considering the availability of content and research on the topic, as well as the opportunities for further study. At the beginning of the project, the lack of data was not known, raising the question of whether it would be feasible to continue with the project.

2.1 Initial Problem Statement

In autonomous vehicles (AVs), a car might find itself in a situation where different authorities give multiple instructions simultaneously, making it difficult for the model to decide which instruction to follow. This type of situation is an edge case, which also makes it more critical for the AV to make the correct decision quickly.

Current technology allows the AV to perceive instructions from all pedestrians in a scene, but it cannot decide which to follow. This project investigates the area of learning-to-rank in the context of autonomous driving. Related work presents authorized traffic controller (ATC) detection [32] but does not utilize intelligent prioritization of the different pedestrians in the given scene.

This project studies pedestrian-to-vehicle communication from the perspective of civilian drivers, including casual drivers, taxi drivers, and bus operators. In contrast, emergency service drivers, such as police officers, firefighters, and paramedics, operate under a distinct hierarchy of authority. They follow a separate system of regulations, which allows for more advanced driving behavior, hence another adherence to incoming instructions.

From this, I state the initial problem statement as follows:

How can autonomous vehicles learn to prioritize incoming instructions from multi-hierarchical pedestrian crowds to decide on the safest instruction to execute?

2.2 Initial Data & Problem Analysis

To determine the type of data suitable for this project, an initial data analysis is conducted, followed by a further study of the collected data to optimize both the data-gathering process and the data itself.

A first thought on how the model would prioritize different personnel and objects could be the following list, keeping in mind that it is surely incomplete and biased.

1. FBI and police officer
2. Firefighter and medic
3. Reflective safety vest and authorized traffic controller (ATC)
4. Injured and personal in proximity to said person

5. Traffic control devices (traffic signs, cones, etc.)
6. Car owner and direct approach (eye contact, pointing)
7. *any other personal: ignored*

This also includes different levels of data abstraction to make the data easier to comprehend and collect, but at the expense of realism.

2.2.1 Initial Data Analysis

The dataset primarily aims to train the prioritization of pedestrians and distinguish their guidance. A short brainstorming session to understand the aspects of the data needed for the project led to the following:

- **Occurrence**

How often does the actor occur in a scene, representing importance in rarity?

- **Situational**

The chosen actor depends on the scenes in which they appear. If police officers often find themselves in chaotic scenes. It is, however, challenging to know what kind of scene is present.

- **Gesture priority**

Certain gestures rank the personal, not the actor. Said ‘Stop’ from a civilian could overrule a ‘Go’ from a police officer.

- **Following instructions**

The driver follows the instructions of the highest-prioritized actor in the scene.

The most promising type of data to utilize appears to be the “Following instructions,” which has a clear input (instructions from multiple pedestrians) and output (response action to the highest-prioritized instruction) for each scene. The other types of priorities are after rarity in the data.

Directional instructions could be given through signs, verbal messages, or gestures. Incoming blockages, like humans and roadblocks, are also considered when prioritizing instructions. For example, is it possible for a police officer to instruct an AV to drive into a tree? Verbal or visual emotions may increase the priority of an actor, and could this be the primary use when dealing with unknown classes? When training on data, what does it mean if the driver does not follow the given instructions? This may down-prioritize the authority but go against the law. This could occur in a scene showing the “gesture priority” rule.

2.2.2 Three-stage realism

Three different degrees of realism make it easier to synthesize the data; however, it is unfeasible to conduct such research. This provides an idea of what the dataset requires as a minimum and an optimal.

1. Naturalistic

Video with directions from multiple accurate pedestrians and patterned action responses in a natural environment. *Ex. A police officer in uniform points left, and an ATC points right on the road. The car turns left when both are visible and right when only the ATC is visible.*

2. Simplified

Video with directions from multiple easily recognizable pedestrians and actions in any environment. *Ex. One person wearing red points left, and another wearing a green hat points right on any street. The camera moves left when both are visible and right when only the person wearing the green hat is visible.* This allows the model to match the priority recognition with the gesture.

3. Minimalistic

Image with directions from multiple canonical gestures and actions. *Ex. "A" and a red circle appear together, and "B" if only a green circle appears.* This allows for matching the property, but not matching any gesture.

It is trained and evaluated on zero-shot paired subjects to stop the model from recognizing paired subjects instead of constructing an abstract priority listing. This means that the evaluation scenes include unseen pairs of subjects, such as *A* and *C*, whereas the training scenes include *A* and *B* and *B* and *C*.

2.2.3 Data Gathering Proposal

An initial major issue of this project could be collecting the dataset. The project's approach would be to research methods for gathering a pilot dataset to conduct initial research and gain an understanding of the problem and the domain. Further insights and clarification of the data and the project period will help inform estimates of additional data gathering and prioritization.

The following lists methods to acquire the dataset, but they increase in difficulty, leading to a smaller dataset. Given the project's scope, obtaining the dataset beyond the 6th method is highly unlikely.

1. Obtain the dataset with the correct elements
2. Scrape other datasets with data overlap
3. Virtual recording from pre-made media, such as video games
4. Virtual synthetic made in a video game environment
5. Realistic synthetic with minimal requirements
6. Realistic scenario with acting pedestrians
7. True scenario with real-life pedestrians

The dataset must contain multiple instructions, such as signs or gestures, and include the desired action. It is preferably in a video, but an image can be used if all the information is perceivable.

2.2.4 Time-period

To get started with the project and to better understand the dataset, minimalistic data will be collected from taken or synthesized pictures. This allows for a straightforward approach to testing the zero-shot paired subject aspect. Along with this, I will be looking into possibilities of scraping data, as this can give some good quality data, but it's uncertain how much.

Without any scraped data, more realistic data will be needed. With more insight, more realistic or virtual videos and images will be created using simplified data types to match the selected urban actor's gesture.

Creating more realistic or virtual naturalistic data may be possible depending on the period and resources. Hopefully, at this time, there will be a better understanding of what is essential for the data, and perhaps only some of the data will be needed. Otherwise, this is a point where more resources are required, highlighting the potential of this project.

2.2.5 Further Data Analysis & Research

Reaching beyond the problem itself could potentially include researching XAI in terms of how having the model explain its decision-making process. This would enhance the model's reliability.

3 Problem Context and Study Scope

Many questions regarding the system's behavior and decision-making in varying scenarios were raised during the initial research period. This widened the understanding of the problem's complexity and helped narrow the scope of this specific problem to a particular research area. It opened up opportunities for future research. Here, I describe the abstract problem, considerations, and scope I seek to study during the project. I wish to understand the issue as a whole, but instead of researching the entire project, I will focus on a specific part.

3.1 Comprehensive Problem Discussion

Analyzing this problem, the number of parameters to be considered in the driver's decision-making process continues to increase. This problem can quickly be oversimplified or overcomplicated. Looking at it in small pieces can help understand where to start, where to head, and when the problem is too complex for now. Before attempting to solve the entire issue, let's define the problem and identify the area to begin with.

To develop a system that can handle all scenarios, it must have a more complex understanding than only how the different authorities outrank each other. The system needs to understand the scene as a whole. Parameters that could influence the driver's decision include the number of pedestrians, their attention, emotions, authority, gestures, the recipient of the gesture (which may also affect the ego driver's decision), verbal communication, obstacles, other vehicles, idle pedestrians, and many more. This raises many questions about how such a system should react in various cases. Some aspects are listed separately to facilitate easy reference as a checklist and for future research purposes, in case they are excluded from the scope of this study.

A1. **Instruction Ambiguity**

'Stop' gesture can mean 'immediately', 'a specific location', or 'at any speed'. How is the driver supposed to act accordingly?

A2. **Ground Truth Discrepancy (Label Bias)**

Will the AV be able to decide on safer actions than the ground-truth driver's actions, and how can I make sure the evaluation is not penalized for that? Maybe it is due to more knowledge, quicker decision-making, and irrationality.

A3. **Gesture Target Ambiguity**

Gesturing towards other vehicles. This can be approached only by examining the gestures towards the ego driver, including additional abstract gestures, or by considering the receiver's gestures in the decision-making process.

A4. **Driver Action Variability**

How will the system learn the correct action in given scenes, if the same or different drivers react differently?

A5. Driver Authority

The priority depends on the driver's authority. An ambulance may have the highest authority, or a police officer's gesture can authorize a vehicle to crash.

A6. Indirect Occlusion Insight

Pedestrians can provide the driver with information about the environment. The driver could rely more on pedestrian gestures in very compact environments, such as a parking area. Compared to empty spaces, where the driver can view all possible objects.

A7. Pedestrian Reliability

Considering all other variables, the ambiguity of human trust and error, and the ability to question and reason about the given gesture's reliability.

A8. Ego Clarification Feedback

Feedback from the ego vehicles is suggested to clarify the understanding of the pedestrians' gestures.

A9. Human-Robot Interaction Bias

Pedestrians may interact differently with the vehicle, knowing it is a non-human driver and thinking it is incapable of reacting to gestures.

These questions surround the idea that the AV needs a complex understanding of the scene, not only the pedestrians' hierarchy as initially thought. The AV must understand direct instructions intended for the driver and interpret them precisely enough to fully execute the intent, solely based on hand gestures and body language. Even if auditory input were included, that could add another level of complexity to the system.

3.1.1 Possible Problem Scopes

Where should this project head now? In *Chapter 2 Preliminary Research*, obtaining proper data for this problem seems complicated. It is not an edge case, as pedestrian-driver communication is quite common. However, obtaining real-life scenarios is not as easy as regular driving. Real-life data is needed, as deciding on the action looks more complex than initially presumed. Therefore, it needs to be real scenarios to train accurately. I list three aspects of this problem that each can be a focus for this project, keeping in mind the lack of this type of dataset.

S1. Training Conflict Classification Pipeline

Develop a pipeline to train on real-life data once it has been collected and processed. It is reasonable to collect 'fake' data to develop this pipeline, but it is insufficient for training, as some situations may not accurately represent real-life scenarios.

S2. Zero-shot Conflict Classification

Develop a pipeline that evaluates zero-shot conflict classification with medium-realistic data mentioned in Chapter 2. This includes pedestrian authority and gesture recognition, but without training a model to learn the hierarchy of authorities and gestures.

S3. Navigation Gesture Execution

Research the execution aspect of the individual gesture. This involves training a model to execute the presented gesture. This doesn't need 'real' data and can be developed with 'fake' data.

3.2 Final Project Scope and Formulation

The focus of this project will be **S2. Zero-shot Conflict Classification** - *autonomous vehicle decision-making in conflicting scenarios given pedestrian gestures, authorities, and obstacles*. As stated earlier, this problem is more complex than just these parameters, so the remaining parameters will also be taken into consideration. Still, they will not be the primary focus of this project. Researching this aspect can help to understand what is needed to continue this work and advance the system to higher complexities.

Hence, I propose the theoretical equation that will be included in this study. I assume the current standard formula for optimal trajectories with the notation ξ . This consists of the parameters of the ego vehicle, including the speed and distance to the ego vehicle by other moving objects, such as pedestrians, bikers, and vehicles, as well as static objects like cones and buildings, and the general environment and road as an abstract concept.

In this project, I propose additional considerations for the formula to compute the safest and optimal trajectory, with further information. The symbols are explained in Table 3.1. I assume the gesture and authority are recognized correctly and only consider the gestures directed towards the ego driver:

$$a^* = \underset{a \in \mathbf{A}}{\operatorname{argmin}} \left(\xi + \sum_{i \in \mathbf{P}} \mathbf{G}_i \cdot \mathbf{A}_i \mid \mathbf{O} \right) \quad (3.1)$$

Symbol	Meaning	\mathbf{v}	Domain
a^*	Optimal action	\hat{y}	$a^* \in \mathcal{A}$
\mathcal{A}	Action labels	y	\mathbf{C}
\mathbf{P}	Pedestrians	x	\mathcal{X}
\mathbf{A}	Authorities	x	\mathbf{C}
\mathbf{G}	Gestures	x	\mathbf{C}
\mathbf{O}	Obstacles	x	\mathcal{X}

Table 3.1: Definitions of key symbols. (Where \mathbf{C} is class, \mathcal{X} is set.) This is explained more in depth in Sec. 6.

We formulate the scope of the included questions listed earlier.

A1. Instruction Ambiguity

The corresponding action can be executed in different degrees. I aim to classify the gesture and action as nominal data.

A2. Ground Truth Discrepancy

This issue will be researched as part of the evaluation method for this problem.

A3. Gesture Target Ambiguity

I only consider the gestures directed towards the ego driver.

A4.. Driver Action Variability

This will be taken into consideration when creating the dataset.

A5. Driver Authority

This project is based on civilian vehicles, as the correct action will change depending on the driver's authority.

A6. Indirect Occlusion Insight

This aspect is excluded from the project. It is briefly considered the possibility \mathcal{P} of obscured objects $\tilde{\mathbf{O}}$ of objects \mathbf{O} , taking into account the given pedestrians' \mathbf{P} gestures \mathbf{G} and authority status \mathbf{A} , as well as the occluded areas in the environment \mathbf{E}_0 . $\tilde{\mathbf{O}} = \mathcal{P}(\mathbf{O} | \mathbf{P} \cdot \mathbf{A} \cdot \mathbf{E}_0)$.

A7. Pedestrian Reliability

I consider all gestures accurate and to be with genuine intentions.

A8. Ego Clarification Feedback

It is excluded, since it focuses more on responding and not perceiving,

A9. Human-Robot Interaction Bias

It is excluded, since it focuses more on the cognitive aspects of the user.

3.2.1 Problem Statement and Objectives

This concludes the final problem statement for the project, along with three sub-problem statements.

“How can autonomous vehicles effectively perform safe and reasonable actions in scenarios with conflicting pedestrian-to-driver gestures in a hierarchical crowd?”

RQ1. Recognition

How can we recognize pedestrians' navigation gestures and authority in zero-shot settings?

RQ2. Decision-making

How can we infer a model to decide the safest action given the current environment?

RQ3. Evaluation

How can we evaluate whether the final system's action is correct?

4 Conflict Data Analysis & Definition

This section delves further into the details of the data for this project. It is essential to understand what already exists and what is needed to find or generate it. I strive to understand the various possible scenarios, ranging from common to edge cases. This follows an extended analysis and formulation of the data. I also look into the categories of different scenarios and propose a model pipeline and data format.

The type of conflict data is analyzed and defined to ensure that the gathering encompasses the various kinds of scenarios in this domain. I aim to explore possible scenarios in theory, to gain a deeper understanding of the relationship between the parameters, what is physically feasible, and how a human driver would react, or whether the driver is presented with a dilemma.

An optimal data analysis would be processed based on real-life data in the correct context. However, it must be done in theory since finding enough data for all possible scenarios is difficult. Gathering data by driving around the city could be feasible, but it would be time-consuming, as some scenarios are edge cases.

To grasp all scenarios by widening the view, I list all thoughts on possible properties of the different parameters' authority and gesture. 'Pedestrian state' is added as implicit communication input.

- **Authority Categories**

- **Law Enforcement** - FBI, military, police.
- **Emergency** - Firefighter, medic, paramedic.
- **Traffic Management** - Traffic warden, reflective safety vest (may be unauthorized), authorized traffic controller (ATC), contextual parking attendant (theme parks, store).
- **Signage** - 'Stop'-sign, yield lines, traffic light, barricades.
- **Civilian** - Stranger, owner, passenger.
- **Miscellaneous** - GPS.

- **Gestures**

Go/proceed/forward, stop, turn right, turn left, turn left waiting, lane change, straight ahead, wait, pull over, slow down, reverse/"go back", point (orientation, directional, location), idle, social (thanks, sorry), '*with a stick'.

- **Additional Input**

- **Solid blockage** - Tree, big animal.
- **Soft blockage** - Cone, small animal.
- **Pedestrian State** - Injured, angry, scared, eye contact, direct.

We perceive the authorities as ordinal data, as they have a social concept of hierarchy, but it isn't

easy to measure. Later research can perceive the data more as a ratio with a complex understanding of the hierarchy.

Scenario	Description	Scene	Action	Obey
Unique	Pedestrians' authority and gesture differ, prioritizing one pedestrian.	$A_i G_i$	α	i
Shared	Distinct authorities share gestures, prioritizing one or multiple authorities. <i>Misleading or 'crowd effect' increases reliability.</i>	$A_{i,j} G_\alpha$	α	$i \vee j$
Conflict	Similar authorities' gestures differ, obeying one pedestrian over the others.	$A_i G_{\alpha,\beta}$	$\alpha \vee \beta$	i
Ignore	Acknowledges gestures are disobeyed, obeying an unknown source.	$A_i G_i$	$\notin G$	<i>unknown</i>
Disagree	Pedestrians' authority aligns, but their gestures contradict, simultaneously obeying and disobeying the same authority type.	$A_i G_\alpha \wedge A_j G_{\bar{\alpha}}$	$\alpha \vee \bar{\alpha}$	$i \oplus j$

Table 4.1: Abstract scenarios involving distinct authority types A and gestures G , highlighting how authority obedience is determined. i and j denote the corresponding gesture and action to determine the correctly obeyed authority of α and β . Symbols are explained in Tab. 3.1 in Section 3.2. Concrete examples are visualized in Fig. 4.1.

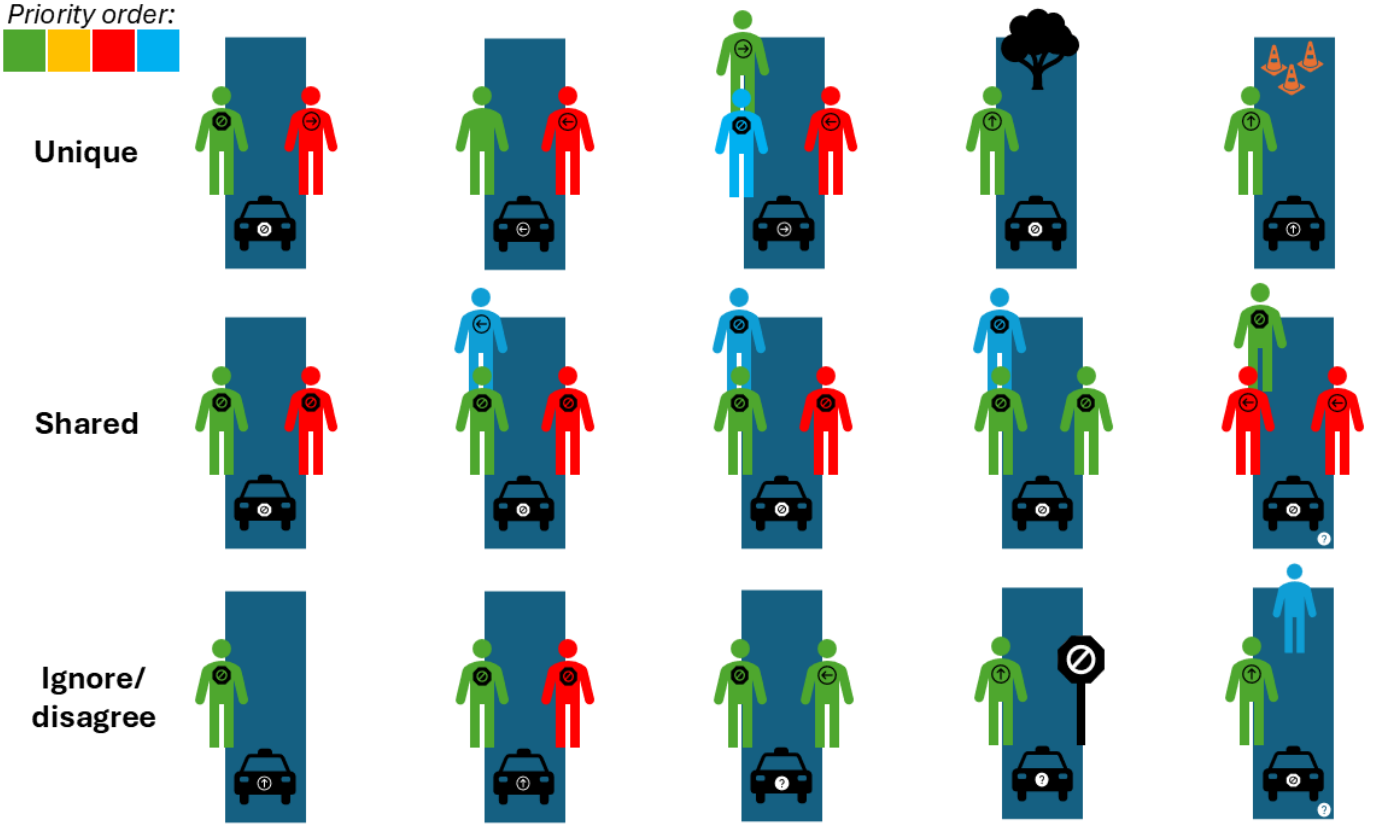


Figure 4.1: A visualization of the possible conflicting scenarios considered. The priority is fixed as shown in the legend *Priority order*, with the gesture indicated as a symbol on each pedestrian. The categories are explained in Tab. 4.1.

Possible ‘flaws’ or ‘errors’ to look out for in the dataset to ensure it is clean and accurate when evaluating are listed as follows. They help to maintain a clean dataset considering the pedestrians’ gestures, authority, and the driver.

- A single pedestrian provides multiple gestures.
- The pedestrian provides an incorrect gesture, while the driver executes the correct action.
- Delayed gestures.
- Misinterpret gestures from another driver.
- Approaching/asking pedestrians for directions. The pedestrian could ‘Point’ while the driver listens instead of executing the gesture shown.

I distinguish the various settings in which an agent’s intentional gestures may be observed. *Informal* traffic settings involve intent, querying, and instructional gestures. These settings require comprehensive scene understanding to ensure a safe response, as other subjects are more unpredictable. In contrast, *formal* settings are typically more predictable as they do not involve intent-based gestures, and follow a more concrete set of rules and traffic flow. They have a specific focus point, such as law enforcement or parking enforcement, which reduces the complexity of the scene.

4.1 Initial Pipeline Proposal

I propose a pipeline for training and evaluating the decision-making model using videos as input.

I list the pipeline's required elements to be located or developed. This list will serve as a checklist and will be updated as further research and development progress.

1. Pedestrian Detection
2. Authority Classification
3. Gesture Classification
4. Action Classification
5. Prioritization Model

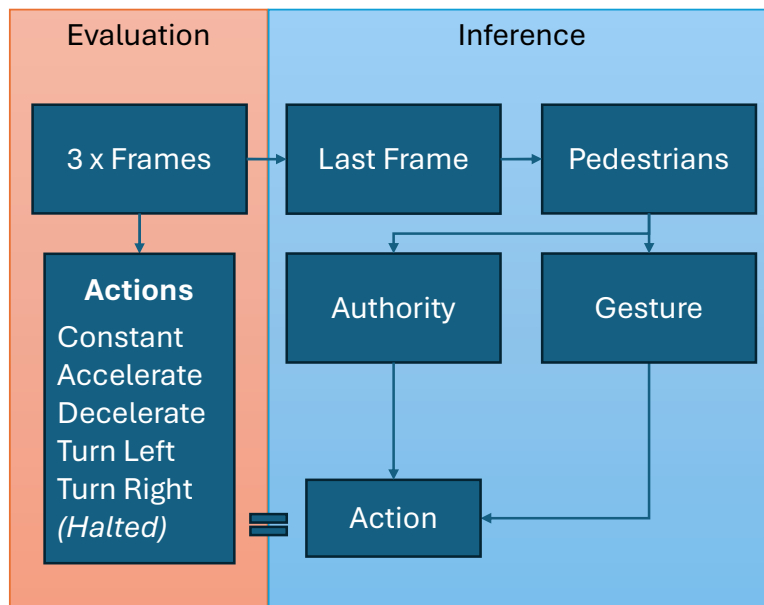


Figure 4.2: Initial pipeline infers each frame with the authority and gesture of each pedestrian. The predicted safest action is classified and compared with the label action, which is classified from the previous three frames.

4.1.1 Annotation Format

Annotations are made using the CVAT online software. The annotations must be for each frame, as they can change over time. This annotation includes the pedestrian's video frame and bounding box, along with their authority and gesture. For future work, *ego* is added too, which indicates if the gesture is pointed towards the ego driver.

video_name	frame	bbox	authority	gesture	ego
video1.mp4	1234	(50, 60, 200, 300)	civilian	stop	1
video2.mp4	5678	(100, 120, 250, 320)	police	go	1
video3.mp4	9101	(80, 100, 220, 280)	warden	left	0

Table 4.2: Example of annotation file with *authority*, *gesture*, and *ego*

5 Background Research

This chapter examines related work and state-of-the-art technology, seeking available models and datasets. This will offer a quicker and easier implementation, as well as a faster launch in the research, rather than starting from scratch. The ideal models would be out-of-the-box models, but transferable models combined with datasets would also be efficient. Models for both zero-shot and few-shot settings are explored to facilitate research in varying accuracy and data requirements. The required components and research are listed in **Sec. 4.1 Initial Pipeline Proposal**. This list serves as a checklist and will be updated with the most available and suitable method. Additional models that were found out later in the project are also listed. Along with each component, the technical aspect of its functionality is explained. The technologies, which are newer to me, are explained in more depth, as the others have been described before.

The ideal scene features pedestrians who communicate with the driver through gestures, and preferably, multiple subjects are present in a single scene. Datasets that include some elements of the ideal scene can help with refining the datatype or sub-elements of the pipeline. General websites which was looked into were [PapersWithCode](#), [Kaggle](#), [GitHub](#), [HuggingFace](#), [universe.roboflow](#), and [DatasetNinja](#).

5.1 Autonomous Driving

Autonomous Driving, also known as *self-driving* or *driverless*, is the ability of a vehicle to drive itself without human controls [34]. It is also referred to as an autonomous vehicle, or AV. The term is used broadly to encompass various levels of autonomy, including those that are not fully autonomous driving. 6 different levels of AV from L0 to L5. L0 is a fully human-controlled vehicle with zero autonomy. L1 introduces driver assistance features such as speed and road alerts, but the controls are still in human hands. L2 has partial autonomy, such as acceleration and steering, but humans should always monitor the environment to ensure safety. L3 is a huge improvement in AVs. It allows humans to be unaware of the environment. However, they should always be ready to take over the controls. In L4, the vehicle controls all situations, and human driving is only an option. L5 is fully autonomous, with human driving not even being an option. [44]

5.2 Pedestrian Detection

The most essential element of this project is the detection of pedestrians. For this, the state-of-the-art and out-of-the-box model ‘You Only Look Once’ (YOLO) [22, 39] is available. YOLO is a general object detection model, with ‘Human’ as one of the classes. It is easy to implement in Python 3 using the ‘Ultralytics’ library. YOLO is a single-stage detector. It sees the task as a regression problem. It divides the image into an $S \times S$ grid, which is used to predict the class and bounding box of each

cell. This enables detection of multiple objects in a single image, with less computation. [39] YOLO versions have been trained on both PASCAL VOC (PASCAL Visual Object Classes Challenge) [15] and COCO [27]. COCO has 80 classes, including vehicles and traffic sign detection [27], which makes YOLO suitable for detecting these objects as well.

Additionally, datasets were obtained with pedestrians, if further training is needed. All from *Dataset Ninja*¹: [kitti-object-detection](#), [mots-challenge](#), [bdd100k](#), [argoverse-hd](#), and [citypersons](#),

5.2.1 Authority Classification

Only the pedestrians can be detected, but not the authority. A few datasets were found with police officers, yellow safety vests, and civilians, which can be used for transfer learning.

- [universe.roboflow.com work-safe-project/safety-vest—v4](#)
- [github.com Ansarimajid Construction-PPE-Detection](#)

This could be improved too, by utilizing the paper ‘*Face, Body, Voice: Video-Clustering with Multiple Modalities*’ [5] for identity clustering, and equipment attachment of the user from ‘*Pose guided anchoring for detecting proper use of personal protective equipment*’ [49].

5.2.2 Pedestrian Attention

The priority of each ego driver’s pedestrian also lies in the attention of each pedestrian in the scene. I also include ‘*Eye direction detection*’ [41]², to distinguish this aspect.

5.2.3 Navigation Gesture Recognition

5.2.3.1 Consumer/Amateur Experiments

Chris, with the YouTube channel Dirty Tesla, uploaded a video where his autonomous driving Tesla stopped as a pedestrian walked up to the road. When the pedestrian pointed to a parked car, to show she intended to enter the vehicle, and not cross, the AV started driving again: [Tesla Self Driving Responds to Hand Signal](#). He introduced this evidence on this X profile. He asked, Does FSD understand the gesture of the pedestrian, or is it just registering them not crossing, and interpreting that as to proceed³. Tesla reposts saying “*Slowing down for pedestrian & taking their gestures into account before proceeding*”⁴.

Experiments to evaluate FSD’s capability of recognizing pedestrian gestures were made from Dirty Tesla [Does Tesla FSD Recognize Hand Signals? We Tested It!](#) and another YouTube channel TechkGeek Tesla, [Does FSD 12.4.1 Recognize Hand Gestures? — Real-World Test](#). They had different cases where the pedestrian would walk up to the road or cross the road with and without gesturing

¹[datasetninja.com](#)

²[fkryan/gazelle](#), GitHub

³[@DirtyTesLa](#), X

⁴[@Tesla](#), X

‘Stop’. Or stand on the road and gesture for the vehicle to proceed, but did not respond. They found out that the vehicle sometimes reacted to the pedestrian, but there was no difference in the AV’s reaction whether the pedestrian gestured or not.

These amateur experiments provide a fundamental understanding of the capabilities of AVs currently on the market. This reveals a gap in this domain and suggests a positive outlook for continued studies. They can help shape this project’s data collection and evaluation process, and also include more divergent pedestrian gesture data.

5.2.3.2 Real-life Samples

Additional recordings of real-life cases were found, which can also be used as a foundation for the dataset. Either by combining it or using it as an example, referring to the creation of a new dataset. All videos are on YouTube.

- [Dash Cam - Angry Lancaster City Crossing Guard, Gregory Hripto](#)
- [Toms River NJ: 2 Accident Scenes on Hooper Ave, Delivery Dashcam](#)
- [Stuff Like This Happens Everyday and I Don’t Care to Upload Anymore:, Delivery Dashcam](#)
- [Work Day of a Traffic Cop \(Toms River NJ\), Delivery Dashcam](#)
- [Tesla FSD 12 Recognizing Hand Signals?!, Tesla FSD \(Full Self Driving\)](#)
- [Tesla FSD Conquers Rainy Downtown Drive! V12.3.4 \[9:27\], Dirty Tesla](#)

5.2.3.3 Datasets

1. Qualcomm Jester Dataset

[qualcomm/jester-dataset](#) 22.8 GB HCI gestures

2. HAGrid dataset

[hagrid-cls-150k](#), [Kaggle](#)

Has varying dataset sizes on Kaggle from approximately 1 to 15 GB. They only have the ‘Stop’ gesture, and some also include ‘no_gesture’ HCI gestures

3. Uni ULM Traffic Gesture Dataset

[uni-ulm.de/en/in/mwt/traffic-gesture-dataset](#)

Pedestrian gestures, but the data type is not understood

Traffic Control Gesture Recognition for Autonomous Vehicles [47]⁵

Only HCI gestures like ‘Zoom’, ‘Slide’, ‘Stop’, and ‘Idle’ with hands were found. No hands or full-body gestures such as ‘Proceed’ and ‘Reverse’ were found.

⁵“Dear Tonko, the video data is not part of the dataset. It has just been used for demonstration for the IROS publication. Best regards, Julian”

5.2.3.4 Pose Estimation

1. OpenPose [8]
[CMU-Perceptual-Computing-Lab/openpose](#), [GitHub](#)
2. DEKR - HRNet [54, 42, 45]
[HRNet/DEKR](#), [GitHub](#)
3. Yolov8n-pose - Ultralytics [13, 22]
[docs.ultralytics.com/tasks/pose](#)
4. RTM Pose / mmdet - OpenMMLab [20, 11, 21]
[open-mmlab/mmdet/tree/main/projects/rtpose](#), [GitHub](#)
5. Mediapipe Holistic [17, 29]
[research.google/blog](#)

The selected pose model varied between Mediapipe Holistic and YOLO, as they are the most out-of-the-box, lightweight, yet still accurate, options, which increased development iterations. Mediapipe Holistic is excellent for hands and faces.

5.3 Ego Driver Action Classification

To avoid annotating each driving sample with the driver's action, it is hoped to utilize an 'Action Classification' model on video data.

1. Optical Flow, self-implemented
2. [github.com/commaai/openpilot](#) [7]
3. [github.com/CIFASIS/ORB_SLAM3](#) ([github.com/UZ-SLAMLab/ORB_SLAM3](#))
4. [github.com/MaybeShewill-CV/lanenet-lane-detection](#)
5. [github.com/ooooverflow/BiSeNet](#) [48]

5.4 Video-Language Model (VLM)

"Video-to-text" or video caption models were researched to gain an understanding of the scene. Using pre-trained caption models enables a general understanding of an image in zero-shot data, which is highly resourceful when a low amount of data is available. As there was limited data available for this project, various caption models were researched for their potential application in training the system

and for the final implementation. VLMs⁶ researched are LLaVA [25, 52], BLIP [24], ViLA [28, 9, 26], VideoLLaMA2 [25, 10], VideoLLaMA3 [25, 51], and Qwen [46].

Here, I explain the theory and details of a VLM, how it works, and why it excels in zero-shot settings. The purpose of a VLM is to generate natural-language captions for the given images or videos. The input may also include prompts to guide the captions in a more specific direction. [3] It is a *multi-modal* foundation model, meaning the model's input consists of varying media.

VLMs have been the primary focus throughout this project due to their potential abilities in a zero-shot setting, compared to traditional classifiers. Traditional classifiers are limited only to the data they have been trained upon, making them suitable for multi-shot or few-shot classification when fine-tuned. VLMs are well-suited for zero-shot classification due to their immense training dataset size. In theory, the model has been trained on millions of data samples, making it a subset of a foundation model. A foundation model is an end-to-end, general-purpose model that encompasses all aspects of a system. They are trained to generalize the perception of data from a single training stage. This enables the classification of unseen data excluded from the training data. An example could be 'A *penguin in a shoe store*'. Most likely, no images of a penguin in a shoe store exist.⁷ However, the foundation model learns the visual meaning and context of 'a penguin', 'in', and 'a shoe store' from previous images (e.g., '*penguin on ice*', '*man in a shoe store*'). The model embeds the image-text pair in the same or similar embedding space. So with an unseen image, the embedding resembles the 'correct' text most accurately. This is the term *Contrastive LanguageImage Pre-training* (CLIP) [36]. CLIP utilizes contrastive learning, a self-supervised learning method, to contrast the embedding features of images. The contrastive loss *InfoNCE* uses an anchor, positive, and negative sample to learn the construction. A current selected data sample is used as the *anchor* or a reference point. Similar-looking samples or augmentations of the anchor are used as positive samples with a similar 'label' as the anchor. Negative samples are other samples that differ from the anchor's features and label. The samples do not actually have labels, but this helps to understand the loss. Before the text-image pair can be constructed, both the text and the image need to be embedded using a text encoder and an image encoder. In the CLIP paper, as the text-encoder, they use a GPT (*Generative Pretrained Transformer*) [37]-like transformer architecture. For the image-encoder they both tried a ResNet-50[19] and a Vision Transformer (ViT) [14]. [36] ViT works by splitting up the image into patches, which, along with a position encoder, are parsed through a multi-head self-attention transformer. [14]

5.4.1 BLIP

BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models [24] This model was initially experimented with, leading to the decision to place VLMs on a shelf in this domain. It was not very precise. Multiple trials were tried with varying prompts, but they did not show any consistent success.

⁶huggingface.co/docs/transformers/en/tasks/video_text_to_text

⁷There are.. Apparently, *Beach Donkey got a pair*, [boston.com](https://www.boston.com)

5.4.2 VideoLLaMA & Qwen

The main VLM I use in this project is the *VideoLLaMA2* [10], *VideoLLaMA3* [51], and Qwen [46]. They are elaborated further in the *VLM Evaluation* paper [4]. The VideoLLaMA models were selected due to their proven potential in the CaVLA model [2]. Qwen was selected after the VideoLLaMA models failed to show any success, as suggested by the lab.

5.5 Conflict Scenario Data

To obtain the data type explained in Cap. 4 [Conflict Data Analysis & Definition](#), multiple ways are explored, since it looks like there is a lack of this. One idea could be *mining* enormous datasets to detect pedestrians and, either manually or with a script, identify movement patterns other than walking. With multiple pedestrians detected in this manner, it could potentially reveal conflicting gestures. The datasets which could be used for this could be: [nuScenes](#) [6], [OpenDriveLab/DriveAGI](#) [50], or [Challenge Of Out-Of-Label \(COOOL\) in Autonomous Driving](#) [1]

Another approach could be simulating this data. Software for this could be *Searchable Database of Potential Crash Scenario Models for CARLA and the Corresponding Sensor Data Feeds* ^{8 9} or less complicated using individual online models in both 3D or 2D animation ¹⁰

5.6 Signs detection dataset

Located signs in the traffic environment also have a significant impact on the correct and safe decision-making for the ego driver. A short list of datasets has also been compiled to fine-tune object detection models, if necessary. They also assist with understanding potential scenarios. They are all at Dataset-Ninja.com [lisa-traffic-light](#), [vietnamese-traffic-signs](#), and [gtsdb](#).

5.7 Large Language Model (LLM)

The selected LLM was *LLaMA2-7B-Instruct-hf* [43, 40]¹¹. The *LLaMA3.3-70B* model was also experimented with, but it required too much RAM. This model was selected due to its smaller size; however, it is still a relatively new implementation. The prompts could be quite long, aiming for a larger model, but a smaller model would be used initially to increase the number of development iterations.

The number at ‘B’ explains the size of the models. It indicates the number of parameters in billions, resulting in a model with 7 billion parameters. The ‘*Instruct*’ tells that the model is fine-

⁸deepdrive.berkeley.edu/node/811

⁹carla.readthedocs.io/en/0.8.4/

¹⁰Google Images: ‘3d stop gesture animation’, [freepik.com/vectors/2d-animation-character](https://www.freepik.com/vectors/2d-animation-character), [vecteezy.com/free-vector/2d-character](https://www.vecteezy.com/free-vector/2d-character)

¹¹[Meta LLaMA](#), [GitHub](#), [CodeLlama-7b-Instruct-hf](#)

tuned for following instructions. Compared with *Chat*’ or *QA*’ models, which are fine-tuned for those purposes. The *hf*’ tells the model is from Hugging Face.

5.8 Trajectory Planning

Common in robotics is the term planning, which involves planning a route rather than predicting what other entities are most likely to do. Instead of categorizing the driver’s action as right or wrong using binary classes, a more reliable measurement is to evaluate it using a trajectory. The evaluation can be done in two ways, supervised and self-supervised. The supervised version computes the difference between the trajectory and the ground truth. This difference can be calculated in various ways using Mean Square Error (MSE), Dynamic Time Warping (DTW), Frchet Distance, and other methods. The self-supervised method utilizes cost heuristics to determine the optimal trajectory (Eq. ??). Typically, in AV, the system penalizes collisions, speeding, and other infractions. [30]

5.9 Findings Summary

The findings are summarized to get an overview of what is needed to research, develop, or gather. This also helps list the chosen models and datasets.

Process	Status
5.2 Pedestrian Detection	Yolov8 (pre-trained)
5.2.1 Authority Classification	Safety vest detection (<i>For pilot study</i>)
5.2.3 Navigation Gesture Recognition	Yolov8n-pose (pre-trained) + class annotation + limited collected data + generated data
5.3 Ego Driver Action Classification	<i>Currently unsuccessful to implement accurately</i> → Annotation / GPS
5.5 Conflict Scenario Data	Annotation + generated data
Prioritization Model	<i>Further Research</i>

This implies explaining the pedestrians through text, so that another large language model could interpret the gesture and have the VLM interpret itself.

This pilot research study uses a small first-hand acted dataset captured with dash-cam footage in contrast to real-life scenarios from a 360° perspective. It is a zero-shot model, using visual-language models (VLMs), Complex scene information is required for the AV to interpret the data from the pedestrian accurately in some cases. Action detection and speed are not processed in the pipeline, as this information is supposed to be obtained from ULM (hardware). It is predicted only in the annotation. In future work, temporal pose classification will aid the model, as it can be easily trained, and pedestrian detection will also be beneficial.

6 Conflict Decision-Making: Design & Experimentation

Initially, the primary focus of the project was centered on the component of having a model draw a decision given a conflicting series of gestures across multiple pedestrians. They would, as an ego driver, either have the same or different authorities, which would indicate the correct commanded gesture to follow. Throughout the project, as knowledge of this problem increased, the data structure and type changed, which also altered the model's architecture. This is the reason for experimentation with varying architectures both in theory and in practice. This makes this chapter contain more experimental content, as most of the experiments were found to be too simple to solve this problem. They were all using a network with the pedestrians as input.

The experimentations of the conflict decision-making models are found at [tbosse20/mercedthesis, Github.com](https://github.com/tbosse20/mercedthesis)

6.1 Synthetic ‘Gesture’ Priority Training with FRIENDS

The idea of this experiment is to convert the information about the ego driver's action in a scene involving the person they obey into a priority model on a synthetic target priority list. A dataset of images featuring specific individuals and a fixed target priority list is used for experimentation. The corresponding action and gesture label will be used to train the model, allowing it to distinguish between individuals prioritized in the given crowd. Fig. 6.1 visualizes the concept using the people from the TV series FRIENDS. This series is selected because a dataset of images was needed that features the same people.

The synthetic dataset is made by assigning a random synthetic pseudo-gesture label to each detected person in each given image. Note that the gesture label can reappear for multiple people in the image, which can confuse the model. The pseudo-gesture is simply a number and an ARUCO marker, but it functions as a specific gesture class. The ARUCO is a type of QR code with a value. Each image is also assigned a pseudo-action label. The action label corresponds with the gesture label of the person in the image with the highest priority, making them the target label. The remaining people are noise labels. In the example under **Synthetic ”gesture” label**, Monica and Rachel are detected. They are both assigned a random pseudo-gesture. The pseudo-action label is set to match Monica's gesture label, since she is the highest-ranked on the *Target priority* list among her and Rachel.

The **Scene priority** shows how all gestures are detected, and the target person is found by matching the action label. This gives the data shown in Tab. 6.1. Here, it is visualized who was obeyed in each crowd or scene. E.g., ‘Ross’ is prioritized highest in a screen with everyone. Or both ‘Chandler’ and ‘Ross’ are obeyed in example 3. This is because ‘Chandler’ has the same gesture label as

‘Ross’. This makes the data more realistic and introduces the ‘*Shared*’ conflict concept, as explained in Tab. 4.1 in Cap. 4.

This is a simple algorithm, but it also marks the beginning of a research effort. It demonstrates a method of matching action-pedestrian pairs based on their gestures. Instead of using specific names, everything could have been pseudo, but using images and people helps prove the concept.

Synthetic “gesture” label priority training

Target priority: Ross, Monica, Chandler, Joey, Rachel & Phoebe

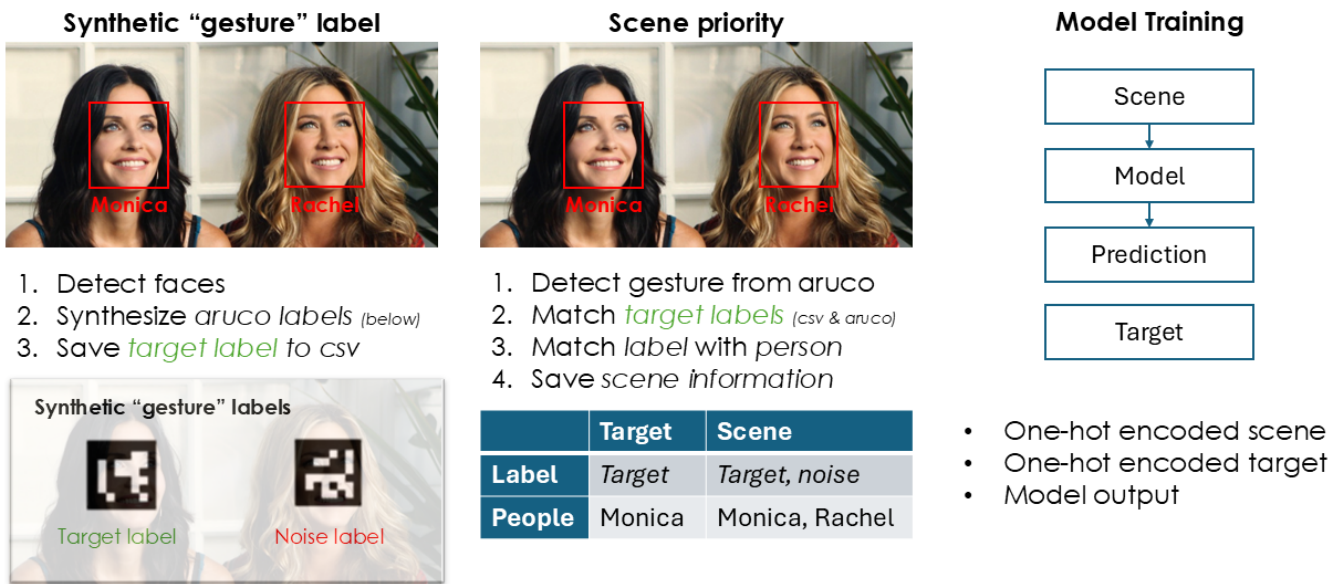


Figure 6.1: Example of synthetic gesture data on FRIENDS. **Synthetic “gesture” label** shows how the gestures are assigned to the target and noise people in the scene, using the **Target priority** list. **Scene priority** matches the action label with the gesture to find the obeyed person in the crowd.

Target priority: <i>Ross, Monica, Chandler, Joey, Rachel, Phoebe</i>	
Obey	Crowd
['Ross']	['Ross', 'Rachel', 'Joey', 'Monica', 'unknown', 'Chandler']
['Joey']	['Joey', 'Phoebe']
['Chandler', 'Ross']	['Ross', 'Rachel', 'Chandler', 'Phoebe']
...	...

Table 6.1: Synthetic gesture labels using the FRIENDS characters

6.2 Model Architectures

As mentioned, the architects of the models change over time. This lies in the input to the model.

6.2.1 Authority multi-class multi-layer perceptron (MLP)

Since multiple classes can be present in a single scene, a one-hot encoding was suggested to merge the classes. This also includes padding for the data to fit, as some scenes could have more authorities than others. So in a scene with A_1 and A_3 , with four different authorities, the input of the model would be:

$$\mathbf{x} = \begin{bmatrix} 0 & 1 & 0 & 1 \end{bmatrix} \quad (6.1)$$

The model's output is the gesture matching the driver's action. This makes the label of the model with \mathbf{G}_1 with one-hot encoding and five different gestures be:

$$\hat{\mathbf{y}} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \end{bmatrix} \quad (6.2)$$

The notation for the output and label is:

$$\mathbf{y} \in \mathbb{R}^{d_o}, \quad \text{where} \quad d_o = |\mathbf{G}| = |\mathcal{A}| \quad (6.3)$$

This suggests that the model should be a fully connected neural network to understand the relationship between the authorities.

6.2.2 Authority-gesture Multi-class Transformer

After gaining a better understanding of the data, it was acknowledged that the scenes can contain multiple subjects of the same type. This makes the multi-class model unsuitable, as it cannot contain more than one of the same authority. Instead, each authority's input will be a vector, making the data suitable for a running model, such as LSTM or Transformer. As the input can contain multiple authorities of the same type, each gesture is added to the input to distinguish the properties of the gesture type and the authority hierarchy. This also allows the system to explore if gesture types and/or several authorities combined can have a higher priority than an entity from a single authority.

For two categorical features that are one-hot encoded, let:

$$\mathbf{A} = \begin{bmatrix} \text{Police} & \text{Civilian} & \text{Firefighter} & \text{Vest} \end{bmatrix} \quad \text{and} \quad \mathbf{G} = \begin{bmatrix} \text{Stop} & \text{Left} & \text{Right} & \text{Go} & \text{Idle} \end{bmatrix} \quad (6.4)$$

For the pair $(\mathbf{A}_C, \mathbf{G}_I)$, the one-hot encodings are:

$$\mathbf{A}_C = \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix}, \quad \mathbf{G}_I = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \rightarrow \quad \mathbf{A}_C \otimes \mathbf{G}_I = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (6.5)$$

The label of this model is the same as that of the MLP model, as shown in equation 6.2. This makes the input of a scene with the subjects A_1 and A_3 with four types of authorities with the gestures

G_1 and G_4 with five types of gestures be:

$$\mathbf{x} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (6.6)$$

The notation for the input \mathbf{x} is then:

$$\mathbf{x} \in \mathbb{R}^{T \times d}, \quad \text{where } T = |P|, d = |A| \cdot |G| \quad (6.7)$$

6.3 Formula Definition

After iterations of data structure understanding and experimentation, a formula is constructed to increase understanding and formulation of what the problem applies to and what is sought to be solved. This formula helps explain the abstract idea of the problem, data, and solution. It includes the inputs of the traffic environment and the corresponding optimal decision made. This is also an extension of the formula proposed in **Sec. 3.2 Final Project Scope and Formulation**. The inputs are defined as follows.

all objects \mathbf{O} in the given frame \mathbf{I} are detected:

$$\text{Detection}(\mathbf{I}) \Rightarrow \mathbf{O} = \{(c_i, b_i, s_i) \mid i = 1, 2, \dots, n\} \quad (6.8)$$

where:

- $\mathbf{I} = \mathbb{R}^{h \times w \times c}$ is the image of the dimensions height, width, and channels.
- $c_i \in C$ is the class label from a set of categories C , here referred to as authorities R .
- $b_i \in \mathbb{R}^4$ is the bounding box coordinates (e.g., (x, y, w, h)).
- $s_i \in [0, 1]$ is the confidence score.

A subset of all the pedestrians \mathbf{P} detected in the scene is made:

$$\mathbf{P} = \{i \in \mathbf{O} \mid i_c \text{ is } c_{\text{pedestrian}}\} \quad (6.9)$$

For each pedestrian, authority \mathbf{A} and gesture \mathbf{G} is classified:

$$\mathbf{P}_{\mathbf{A}}, \mathbf{P}_{\mathbf{G}} = \{i \in \mathbf{P} \mid (\mathbf{A}(i), \mathbf{G}(i))\} \quad (6.10)$$

The optimal action, which is the action the driver should take, is then chosen from the set of possible actions \mathcal{A} by maximizing a scoring function f with the given Pedestrians \mathbf{P} . It is maximized to achieve the highest priority:

$$a^* = \operatorname{argmax}_{a \in \mathcal{A}} f(a \mid \mathbf{P}) \quad (6.11)$$

In summary, given the pedestrians \mathbf{P} composed of authority-gesture pairs (from \mathbf{A} and \mathbf{G}), the model selects the best action a^* to execute from the list of possible actions \mathcal{A} by finding the action that maximizes $f(a \mid \mathbf{P})$.

The action can be represented by both a finite, predefined set of classes, an open vocabulary in zero-shot learning, and a trajectory τ as GPS location, or something similar.

6.4 Findings Summary

These data structures and architectures have predefined authorities and gestures, making them incapable of open vocabulary and zero-shot settings. They rely on training to reveal the nature behind an idea of a certain fixed *priority list*. However, it appears that the decision-making in this domain is more complex than initially thought. Ultimately, it is believed that the problem is too complex for architects like these to handle. This appears to be more aligned with LLMs, as they possess a deeper understanding of legal culture and can be utilized in a zero-shot setting. These models rely on training. However, it remains essential for the process and is worthy of inclusion. As mentioned in the Introduction, the focus shifted, which laid the conflict decision-making to rest, and due to time limitations, left for future exploration.

7 Datasets: Preparation & Creation

Here, I outline the thought behind the datasets' samples, annotations, and structure. In total, there are three different datasets, two of which are also explained in the VLM Evaluation Paper [4]. Only the ATG dataset is utilized in this project, with the ITGI and Act-CANG serving as extensions for further research. The data structure and content are also defined by the experiments from Cap. 6 [Conflict Decision-Making: Design & Experimentation](#).

7.1 Acted Conflicting Authorities & Navigation Gestures (Act-CANG)

Additional data needed to be gathered as part of the project, as it was impossible to obtain enough data to include enough scenarios to enable a detailed evaluation. Sufficient data was collected to gain a better understanding of how elements affect the scene and identify the essential elements to include in the dataset. This section describes the reasoning for creating the dataset.

The dataset will initially consist of 20-25 scenes distributed across the different types of scenarios described in Table 4.1, as these are the scenarios most likely to occur. The dataset can be expanded after researching a system that uses only these. To make it easy, fast, and realistic, the aspect in the "data gathering" section of bullet point 6 in the "problem description" has been selected. This is the second category in the "Three-stage realism" problem description. This allows the system to be tested on real people, but does not require the inclusion of fundamental authorities. This data can be collected from real-life scenarios, such as attending concerts, experiencing traffic jams, visiting airports, attending sports events, and visiting theme parks. However, this will drastically decrease scene control and only slightly increase realism. The authority and gesture in each scene are selected to match real life. Notice that the data cannot be individual images, as the gesture and action classification require temporal information. The realism level aims to be the highest while staying in a safe and controlled environment and setting.

The annotation software, dataset links, and further guidelines for the annotators are found on [tbosse20/navigation-gesture](#), [Github.com](#)

#.	Qty.	Scene Config.
<i>Unique (10 samples)</i>		
1.	3	$\mathbf{A}_{ 3 }\mathbf{G}$
2.	2	$\mathbf{A}_i\mathbf{G}_\emptyset + \mathbf{A}_{i<}\mathbf{G}$
3.	5	$\mathbf{A}_i\mathbf{G}_\alpha + \mathbf{A}_{i<}\mathbf{G}_\beta$
<i>Shared (6 samples)</i>		
4.	2	$\mathbf{A}_{ 2 }\mathbf{G}_\alpha$
5.	1	$\mathbf{A}_{ 3 }\mathbf{G}_\alpha$
6.	1	$\mathbf{A}_{ 2 }\mathbf{G}_\alpha + \mathbf{A}_\beta\mathbf{G}_\beta$
7.	2	$2\mathbf{A}_i\mathbf{G}_\alpha + \mathbf{A}_{i<}\mathbf{G}_\beta$
<i>Ignore (2 samples)</i>		
8.	1	$\mathbf{A}\mathbf{G}_\alpha \neq \mathcal{A}_\alpha$
9.	1	$\mathbf{A}_{ 2 }\mathbf{G}_\alpha \neq \mathcal{A}_\alpha$
<i>Disagree (2 samples)</i>		
10.	2	$\mathbf{A}\mathbf{G}_{(\alpha,\beta)}$

Table 7.1: Authority-gesture Categories. $M \times kX_{|n|}$: M scene variants with k number of X being $|n|$ different types. Non-specification equivalents any or a single value. Symbols explained in Table 4.1. *Example explaining nr. 3:* Defines 5 scene variants, each consisting of two pedestrians. First pedestrian with an arbitrary authority-gesture pair $\mathbf{A}_i \mathbf{G}_\alpha$. The second pedestrian possesses a higher authority and different gesture than the first pedestrian $\mathbf{A}_{i>} \mathbf{G}_\beta$.

7.1.1 Requirements

It is examined what is required to create a well-structured dataset that contains all necessary features and excludes unnecessary data. The section argues and lists the required and needed features regarding authority-gesture pairs, driver action decision-making, and resources for the data creation.

7.1.1.1 Authorities and Gesture

The dataset requires a set of possible authorities and gestures to include. Since the model needs to learn the correct rank of the authorities, a certain number of different authorities must be included to avoid relying on luck. Since only one of these arrangements is in the correct order, the probability of randomly selecting the proper gestures is given by:

$$P_{\text{correct}} = \frac{1}{n!}, \quad \text{where } n \text{ is number of authorities}$$

To make sure the priority list is not sorted by chance, the n is found where the formula drops below $\alpha = 5\%$, being the standard p -value:

$$|A| = \operatorname{argmin}_{n \in \mathbb{N}^+} \left\{ \frac{1}{n!} < 5\% \right\} = 4 \Rightarrow \frac{1}{4!} = 4.16\%$$

Using four types of authorities, the chance of finding the correct order by chance is 4.16%.

7.1.1.2 Scenario Construction Rules

The final selection of scenes is based on a predefined set of rules designed to assess how effectively the model learns a specific system of rules. They decide on the action of the driver. This approach increases the realism and reliability of the scenarios and remains. This only makes sense to do from a trained priority model, since a zero-shot model follows a more abstract cultural rule and laws. The following rules are listed in order, with notes provided below.

1. The car stops for any hard obstacle, no matter the incoming authority and gesture.
 2. The car prioritizes the ‘Stop’ gesture above any authority.
 3. The car follows the predetermined hierarchy:
 1. *Police officer*, 2. *Firefighter*, 3. *Safety vest*, and 4. *Civilian*.
 4. In dilemma and paradox cases, the individual driver participant decides the most appropriate action, based on feelings, intuition, and prior/external knowledge.
- Pedestrians without a gesture are not considered.
 - Crowded gestures do not exceed gestures from higher ranks.
 - Same authorities with different gesture reasons, the decision from other sources, for example, ‘Obstacles’.
 - Gestures other than ‘Stop’ do not affect the priority.
 - Selected authority and gesture tries to match realistic, possible scenarios.

It was challenging to figure out the number of different gestures and authorities after setting up the possible scenes. This was because it could be random, as every scenario is technically feasible. However, to better explain the specific number of authorities and gestures chosen for the scene, an attempt was made to replicate realistic scenes. The best option would be real-life data. I am, however, creating this dataset because of this lack, so the next best thing was to draw on previous experiences or the most likely and common scenarios that have been considered.

An attempt was made to prepare a table for the collection day. This would show the accurate authority and gesture for each scene. This would also quantify the different types of scenes, making it easier to gain an overview of the content and ensure that all relevant authorities were included. This was, however, removed, since it was not very accurate or reliable, and caused a lot of confusion instead.

7.1.1.3 Resources

The ideal location would be a fixed road environment, giving us complete control over the scenarios. It would be efficient with other parked cars, as they could function as obstructions. It needs to include an intersection to allow for cases of turning. This also makes it easier to include conflicting gestures, such as ‘Left’ and ‘Right’. The equipment I need for creating this dataset is 2 of 2-3 different types of authority uniforms, 4 cones, 1 x 2m - 3m measuring tape, 1 dash and rear camera (HD \leq), 2 cars, and 3-4 participants.

7.1.2 Annotation

We provide instructions on how the data is annotated, including explanations of possible misunderstood annotations. Section 10.1 in Discussion argues a more in-depth reasoning for this framework. In short, the data is annotated by category and caption for each newly started movement.

The dataset was annotated in collaboration with undergrad members of the Mi3 Lab at the University of California, Merced. This helped speed up the annotation process, but it took a considerable amount of time to ramp up. Besides video post-processing, it included assigning videos, creating an annotation guide, creating an annotation framework, and creating annotation software. All these elements had to be accurate from the start, since it would be difficult to correct the launch of the annotation process. The annotators were handed this section 7.1.2 *Annotation* and the README file as their guidelines. Of course, they could ask questions throughout the process.

7.1.2.1 Argumentation

Annotation can be applied in two dimensions: *temporal* and *body*, where body includes internal degrees of detail. In terms of frame, the annotation can be applied to every single frame, fixed sequential frames, a dynamic frame or timestamp, or the entire video. The body can be divided into groupings that can be combined as a graph. The *body-grouping* can be every joint, sub-parts (*individual finger*), grouped parts (*all fingers*), appendages (*hand*), limb (*arm*), or the full body. Body sections can also be excluded at each degree. Explaining the body is complex, considering the details a driver wants to obtain useful information. Body detail descriptions can only be qualitative as they are not measured. These details are the following, including methods or aspects to describe: Reference point (*pedestrian, ego, car, road*) in terms of distance, position, speed, orientation (*egocentric, compass, clock, angles, facing*), and physical state descriptors (*tucked, flat*). [4]

A more in-depth study is needed to understand the practical details. Still, it is thought that, to interpret traffic gestures accurately enough to interpret the gesture enough, without too much additional information, the description needs the following information in formulated details:

The description needs temporal information, as some gestures are dynamic. This annotation is separated for each newly interpreted gesture. For example, a pedestrian is transitioning from being idle, lifting their hand to gesture a ‘Stop’ gesture. They lower their arm back to an idle state. This

sequence would have five different annotations: ‘Idle’, ‘Transition’, ‘Stop’, ‘Transition’, and ‘Idle’. This allows for anticipating captions and merging the transitions to pre- and post-gestures for additional use.

The gesture can differ down to the individual finger, but mainly in the upper body. Degrees of position, orientation, and physical state don’t significantly impact the meaning of the gesture. Movement speed emphasizes the intent and seriousness, but doesn’t change the gesture.

What’s more important are the reference points, distance, and direction. Using compass directions from the ego driver’s perspective could provide detailed spatial references, but may lead to ambiguity or misinterpretation.

The body and interpretation descriptions are separated to allow for additional usage of the dataset. They can be used individually if the application focuses on human movement rather than gestures, or merged if needed. As with the transition of gestures, this threshold can potentially be used in future work to teach the model to anticipate specific gestures or respond more quickly. Or it can be merged simply by expanding the class label to adjacent transitions.

The focus is currently on the ego driver. To reduce the annotation process time, we only annotate those gestures. The annotation can be expanded by easily finding excluded annotations. I originally considered using a `ego_mask` to find them again. However, this would mean all current annotations would be True at the `ego_mask` flag, making no distinction anyway. And if there were any False gestures, those would need to be processed anyway. To find the excluded gestures for future annotation or evaluation, we filter out the bounding boxes for each frame that do not have annotations. This will require processing all the bounding boxes, not just the included gestures. *Note: Maybe a script can annotate this quickly anyway.*

7.1.2.2 Guide & framework

The description of the selected method depends on the specific situation, and it is up to the annotator to interpret it accordingly. 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.

7.1.2.3 Caption Instruction

The instruction considered when captioning the ground truth descriptions of the body movement and interpretation is formulated as follows, inspired by [4] (examples can be found in Sub-section 7.1.2.4):

Transcribe the pedestrians' upper-body posture and expressive gestures, specifying the intended recipient (*e.g., gesturing 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.

7.1.2.4 Caption Examples

The following examples provide an understanding of expressing the human body, its movements, and relationships to other objects, as well as the usage of *<skip>*. The examples are inspired by [4].

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.*)

7.1.2.5 Gesture & Authority Classification Labels

The classification is limited to the ego driver and overlooks 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' to mean drive across an intersection, 'Left' and 'Right' to mean turn, and 'Advance' to mean drive wherever. 'Idle' is not being used, as I only focus on direct gestures; however, it is still included for clarity. The gesture classification classes are inspired by [4]:

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

Table 7.2: 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.

#.	Authority	Color
0.	Officer	Blue
1.	Firefighter	Red
2.	Civilian	White
3.	Safety vest	Yellow
4.	<i>Unlisted</i>	<i>Gray</i>
5.	<i>Unclear</i>	<i>Gray</i>

Table 7.3: The available authorities included in the dataset are listed with corresponding ID and color. The authorities may also be annotated separately, with only the authority class for each pedestrian ID. For now, I merge them.

7.1.2.6 Format

The annotation includes bounding boxes and IDs for each pedestrian, if multiple pedestrians are in the scene. To simplify the concept for now, the gesture class only describes the gesture towards the ego driver. A description is provided for each pedestrian separately, while 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 Table 7.4.

video	camera	ped_id	start_frame	end_frame	auth_id	gest_id	body_desc	interpret_desc
<i>str</i>	<i>str</i>	<i>int</i>	<i>int</i>	<i>int</i>	<i>int</i>	<i>int</i>	<i>str</i>	<i>str</i>
video_04	front	24	41	56	1	2	"..flat hand.."	"..stop.."
video_04	front	24	57	63	1	12	"..nods head.."	"..approve.."
video_04	back	71	45	56	0	10	"..points.."	"..go there.."
video_04	back	52	48	46	2	13	"..spins.."	"..unknown.."

Table 7.4: Annotation format example including multiple pedestrians. The annotations contain the features: Name of video (`video`) , Name of camera (`camera`) , Pedestrian ID (`ped_id`) , Start frame at movement (`start_frame`) , End frame at movement (`end_frame`) , Authority class ID (`auth_id`) , Gesture class ID (`gest_id`) , Body movement description (`body_desc`) , Interpretation description (`interpret_desc`).

7.1.2.7 Driver Action

Instead of classifying the data with finite classes, it is continuous data in the form of GPS locations. As a supplement to the videos and annotations, we recorded the location of the ego vehicle. We used the app GPS Logger¹, which records the longitude, latitude, speed, altitude, and direction of the phone. This increases the properties and usability of the dataset. I aim to utilize this as a numerical response and for the action labels in the conflicting samples. The post-processing included syncing the GPS data with the recording. An undergrad was in charge of this task, with guidance.

7.1.3 Description

The complete dataset comprises 38 samples, ranging from 12 to 56 seconds. Each sample includes a thumbnail, 360° view across four cameras, and a class and caption annotation for the specific movement for the ‘Front’ camera, explained in Section 7.1.2. Fig. 7.2 shows an example of a conflicting scenario. To assess the recording of the data, we had a group of four people with one ego vehicle and one other vehicle. The footage is recorded from the ego vehicle using a 2022 Tesla Model Y, equipped with four cameras to achieve 360° view. This is visualized and explained in Fig. 7.1. The actions of the ego driver were decided by the individual driver, instead of following a fixed set of rules. This was selected to increase the realism of the scene. This introduces the prediction ambiguity, where the model can potentially make a safer and better decision than the ground truth.

¹[GPS Logger](#), [BasicAirData](#), [Google Play Store](#)

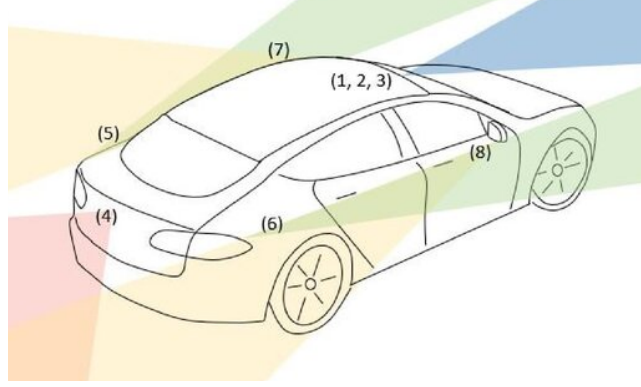


Figure 7.1: Both datasets were recorded using a 2022 Tesla Model Y, which includes four cameras as ‘Front’ (1, blue), ‘Back’ (4, red), ‘Left_repeater’ (7, yellow), and ‘Right_repeater’ (8, yellow). The ‘Front’ and ‘Back’ cameras are angled directly, whereas the ‘Repeater’ cameras are angled to record mainly behind the car. The illustration is a crop of Fig. 1 in [35].



Figure 7.2: Example of Act-CANG ‘video_31 Front’ (crop), showing a supposed ‘police officer’ (left) and a civilian (right) gesturing the ego driver towards opposite directions.

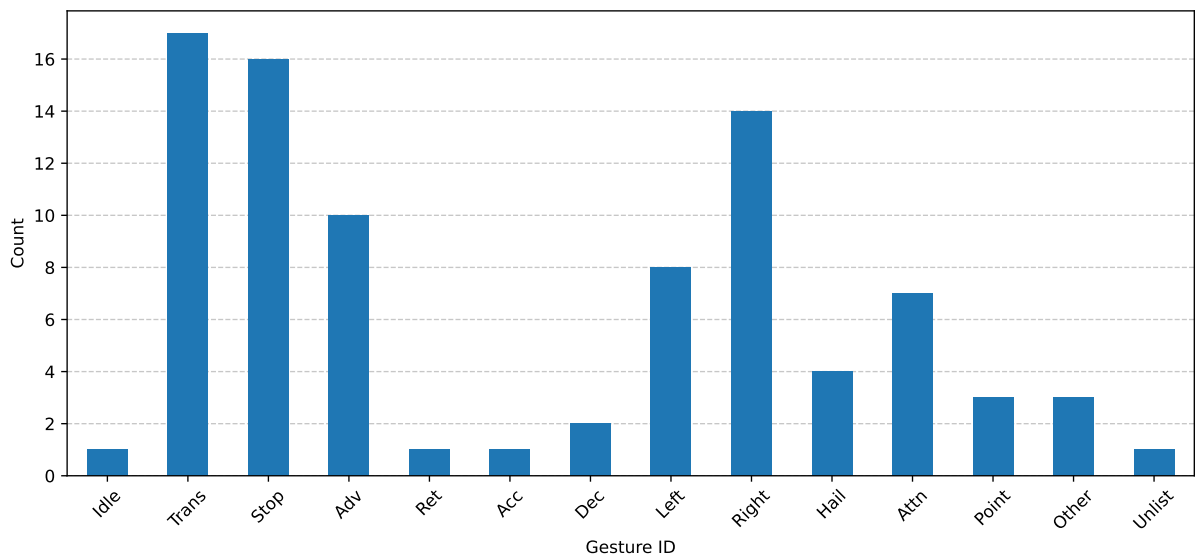


Figure 7.3: Distribution of occurring gestures in the Act-CANG dataset. All gestures listed in Tab. 7.2. Each count is only of every occurrence, and not for every frame.

7.2 Acted Traffic Gestures (ATG)

Due to the limited data available on pedestrian gesturing to the ego driver, additional samples were collected firsthand. As the model learns the relative position of the joint points of the person, the initial dataset will include data from a single person, which should be enough to get an estimate. With the option of extending it to multiple people if the model can not be generalized enough.

The goal for the dataset is to include the selected types of gestures from multiple distances and angles in front of the car and merge the movements with various actions, such as running, walking, startled, etc. As the data is collected in video format, a sufficient number of frames should capture a high diversity of movements.

7.2.1 Pedestrian-camera Distance

To determine how far away recordings of gestures should be collected, a safe distance for a braking vehicle was calculated. This gives a good idea of the distance at which the pedestrian should be detectable to react. The speed limit in urban areas is around 50 km/h, making it a good measurement to utilize in this test. The computing time, referred to as reaction time, is set to 0.1 seconds.

The brake distance is computed at 50 km/h to be 14.5 m. Adding the distance driven for computation/reaction time, with 0.1 s being 1.389 m. A total of 15.889 m traveled. This indicates the model needs to be able to detect and perceive the gesture of a pedestrian from at least this far away. For good measurements, the number is rounded up to 20 m.

The brake distance is computed at 20 km/h to be 2.24 m. Adding the distance driven for computation/reaction time, with 0.1 s being 0.556 m. Traveling a total of 2.796 m.

The lowest driving speed in a parking lot, where this scenario is usually in effect, is around 5 km/h. The brake distance is computed at 5 km/h to be 0.14 m. Adding the distance driven for computation/reaction time, with 0.1 s being 0.139 m. Traveling a total of 0.279 m.

An initial test was conducted to determine the distance at which the camera/car could detect and estimate the pose of a pedestrian. A simple test was conducted with a participant waving at distances of 1 m, 5 m, 10 m, 15 m, and 20 m from the camera. The video was recorded in a lab hall with multiple light sources to illuminate the participant. The pose estimation model was run on this video. The pose estimation distance test using 'yolov8-pose' showed that it can detect the participant with a confidence of 0.8 to 10 meters. However, this decreases to 0.4 at 15 meters, and the waving arm is not detected correctly. The detection fails above 15 meters. This proves that the data should only be collected from 15 meters away from the car. However, a few recordings were still collected from 20 meters away to advance the dataset.

A car's hood is approximately 1.1 meters long, which also makes the shortest distance between the camera and a pedestrian located in front of the vehicle.

The final selected distances between the participant and the camera from which data will be collected are 1 - 5, 10, 15, and 20 meters.

7.2.2 Resources

The ideal location would be a common urban area or a similar setting, with sidewalks close to the road on both sides. This will give a good indication of the pedestrian's realistic position relative to the ego car. Having this or multiple scenes like this, both in and outside of shadow, can make for more diverse data. The road must be long enough to obtain data from a distance of at least 20 meters and 5 meters wide, thereby increasing the number of available positions. The necessary equipment includes cones to mark distances, a high-definition (HD) camera, a tripod, and a minimum of one participant.

7.2.3 Gesture List

The final datasets include the following gestures, which were performed across two iterations. All are directed at the ego driver and include gestures relevant to other hypothetical vehicles.

Category	Gestures
Basic	Stop; Reverse; Advance; Hail; Attention; U-turn
Contextual Depends on the environment or car features	Left; Right; Proceed; Accelerate; Decelerate; Pull over
Head movement Using the head to communicate	Left; Advance; Right; Affirmative; Negative
Simultaneously Multiple gestures and directions	Stop, advance; Stop, stop, advance (head)
Sequential Gestures in sequence	Stop, and drive; Me, there, not; You, stop
Pointing Precise locations	Go there; Look there
Irrelevant / Social Non-navigation gestures	Complain; Thanks; Idle; Shrug; Apology

Table 7.5: Categorization of hand gestures. The ground truth labels are found in Section 7.1.2, as these categories are not the labels.

Other samples were excluded from the recordings due to their increased complexity in interpretation and meaning. The category 'Question or instruct' distinguishes the gesture as a question or an instruction given to the ego driver. An example could be 'Location', where a pedestrian points at a location, either to ask if the ego driver is going to that location or to instruct them to go to that

location. Or ‘Crossing’, where the pedestrian requests or informs the ego driver that they are crossing the road.

It is challenging to consider all possible ways to communicate with a driver. This requires considering multiple scenarios in both locations and cultures, further proving the importance of this application. To assist this list, real-life scenarios should also be collected to replicate or use as optimal data.

7.2.4 Post-processing

The post-processing of the video includes cutting, annotation, and data processing. They are cut to fit only the gesture with a time margin before and after. A script was created to handle CSV files containing the cuts, eliminating the need for manual handling. This was done because the original data needed to be recut. This required reprocessing all videos to find the specific frame, as the annotations were already completed. Using a fixed CSV file would avoid this issue if needed again, and also make it easier to adjust cuts after reviewing the clips. The annotation was initially done for every 8th frame combined. As explained in the paper [4], this was done to increase the rate of specific frames where the gestures were detected. The paper proved that the VLMs would not work with that few frames available, so annotations were added for each clip to evaluate the enhanced VLMs. Additionally, the paper demonstrates VLMs’ inability to classify given classes, so this section of the dataset, for starters, only contains classification annotations.

7.2.5 Description

The described dataset is an extension of the original VLM evaluation paper [4]. I extend this dataset with 118 annotated samples of single gestures, bringing the total to 127 samples for this evaluation. However, I only evaluate on the extended. The extended dataset includes these classes with the amount of each: $3 \times$ ‘Idle’, $26 \times$ ‘Stop’, $22 \times$ ‘Advance’, $19 \times$ ‘Reverse’, $7 \times$ ‘Decelerate’, $3 \times$ ‘Left’, $8 \times$ ‘Right’, $6 \times$ ‘Hail’, $10 \times$ ‘Attention’, and $14 \times$ ‘Other’. The videos named 118 and higher in the extended version are sequential and have not been annotated, as they require a different type of annotation. This is for more complex gesture understanding and will be on hold until the more basic ones can be classified first. There are 22 ‘Sequence’ samples. The extended dataset has a total of 140, making the whole dataset consist of The data is set in varying direct sunlight and shadow in an empty part of a parking lot. A single participant is 1 to 20 meters away from the camera. A cone is marked approximately every 5 meters.

7.3 Instructive Traffic Gestures In-The-Wild (ITGI)

The dataset is a part of the VLM Evaluation Paper [4], but in short, it contains police enforcement performing real navigation gestures in-the-wild traffic. After the submission of the paper, navigation gestures were annotated according to the CANG dataset, with the help of the undergrads at UC

Merced Mi3 lab. The recording vehicle is the same as explained in the Description Section 7.1.3 in the CANG dataset.

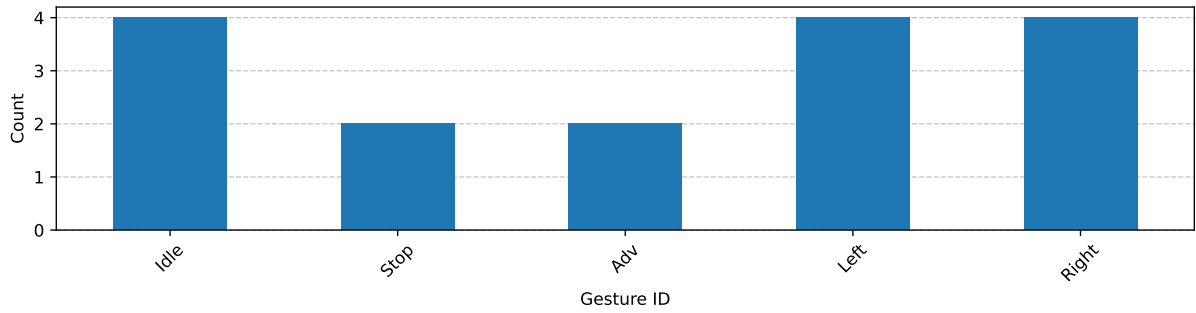


Figure 7.4: Distribution of occurring gestures in the ITGI dataset. All gestures listed in Tab. 7.2. Each count is only of every occurrence, and not for every frame.

8 VLM Enhancing for Navigation Gestures

We showed proof that the current VLMs are incapable of captioning and classifying human gestures in a zero-shot setting. This study is written as a scientific paper ‘*Can Vision-Language Models Understand and Interpret Dynamic Gestures from Pedestrians? Pilot Datasets and Exploration Towards Instructive Nonverbal Commands for Cooperative Autonomous Vehicles*’ [4] by Tonko Bossen (me), with Andreas Møgelmoose and Ross Greer as supervisors. It is attached as Appendix A. The final paper was refined based on the reviews of the submitted paper. The reviews are summarized and analyzed, along with an improvement plan, as Appendix C.

This paper lays the foundation for its primary focus: enhancing and evaluating VLMs for captioning and classifying human navigation gestures in a zero-shot setting and traffic scenario.

The code for the paper is found: [tbosse20/gestyLM_val, Github.com](https://github.com/tbosse20/gestyLM_val)

The enhanced aspect is found in the branch *Enhance*: [tbosse20/gestyLM_val/tree/enhance, Github.com](https://github.com/tbosse20/gestyLM_val/tree/enhance)

8.1 Enhancing Methods

This chapter explains the methods used to ‘enhance’ the VLMs to avoid fine-tuning. I discuss the implementation of each of the two selected methods ‘*Pose Projection*’ and ‘*Supplementary Body Description*’. These methods were selected ...

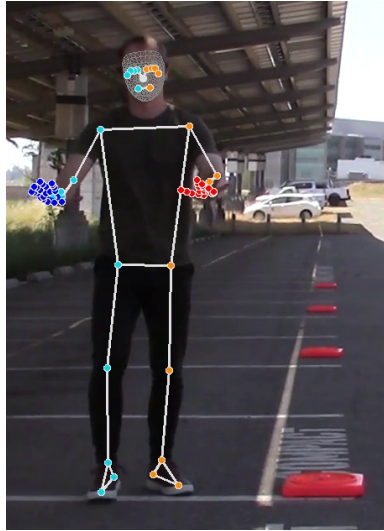


Figure 8.1: Sample video_20, ‘Decelerate’ visualized with the projection- and the SBD-method is: “*They’re face’s tilted down. Left hand is right and below their face with the palm facing down. Raised fingers are: Thumb, Pinky. Right hand is left and below their face with the palm facing up. Raised fingers are: Thumb, Index, Middle, Ring, Pinky.*”. The cones each indicate a distance of approximately 5 meters.

The enhancements are built upon the VLM VideoLLaMA3-7B [51], since it showed the best results in the evaluation [4]. Each method supplements the VLM to enable it to caption pedestrian

navigation gestures. The primary enhancement methods are *Pose Projection*, *Supplementary Body Description (SBD) with VLM*, and the two combined.

The prompt aims to endow the caption to classify the navigation gesture using the available classes in Section 7.1.2. However, I removed some classes since they were more complex. The ‘Transition’ class, since the video is annotated as a whole, and ‘Pointing’, since some gestures include pointing, which can be misleading, since this study excludes pointing. *Early experiments showed a strong tendency to over-predict ‘Pointing’, also resulting in removing this.*”

We evaluate using an NVIDIA GeForce RTX 4090 24GB with float16 precision and Flash-Attention 2.0 [12]. The VLM’s hyperparameters are temperature at 0.2 and 512 maximum new tokens same setup used in the evaluation [4].

8.1.1 Supplementary Body Description with LLM

A preliminary experiment was conducted, parsing a manually corrected description of the participant using these degrees to the large-language model, ChatGPT-4o. The anticipated classification of video_20 should be ‘Decelerate’, but was instead ‘Stop’. Another sample video_27 where the participant waved their arms above their head should be classified as ‘Attention’, but again it was classified as ‘Stop’. This gives insight into language models’ capability to understand human bodies and the meaning of their gestures. Still, body descriptions should be explored using VLMs since they can merge temporal visual and textual information and hopefully interpret gestures correctly. It was thought that this setup would also be used for a complete experiment, but due to the preliminary experiment, it was under-prioritized.

8.1.2 Pose Projection

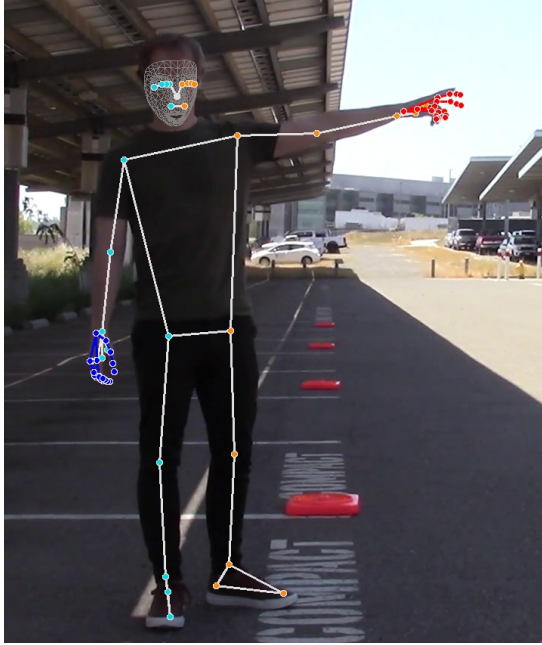
This method enhances the VLM by projecting visual representations of pose estimations onto the person. It includes the edges and vertices of the face mask, hands, finger joints, and general body pose.

The selected pose estimator is MediaPipe Holistic [29], as it is an out-of-the-box model that projects and estimates both the face, body, and hands. Other methods, such as Ultralytics YOLOv11-Pose [22], only estimate the body and face without the hands, and OpenPose [8] is too large a model for this initial single-person evaluation.

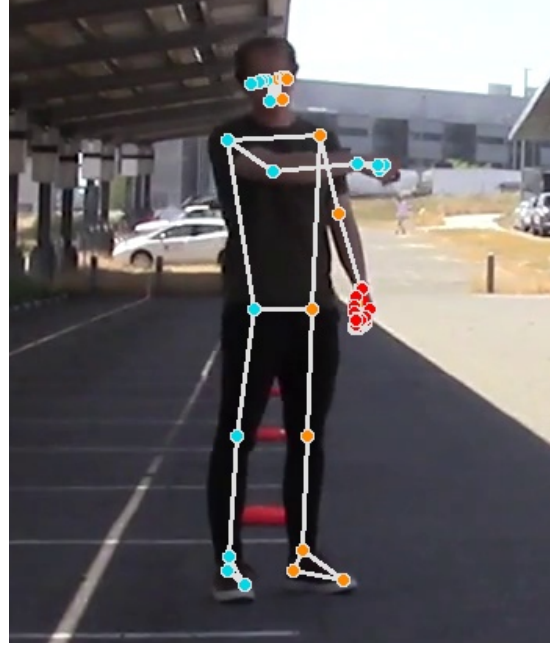
This method extends the prompt with the phrase: *“The pose is projected upon the person, to help understand their pose.”* for the VLM to understand the usage of the pose skeleton.

8.1.3 Supplementary Body Description with VLM

This method, referred to as SBD, enhances the VLM by supplementing the prompt of each frame with ‘natural language’ to describe the person’s body position roughly. As I advance the VLM in zero-shot classification, this method purely describes the body without interpreting any specific details. Then,



(a) Sample video_100, ‘Right’. “They’re face’s tilted down. Left hand is left and below their face with the palm facing the camera. Right hand is right and below their face with the palm facing the camera.”



(b) Sample video_102, ‘Right’. “Left hand is right and below their face. Right hand is right and below their face with the palm facing right.”

Figure 8.2

the interpretation is left for the VLM or an LLM. As this method employs natural language, it provides a rough description of the body rather than using precise distances (this can be explored in a future experiment).

We use pose landmarks to hard-code descriptions of the person’s face and hands. Initially, I used the refined face and hand landmarks, but due to low detection (argued in Subsec. 8.2), I chose to use the face and hand landmarks from the pose, since it was detected more often. The description details in-capture the essence of the essential body parts to interpret gestures. I describe the following. 1. The faces’ and palms’ direction in six states (*facing*: ‘Camera’, ‘Left’, ‘Right’, ‘Up’, ‘Down’, ‘Back’). 2. Each hand’s position relative to the face vertically in three states (‘Above’, ‘Horizontal’, ‘Below’) and horizontally in three states (‘left’, ‘vertical’, ‘right’). The left and right are described from the ego driver’s point of view. I flip the z-axis to ensure this.

Initially, the depth of the hand and the fingers raised were also included. The depth of the hand was in three states (‘In front’, ‘Beside’, ‘Behind’), using the ratio of the hand and face. However, this was inaccurate in cases where the hand was tilted, resulting in a smaller bounding box. The raised fingers were removed since the fingers of the hands were unreliable and were only detected in approximately 45% of the frames (Subsection 8.2). We had each hand’s fingers in multiple-label classification (‘Thumb’, ‘Index’, ‘Middle’, ‘Ring’, ‘Pinky’), making the ‘Peace’ gesture labeled as ‘Middle’ and ‘Index’ fingers raised. I trained a graph neural network (GNN) on a converted class-annotated gestures dataset [23] to match the fingers raised in the gesture. We increased the diversity of the training data, as most samples were directed toward the camera, making them nearly “perfect”. This was done by

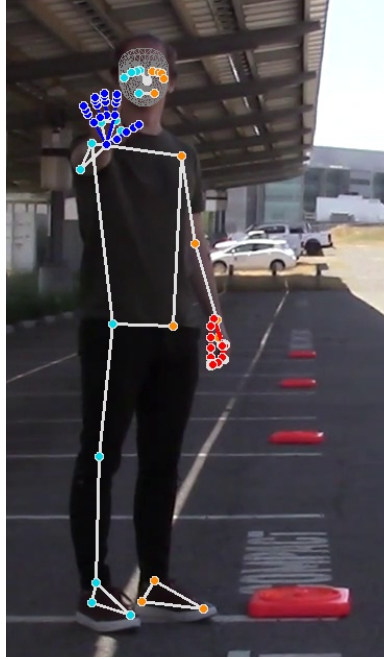


Figure 8.3: Sample video_137, ‘Decelerate’ visualized with the projection and the supplementary body description is: *“They’re facing the camera. Left hand is right and below their face with the palm facing left. Making a fist. Right hand is left and below their face with the palm facing the camera. Raised fingers are: Thumb, Index, Ring.”*

augmenting the data by applying geometric and photometric transformations: random rotations of π radians, isotropic scaling of 20%, and 1% Gaussian.

8.2 Enhancing Methods Validation

To ensure the methods’ functionalities, a ‘short’ validation was conducted by analyzing them in isolation. We validate the pose projection method in conjunction with the pose estimation functionality to accurately predict and project poses. The most reliable method to validate this would be to annotate each frame with the accurate pose, or, to be less precise, each subpart, such as face, hands, and general pose. However, it was thought that avoiding the annotation of each frame would be more efficient, as it would be too time-consuming and is not the priority of this study. To gain a general understanding of the functionality without annotating each frame, I found it helpful to count the number of frames for which each body part was estimated, as the participant remained on screen throughout the entire video. As mentioned earlier, this does not validate the correct projection but rather quickly estimates how many of the frames the pose estimation is functioning correctly. This will indicate whether the pose-estimation model is compatible enough for this purpose. With approximately 80% of frames estimated with the participant, the VLM could potentially have a chance to classify the gesture correctly. Still, with fewer pose-estimations, it will be less likely to prove the pose-estimation model as the flaw instead of the enhancement method itself. Since the SBD-method depends on pose estimation and needs frame-based annotation to perform more accurate validation, I do not investigate further validating the body description.

The validation shows that out of 9,150 frames across the 118 videos, the participant's body was detected in 86.28% (7,895 frames), detailed face in 34.96% (3,199 frames), left hand is detected 47.33% (4,331 frames), and right hand is detected 44.73% (4,093 frames). This indicates that only the body is projected in most frames, which should be sufficient for the projection method. However, since the SBD-method depends on the pose detection of each body part mentioned, it only makes a complete description of the body in less than 34% of the frames. This decreases the reliability of the final results of the SBD-method.

To fix this, I used the face landmarks of the face and hands to compute the relative position instead. I also tried this for the orientation of the face and hands, but the landmarks were too close to compute in 2D.

9 Results

The evaluation method used is inspired by the Classification evaluation method in VLM evaluation [4]. They ... To further analyze these methods' capabilities, I distinguish the classifications by the parameters' distance', 'lighting', and 'complexity'. Since the palm direction was unreliable, shown in Fig. 8.2, I evaluated both with and without the hand direction description.

	Accuracy ↑	Precision ↑	Recall ↑	F1-Score ↑
Plain	0.04	0.11	0.04	0.05
Projection	0.07	0.14	0.07	0.06
SBD	0.09	0.01	0.09	0.02
SBD + Hand	0.10	0.01	0.10	0.02
Proj. + SBD + Hand	0.08	0.23	0.08	0.03

Table 9.1: Classification results as weighted average from the extended version of ATG. The results across 118 videos show that all methods **decrease** the VLM's ability to classify static or dynamic navigation gestures. Even the static navigation gestures were not classified correctly. The weighted average F1-scores of the enhanced methods are ± 0.03 of the plain VLM's weighted average F1-score of 0.05.

Preliminary experiments revealed a significant skew towards predicting the classes 'Other' and 'Pointing'. They were removed not only due to this skew but mainly because none of the samples were labeled with these classes as their ground truths. They could be an 'easy' classification for the model, even if it's technically correct.

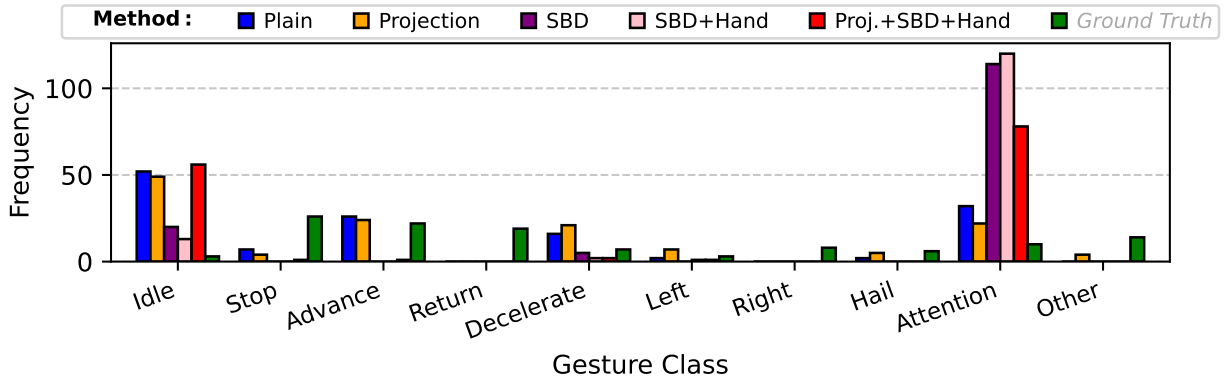


Figure 9.1: The predictions are distributed as follows, with each model including the ground truth. The predictions are heavily biased towards the 'Idle' and 'Attention' classes. The methods predict only these classes combined from 50% in 'Projection' to 99% in 'Describe', whereas the data contains 11% of these two classes combined.

We identified a pattern in the videos that was accurately predicted by all methods. None of the samples was classified correctly by all four methods. The term 'clear' refers to how visible and distinguishable the movement is. The samples classified correctly by three of the methods were video_26, video_28, and video_29, which are all 'Attention'. The movements are similar and clear. Across the samples, the participant is centered at 10 m, left at 10 m, and left at 2.5 m. However, video_27 has almost similar movements, if not slightly clearer and centered 5 m. video_98, 'Idle'

is the participant looking at their phone to the left and 7.5 m from the camera. The only video that two or more methods have correctly classified, which is not a ‘Idle’ or ‘Attention’ class sample, is video_20, ‘Slow’. This video is approximately 5 meters away, with a clear view of the arms.

It seems more or less as if the methods are guessing only with a bias towards specific classes. To reduce ambiguity in guesses, I reran the experiment to decrease random classification.

10 Discussion

Pedestrian-gesture classification remains an underexplored field due to the limited availability of data, although extensive research on human gestures makes integrating these models into pedestrian detection feasible.

10.1 Annotation Framework

The data processing, handling, and annotation process must be highly structured and have a clear set of guidelines. The data names should be consistent and rounded to zero for every process step. For example, videos and frames would need to be zero-indexed every time. Later renaming could potentially lead to mistakes and be tedious due to the large amount of data. The entire data structure is also essential, especially when others need to use it. The datasets were well-structured in their various forms, including individual videos and clusters of videos. The code was optimized to handle and understand the different structures. This took some time to implement, but it was worth it, as it made later adjustments easier. Seeing the annotation used by the undergrads made me change the whole structure. It was initially thought to work dynamically with raw, clean individuals and concatenated files, so the code would not have to be adjusted or re-implemented for each step of the data process. It ended up being tailored specifically for individual samples by locating the corresponding data file for each visualization. This made it time-consuming to use, since the annotator would have to write in the specific location of each file. However, this approach seemed more straightforward to understand than a specific data structure, where the system would search for a particular file without knowing its name. This can be ineffective when examining the data as a whole, but since the person can only view one sample at a time anyway, it should not be a significant issue. A potentially better structure of the data could be to have all videos in the same folder, labeled by both the video and camera names, rather than issuing individual folders for each cluster. It makes sense to work with them in the clusters, by concatenating and editing them, but then merging all clusters into one folder. The issue of having clusters as subfolders is, first of all, locating the specific file. This is not the biggest issue, but it can become a problem if the data structure is not maintained consistently. This happens if you were only to download a single sample from the data. Here, it also becomes an issue, as each ‘front’ camera is named accordingly in the cluster folder, as they are specified by video name in the cluster folder name. However, when downloading multiple front files, each one must be renamed manually. Keeping all files in a single folder would already be named by video and camera.

10.2 Evaluation Paper

Extended discussion of the paper: Overall, it demonstrates that even expert-generated captions struggle to achieve a high cosine similarity score. This suggests that there should have been more par-

ticipants to provide ground-truth annotated captions for comparison, to generalize the results more effectively, as was attempted in the second comparison, or to include more expert-generated captions to filter out noise. It does not seem like it would make a difference in the results, but the prompt for the VLMs and the expert captions should have been the same or more similar.

Reconstruction could have used another metric, as the temporal difference would significantly impact the result. Although the movement could be precisely accurate, it might not be reenacted in the same frame as the original video. Reconstructing the caption from multiple sources should have been done in a different order to avoid previous knowledge. This would require more data to show a significant difference in the generation methods.

The difficulty of comparing sentences as a source of singular ‘truth’ is well-known, making the optimal annotation of ground truth captions difficult to construct. With this in mind, I evaluate VLM output in two ways: the comparison of the similarity of the model output to ground truth captions, and the measurable utility of the output toward the success of the driving task.

10.3 Enhancing

It was hoped that the methods demonstrated capabilities in the samples that were at least 1-5 m away from the camera. The results and post-analysis indicated that the predicted classifications are largely random, with a predominance of ‘Idle’ or ‘Attention’. The methods do increase the capabilities of the VLM, but not very significantly. The sample `video_20` could be predicted more, since it is closer to the camera, but since it is only one close sample, it might as well be random.

The findings indicate there is no foundation for VLMs to caption gestures with or without enhancement methods, especially given that the old, flawed SBD method is more accurate than the corrected SBD.

A significant aspect of the VLM that will help it succeed in this task is to perceive and utilize natural language. The VLM is potentially trained only on single-word or a few-word captions. This is unsuitable for selecting a classification in the prompt or describing the entire human body in both static poses and motion.

The paper generates longer paragraphs of text in the examples. Figure 9 shows a basketball court with the prompt *“What part of a basketball game is this?”*. The caption explains how a basketball player is holding a ball relative to themselves *“He’s holding the ball and preparing to take a shot, with his arms raised above his head in the classic free throw stance.”* [51]. This makes a case where the VLM is actually performing the task I study, even without being prompted.

It is crucial for any model during this task to fully perceive the participant’s body in relation to both the participant and the ego driver. As the participants moved in many directions from the perspective of the ego driver, it became complicated to recognize movement and position. This is especially true for gestures directed towards the ego driver, which can be even more challenging to capture using a mono-camera. An improvement to this could be utilizing 3D pose estimation of the subject. This would not only potentially be more accurate poses, but it would also ease perceiving

gestures towards the ego driver.

Since I mainly use the body pose in the final version, I could use Ultralytics YOLOv11-Pose [22], or increase the accuracy with OpenPose [8]. However, these poor results do not indicate that it would make a huge or any difference.

10.4 Research Questions

The research questions are addressed and discussed. RQ1 has been discussed in the enhancement section. RQ2 and RQ3 are both difficult to answer, as they involve decision-making, which was halted. After the shift in the project, a new problem statement and set of research questions would be optimal to construct to answer, rather than those that focus on decision-making. RQ2 begs the question of inferring a model to decide these scenes of conflict. It cannot be concluded, but a lot of research in this study suggests the need for a more complex model than an attention model that includes each pedestrian. This is not only to avoid training the model or doing it in a zero-setting, but also because so many parameters go into this decision. It does not depend only on the authority, but also on the pedestrians, and everything about it. It can vary from driver to driver. An example of this is the recording of the Act-CANG dataset. It was thought that the driver would obey the civilians' command, rather than the police officer's, in the scene. This does not align with the law, but the thought process was that the civilians had additional information or empathy that the officer did not possess. However, the driver still followed the police officer, claiming the civilian did not seem severe enough since they were smiling. This made the civilian untrustworthy.

To evaluate the ego drivers' decision-making, does not come down to the individual priority, as initially thought. This comes in the nature that the more complex decision does not lie only with the authority, but can vary depending on a countless number of variables. To better plan the vehicle's action and evaluate it, it is thought to utilize trajectories instead of finite classes. This is, however, first suggested when there is an improvement in the decision-making.

10.5 Limitations and Challenges

In some cases, it was manually observed that the VLM would prefer to 'Stop' or 'Decelerate' the car, where it was supposed to maintain a constant speed, by arguing that driving in a narrow street would be safer through 'Decelerate'. However, by stating it was only driving 5 km/h and even 0 km/h in the prompt, it still preferred this as the safest course of action. This reminds us that the loss should minimize travel duration, as the safest option would probably be not to drive at all.

11 Conclusion

The overall problem was bigger than initially thought. It was initially thought to be easy to find datasets of pedestrians gesturing towards the driver, since it seems pretty familiar when driving as a private person. Gestures in conflict, however, are a more uncommon scenario and an edge case. As the report shows, that was not even the most significant issue. The first task, recognizing pedestrians' gestures, was the biggest issue in this project. Taking over the focus, saving the decision-making in conflict for another time.

In this study, the primary focus was to detect pedestrian-to-driver navigation gestures in a zero-shot setting using VLMs. It was first proven in the *VLM Evaluation* paper [4] that they were not capable of this task. The hypothesis lies in improving the ability by enhancing the VLMs using the methods *Supplementary Body Description with VLM* and *Pose Projection*, to avoid fine-tuning. None of these methods showed any increase in accuracy in classifying the extended dataset videos of ATG. Across 118 samples, each method only had an F1-score between 0.02 and 0.06. This again proves the incapability of the VLM VideoLLaMA3, as well as the insufficient increase in enhancement methods. This calls for further research in this domain.

Regarding conflicting scenarios in traffic, further research is needed to conclude this matter. This study, however, highlights the complexity of this problem and the scarcity of available data. This study provides a foundation for understanding this problem through data definition and a dataset.

The problem statement and research questions can not be concluded upon. This study gives an idea for further research to solve this problem. With the constructed datasets and defined problem, this study is just a small piece of a bigger domain.

12 Future Work

CANG Dataset

In this annotation, we skipped gestures that are not directed to the ego driver. To complete the annotation, all gestures should be annotated. I suggest using a binary flag `ego_mask` to indicate gestures directed towards the ego driver. Additionally, this enables the possibility of expanding the annotation to include gestures directed at subjects other than the ego driver. In a more complex scene understanding, the direction could be specified by the individual subject's ID or location instead of using a binary direction feature. This may be more accessible using 3D scenes and object permanence or continuity.

Additional Enhancing Methods

This study highlights the limitations of enhancing VLMs, suggesting that other methods may likely follow a similar path. However, another idea for a method, that was not included in the evaluation, was *Chain-of-Thought*. This method involves the concept of recurring the VLM to enforce itself. The idea lies in having it ask itself questions about what more it should be looking for about the pedestrian.

Zero-Shot Navigation Gesture Recognition

Train a foundation model encoder that embeds descriptive body movement, which an LLM would then interpret. This idea came from reading about MotionBERT [53].

End-to-end

Instead of classifying or captioning the gesture, have the image and pose as input, and have the trajectory as the output. The pose can be complex to classify and describe accurately in relation to another model to understand it fully. With sufficient navigation gestures, the model should be able to be trained end-to-end. This avoids middle steps, where the meaning can be lost in translation between models.

Bibliography

- [1] Ali K AlShami, Ananya Kalita, Ryan Rabinowitz, Khang Lam, Rishabh Bezbarua, Terrance Boulton, and Jugal Kalita. “Cooool: Challenge of out-of-label a novel benchmark for autonomous driving”. In: *arXiv preprint arXiv:2412.05462* (2024).
- [2] Hidehisa Arai, Keita Miwa, Kento Sasaki, Yu Yamaguchi, Kohei Watanabe, Shunsuke Aoki, and Issei Yamamoto. “Covla: Comprehensive vision-language-action dataset for autonomous driving”. In: *arXiv preprint arXiv:2408.10845* (2024).
- [3] Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, et al. “An introduction to vision-language modeling”. In: *arXiv preprint arXiv:2405.17247* (2024).
- [4] Tonko E. W. Bossen, Andreas Mgelmoose, and Ross Greer. *Can Vision-Language Models Understand and Interpret Dynamic Gestures from Pedestrians? Pilot Datasets and Exploration Towards Instructive Nonverbal Commands for Cooperative Autonomous Vehicles*. 2025. arXiv: 2504.10873 [cs.CV]. URL: <https://arxiv.org/abs/2504.10873>.
- [5] Andrew Brown, Vicky Kalogeiton, and Andrew Zisserman. “Face, body, voice: Video person-clustering with multiple modalities”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 3184–3194.
- [6] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. “nusenes: A multimodal dataset for autonomous driving”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 11621–11631.
- [7] Carlos Campos, Richard Elvira, Juan J Gómez Rodríguez, José MM Montiel, and Juan D Tardós. “Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam”. In: *IEEE transactions on robotics* 37.6 (2021), pp. 1874–1890.
- [8] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. “Realtime multi-person 2d pose estimation using part affinity fields”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 7291–7299.
- [9] Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. *LongVILA: Scaling Long-Context Visual Language Models for Long Videos*. 2024. arXiv: 2408.10188 [cs.CV].
- [10] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. “Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms”. In: *arXiv preprint arXiv:2406.07476* (2024).

- [11] MMPose Contributors. *OpenMMLab Pose Estimation Toolbox and Benchmark*. <https://github.com/open-mmlab/mmpose>. 2020.
- [12] Tri Dao. “Flashattention-2: Faster attention with better parallelism and work partitioning”. In: *arXiv preprint arXiv:2307.08691* (2023).
- [13] Jing Ding, Shanwei Niu, Zhigang Nie, and Wenyu Zhu. “Research on human posture estimation algorithm based on YOLO-pose”. In: *Sensors* 24.10 (2024), p. 3036.
- [14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).
- [15] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. “The pascal visual object classes (voc) challenge”. In: *International journal of computer vision* 88 (2010), pp. 303–338.
- [16] GitHub, OpenAI, Microsoft. *GitHub Copilot*. <https://github.com/features/copilot>. 2021.
- [17] Google Research. *MediaPipe Pose*. <https://github.com/google/mediapipe>. 2023.
- [18] Grammarly Inc. *Grammarly*. <https://www.grammarly.com/>. 2023.
- [19] 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*. 2016, pp. 770–778.
- [20] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. *RTMPose: Real-Time Multi-Person Pose Estimation based on MMPose*. 2023. DOI: [10.48550/ARXIV.2303.07399](https://doi.org/10.48550/ARXIV.2303.07399). URL: <https://arxiv.org/abs/2303.07399>.
- [21] Tao Jiang, Xinchun Xie, and Yining Li. “RTMW: Real-Time Multi-Person 2D and 3D Whole-body Pose Estimation”. In: *arXiv preprint arXiv:2407.08634* (2024).
- [22] Glenn Jocher and Jing Qiu. *Ultralytics YOLO11*. Version 11.0.0. 2024. URL: <https://github.com/ultralytics/ultralytics>.
- [23] Alexander Kapitanov, Karina Kvanchiani, Alexander Nagaev, Roman Kraynov, and Andrei Makhliarchuk. *HaGRID – HAnd Gesture Recognition Image Dataset*. Kaggle Dataset, <https://www.kaggle.com/datasets/kapitanov/hagrid>. 2022.
- [24] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models”. In: *International conference on machine learning*. PMLR. 2023, pp. 19730–19742.
- [25] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. “Video-LLaVA: Learning United Visual Representation by Alignment Before Projection”. In: *arXiv preprint arXiv:2311.10122* (2023).

- [26] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. *VILA: On Pre-training for Visual Language Models*. 2023. arXiv: [2312.07533](https://arxiv.org/abs/2312.07533) [cs.CV].
- [27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. “Microsoft coco: Common objects in context”. In: *Computer vision—ECCV 2014: 13th European conference, zurich, Switzerland, September 6–12, 2014, proceedings, part v 13*. Springer. 2014, pp. 740–755.
- [28] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Vishwesh Nath, Jinyi Hu, Sifei Liu, Ranjay Krishna, Daguang Xu, Xiaolong Wang, Pavlo Molchanov, Jan Kautz, Hongxu Yin, Song Han, and Yao Lu. *NVILA: Efficient Frontier Visual Language Models*. 2024. arXiv: [2412.04468](https://arxiv.org/abs/2412.04468) [cs.CV]. URL: <https://arxiv.org/abs/2412.04468>.
- [29] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. “Mediapipe: A framework for building perception pipelines”. In: *arXiv preprint arXiv:1906.08172* (2019).
- [30] Ana-Maria Marcu, Long Chen, Jan Hünemann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, et al. “LingoQA: Visual question answering for autonomous driving”. In: *European Conference on Computer Vision*. Springer. 2024, pp. 252–269.
- [31] Kaouther Messaoud, Itheri Yahiaoui, Anne Verroust-Blondet, and Fawzi Nashashibi. “Non-local social pooling for vehicle trajectory prediction”. In: *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE. 2019, pp. 975–980.
- [32] Ashutosh Mishra, Jinhyuk Kim, Jaekwang Cha, Dohyun Kim, and Shiho Kim. “Authorized Traffic Controller Hand Gesture Recognition for Situation-Aware Autonomous Driving”. In: *Sensors* 21.23 (2021). ISSN: 1424-8220. DOI: [10.3390/s21237914](https://doi.org/10.3390/s21237914). URL: <https://www.mdpi.com/1424-8220/21/23/7914>.
- [33] OpenAI. *ChatGPT*. <https://chatgpt.com/>. 2023.
- [34] Oxford University Press. *self-driving*. 2025. URL: <https://www.oxfordlearnersdictionaries.com/definition/english/self-driving>.
- [35] Wing Yi Pao, Long Li, and Martin Agelin-Chaab. “Perceived rain dynamics on hydrophilic/hydrophobic lens surfaces and their influences on vehicle camera performance”. In: *Transactions of the Canadian Society for Mechanical Engineering* 48.4 (2024), pp. 543–553.
- [36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. “Learning transferable visual models from natural language supervision”. In: *International conference on machine learning*. PmLR. 2021, pp. 8748–8763.

- [37] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. “Improving language understanding by generative pre-training”. In: (2018).
- [38] Vasili Ramanishka, Yi-Ting Chen, Teruhisa Misu, and Kate Saenko. “Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7699–7707.
- [39] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. “You only look once: Unified, real-time object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.
- [40] Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, et al. “Code llama: Open foundation models for code”. In: *arXiv preprint arXiv:2308.12950* (2023).
- [41] Fiona Ryan, Ajay Bati, Sangmin Lee, Daniel Bolya, Judy Hoffman, and James M. Rehg. “Gaze-LLE: Gaze Target Estimation via Large-Scale Learned Encoders”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2025.
- [42] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. “Deep High-Resolution Representation Learning for Human Pose Estimation”. In: *CVPR*. 2019.
- [43] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. “Llama 2: Open foundation and fine-tuned chat models”. In: *arXiv preprint arXiv:2307.09288* (2023).
- [44] u-blox. *Autonomous driving levels: from unassisted to hands-free driving*. 2024. URL: <https://www.u-blox.com/en/blogs/insights/autonomous-driving-different-levels>.
- [45] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. “Deep High-Resolution Representation Learning for Visual Recognition”. In: *TPAMI* ().
- [46] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. “Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution”. In: *arXiv preprint arXiv:2409.12191* (2024).
- [47] Julian Wiederer, Arij Bouazizi, Ulrich Kressel, and Vasileios Belagiannis. “Traffic control gesture recognition for autonomous vehicles”. In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2020, pp. 10676–10683.
- [48] Dong Wu, Man-Wen Liao, Wei-Tian Zhang, Xing-Gang Wang, Xiang Bai, Wen-Qing Cheng, and Wen-Yu Liu. “Yolop: You only look once for panoptic driving perception”. In: *Machine Intelligence Research* (2022), pp. 1–13.
- [49] Ruoxin Xiong and Pingbo Tang. “Pose guided anchoring for detecting proper use of personal protective equipment”. In: *Automation in Construction* 130 (2021), p. 103828. DOI: <https://doi.org/10.1016/j.autcon.2021.103828>.

- [50] Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, Jun Zhang, Andreas Geiger, Yu Qiao, and Hongyang Li. “Generalized Predictive Model for Autonomous Driving”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.
- [51] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. “VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding”. In: *arXiv preprint arXiv:2501.13106* (2025).
- [52] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiayi Cui, Hongfa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. “LanguageBind: Extending Video-Language Pretraining to N-modality by Language-based Semantic Alignment”. In: *arXiv preprint arXiv:2310.01852* (2023).
- [53] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. “Motionbert: A unified perspective on learning human motion representations”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 15085–15099.
- [54] Bin Xiao, Haoxiang Zhang, Jingdong Wang, Zengke Geng, Ke Sun. “Bottom-Up Human Pose Estimation Via Disentangled Keypoint Regression”. In: *CVPR*. 2021.

A VLM Evaluation Paper [4]

Can Vision-Language Models Understand and Interpret Dynamic Gestures from Pedestrians? Pilot Datasets and Exploration Towards Instructive Nonverbal Commands for Cooperative Autonomous Vehicles

Tonko Bossen
University of California Merced
tbosse20@student.aau.dk

Andreas Møgelmoose
Aalborg University
anmo@create.aau.dk

Ross Greer
University of California Merced
rossgreer@ucmerced.edu

Abstract

In autonomous driving, it is crucial to correctly interpret traffic gestures (TGs), such as those of an authority figure providing orders or instructions, or a pedestrian signaling the driver, to ensure a safe and pleasant traffic environment for all road users. This study investigates the capabilities of state-of-the-art vision-language models (VLMs) in zero-shot interpretation, focusing on their ability to caption and classify human gestures in traffic contexts. We create and publicly share two custom datasets with varying formal and informal TGs, such as ‘Stop’, ‘Reverse’, ‘Hail’, etc. The datasets are “Acted TG (ATG)” and “Instructive TG In-The-Wild (ITGI)”. They are annotated with natural language, describing the pedestrian’s body position and gesture. We evaluate models using three methods utilizing expert-generated captions as baseline and control: (1) caption similarity, (2) gesture classification, and (3) pose sequence reconstruction similarity. Results show that current VLMs struggle with gesture understanding: sentence similarity averages below 0.59, and classification F1 scores reach only 0.14–0.39, well below the expert baseline of 0.70. While pose reconstruction shows potential, it requires more data and refined metrics to be reliable. Our findings reveal that although some SOTA VLMs can interpret zero-shot human traffic gestures, none are accurate and robust enough to be trustworthy, emphasizing the need for further research in this domain. We make our code publicly available at github.com/tbosse20/gest_VLM_eval

1. Introduction

Scene understanding and decision-making in autonomous driving rely on the ability of systems to predict the future location of moving objects [1–3]. Still, a limitation of safe autonomy lies in understanding the gestures of surrounding humans. This decreases the safety and trust of these systems in interactive traffic scenarios.



Figure 1. In autonomous driving scenarios, navigation instructions may come from pedestrians’ dynamic, nonverbal gestures. Interpreting and responding to such gestures is vital for safe autonomous driving.

While physical constraints of motion may inform trajectory prediction methods for dynamic objects, in this research, we approach the challenge of intent prediction [4], where motion must be anticipated before it begins. However, our research considers not only the intention of an individual agent but also the intentions the agent imposes on others in the form of instructions [5]. We make a clarifying distinction in uses of the word ‘intention’: borrowing from attention-based learning architecture terminology, an agent’s self-intention describes their intended future actions, while an agent’s cross-intention describes the future actions of other agents as intended by the observed agent.

An agent’s intentional gestures may communicate intent, querying, and instruction, and often require comprehensive scene understanding to ensure a safe response, especially when other independent agents are present in the scene, making the complete scene motion less predictable. In ideal settings, a scene may have a formalized focal point, such as a law enforcement or traffic-directing officer, such as in Fig. 1, whose authority reduces the complexity of interpreting the scene and selecting gestures to follow. Understanding

traffic gestures is crucial for autonomous vehicles (AVs) to function in these *formally-directed* settings, where the ability to follow instructions is essential. In non-authoritative or *informal* situations, though there is no direct jurisdiction of control. Pedestrians can signal their own intent, expectancy of the ego drivers’ intent, or information about the scene. For example, acting as a vantage point for occluded areas. While explicit communication from pedestrians occurs in only 2.7% of road-crossing events [6], it can enhance the experience and safety of pedestrians and drivers by increasing the scene understanding. In this research, we consider that the pedestrian is not just a passive agent in the traffic scene, but a source of deliberate, intentional information and instructions to the ego vehicle, whether or not they are ordained with authority.

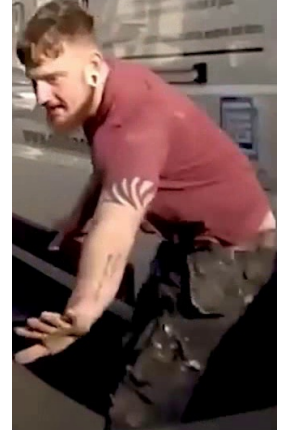
Human gestures can be subtle, yet through common sense and experience, people can often understand one another using only gestures. However, the interpretation of gestures can vary between countries and cultures, adding to the complexity of intention understanding [7]. The wide range of subtle variations in gestures may be difficult to capture in any limited-size training dataset, and further may be difficult to extrapolate in meaningful ways through pose estimation subtask modules towards gesture understanding. So, in this study, we seek to understand if vision-language models (VLMs) trained on foundation-model-scale data may encode this information to be recovered in a zero-shot manner. Teaching AV systems to interpret pedestrians’ incoming gestures can create a more maneuverable, efficient, and safer road environment. This research explores how AI can perceive and utilize pedestrian gestures to make better decisions.

We perform an initial evaluation using a few online demos to gain insight into this task. Figures 2a and 2b show two pedestrians extracted from left-to-right from a real traffic scene of the COOOL dataset [8]. Fig. 2b provides a preliminary example that hints toward the general incapability of VLMs to caption traffic gestures. The VLM received the image along with the prompt: “*What is this pedestrian gesturing?*” using available online VLM demos BLIP2¹, VideoLLaMA2², VideoLLaMA3³, VideoLLaMA3-Image⁴, and ChatGPT-4o⁵.

The output prompting these images varied from “*The person in the video is gesturing something that resembles a farting sound or action.*” from VideoLLaMA2, to “*...their body posture suggests they might be trying to stop something or someone, maintain balance ...*” from ChatGPT-4o. However, even the slightly more promising ChatGPT model



(a) Girl pedestrian walking in front of the ego car, looking down the road, with her arms down her side (crop).



(b) Man pedestrian gesture ego driver to ‘Stop’ with one arm (crop).

Figure 2. Pedestrians from frame 71 from video 0153 of the COOOL dataset [8]. Zero-shot analysis of these pedestrians using VLMs fails to capture the significance of their gestures (or non-gestures) toward the traffic scene.

was quick to hallucinate or misinterpret gestures when presented with the same prompt accompanying Fig. 2a. This image was a crop of a little girl walking in front of the ego driver, facing away from the ego driver with her arms down her sides. Still, ChatGPT’s output stated she was raising one arm and waving or pointing. This output expressed further insecurity due to image quality despite the visual clarity. These brief examples highlight the impetus for a more thorough research into the limitations of these models in recognizing gesture-based human communication.

2. Related Work

Existing VLMs can generate various output forms supporting autonomous driving, ranging from natural language captions that might inform trajectory generation [9, 10], selection of specific control commands [11], novelty detection [12, 13], or even end-to-end learning of direct way-point trajectories [14, 15]. A particular class of models that are most relevant for gesture recognition are Video Foundation Models (ViFMs) due to the temporal nature of gesture-based communication [16]. In this section, we elaborate on some existing VLM techniques and datasets toward the safe navigation task.

CoVLA combines VLMs and object detection to generate a caption of the scene, which is used to learn a trajectory projected upon the scene image [17]. The usage of VLMs is considered for the end-to-end training of networks, which may naturally include gesturing pedestrians. Still, we suggest that scene agent gestures can be so fundamental to control decisions that implicit end-to-end learning

¹huggingface.co/spaces/hysts/BLIP2

²huggingface.co/spaces/lixin4ever/VideoLLaMA2

³huggingface.co/spaces/lixin4ever/VideoLLaMA3

⁴huggingface.co/spaces/lixin4ever/VideoLLaMA3-Image

⁵chatgpt.com

may not be sufficient. Further, the existence of confounding factors (e.g., a green light with a traffic controller simultaneously indicating forward motion) may inhibit the model from learning causality between scene object patterns and control decisions [18–20]. Further, models like CoVLA and DriveLLaVA [21] learn precise trajectory outputs rather than abstract direction commands; the alignment of the latter may be more suited to learning from gesture, though it is our intention in developing methods for such gesture feature extraction that this may also be used modularly within trajectory learning pipelines.

While existing RGB-media datasets around pose and gesture focus on ‘action’, ‘pose estimation’, ‘sign language’, and ‘hand gesture’ [22–25], there is a gap in data within the navigation/traffic gestures domain, which predominantly utilize the upper-body. Non-domain datasets still provide utility for general pose classification using VLMs, expanding the general understanding of the capabilities of these models. Within the domain, the Traffic Control Gesture (TCG) dataset includes 250 sequences of 3D body skeleton [26]. While this dataset is highly pertinent to our study, it solely provides 3D pose annotations without the accompanying visual data, limiting its applicability in VLMs. While the anomaly detection dataset COOOL [8] contains a few scenes of pedestrian gesture-based communication to the ego driver, the scenes are unannotated and are too few in number for a gesture-specific analysis.

A promising advancement in zero-shot recognition of hand gestures using image data is GestLLM [27], which integrates large-language models with pose-based feature extraction. GestLLM system showed robust performance in hand gesture recognition, providing one path towards improving the zero-shot VLM issues studied in this paper.

3. Methodology

This study seeks to understand the zero-shot capabilities of VLMs in recognizing and responding to static, dynamic, and composite human traffic gestures in RGB videos with physical body descriptions and contextual interpretations within driving scenarios. To do this, we utilize three evaluation methods 1. *Embedded Similarity*, 2. *Classification*, and 3. *Reconstruction* to quantify the performance of models in converting human gestures and motion to text and derivative meaning.

By using VLMs, *intentional* gestures (e.g., hand gestures) are emphasized with *accompanying* (e.g., body language and facial cues) second-hand information, which may or may not agree with the manual (i.e., hand-communicated) intention.

3.1. Dataset

We create and publish two datasets for this study⁶: 1. ‘Acted Traffic Gestures’ (ATG) with a single actor portraying various gestures to the camera as a hypothetical ego vehicle, used for this evaluation. 2. ‘Instructive Traffic Gestures In-the-Wild’ (ITGI) is a real-world encounter of an ego vehicle with traffic conductors, filmed from four synchronized dash cameras for a multi-directional surround view, added as additional data.

Acted Traffic Gestures (ATG) The ATG dataset features a single actor gesturing towards a static camera recorded at 30 FPS. The camera is 1.6 meters above the ground and 1 - 2 meters from the participant, acting as a vehicle dash-cam. It is recorded inside a closed room against a white wall, to lock parameters and reduce noise. The dataset includes 8 short videos with gestures for ‘Idle’, ‘Reverse’, ‘Go’, ‘Stop, pass’, ‘Follow’, ‘Forward’, ‘Stop, go’, and ‘Hail’, ranging from 1 to 4 seconds. The ‘Stop, pass’ and ‘Stop, go’ are composed gestures of ‘Stop’ to the ego driver, and ‘Pass’ or ‘Drive’ to other vehicles. These gestures are acted out as an unofficial but naturalistic traffic guide rather than following any municipality’s official traffic warden gestures. *In this report, we detail the initial dataset properties at the time of writing, and the dataset continues to be extended with varying environments, distances, and gestures to enable research across a wider range of scenes and parameters. Updated dataset details are available in the repository README.*

Ground truth annotations were made by a licensed driver with oracle knowledge of the underlying gestures instructed to the actor. This annotator reviewed each video at 8-frame intervals, without overlapping segments. The traffic gesture label is described from both the pedestrian’s and the driver’s perspectives combined in each annotation, and it is interpreted in terms of the pedestrian’s intended communication towards the ego driver or other drivers. Additionally, *expert-generated* captions were made by additional licensed drivers. They serve both as a ‘baseline’ to assess the overall effectiveness and accuracy of the evaluation method, and as ‘supplementary’ ground truths to emphasize the intended meaning of the gestures, rather than the specific wording. The ‘instructions’ considered when generating the ground truth is formulated as follows: *“Describe the pedestrians’ body posture focusing on their arm position and movement relative to both themselves (e.g., at their side, in front of them) and the ego driver (e.g., towards the ego driver, left of the ego driver), their hand position and shape (e.g., flat hand faced downward), and the orientation of their body and face (e.g., facing to the left). Include an interpretation of potential gestures and their intended recipient (e.g., signaling to stop, requesting to pull over).”* A ground truth cap-

⁶Link to datasets in README: github.com/tbosse20/gest_VLM_eval



Figure 3. Frame 18 from the “Reverse” command gesture video. This frame’s set was annotated with the caption, *“The pedestrian is standing in front of the ego driver. They are facing their torso and head towards the ego driver. They are moving their flat palms back and forth towards the ego driver, gesturing for it to reverse. They are moving slightly to the side.”*

tion example and a corresponding frame from an 8-frame sequence are provided in Fig. 3. In addition to the short-term motion analysis annotations, we provide a complete caption describing the gesture for each complete video.

Instructive Traffic Gestures In-the-Wild (ITGI) The dataset is collected while driving around town during a bike race, which involved police enforcement guiding in intersections. It varies from minor intersections with a single idle police officer and a few cones, to light intersections blocked by police vehicles with multiple officers guiding cars. The scenes are all set in formal settings, including official traffic regulation gestures and more casual/unofficial gestures. It was recorded from a Tesla with four built-in cameras: front, back, right-back, and left-back at 36 FPS. The data consists of 18 videos ranging from 8 seconds to 2 minutes. We provide this dataset without annotation.

3.2. Models and Setting

The selected models evaluated in this study are *VideoLLaMA2* [28], *VideoLLaMA3* [29], and *Qwen2* [30]. The VideoLLaMA models were selected for their recent notable performance in the CoVLA paper [17], and Qwen’s near relationship as a subset of VideoLLaMA. We limited our evaluation over models which are available to run locally on an edge device without internet access or API charge, which excludes candidate model ChatGPT-4o. All evaluations were run on one NVIDIA® GeForce RTX™ 4090 24GB compatible with `float16` precision and Flash-Attention 2.0 [31]. We use a model temperature of 0.2 and 512 maximum new tokens.

We use 8-frame samples per caption. This window span was selected as a baseline for experiments to engage frame-to-frame captioning with temporal context. The frame rate was chosen as an estimate of the necessary temporal information to reasonably represent a gesture, an area we high-

light for future research. The correct frame rate is a study in itself, as traffic gestures in this dataset can vary from 0.1 to 5 seconds long, and multiple gestures can be combined into one command interpretation, making it challenging to capture a complete gesture and its communicative intent within a short window of eight frames.

3.3. Prompting

For the purpose of this research, we craft a series of prompts intended to extract varying types of information from the VLM. We acknowledge that partial information may enable chain-of-thought reasoning as opposed to zero-shot task success, and that analysis of subtasks may be valuable toward ongoing research. Two central pieces of information included in the prompts are the ‘Context’ (a driving autonomous vehicle from a dash-cam perspective), and the ‘Objective’ (retrieving information for safe, intention-aligned decision-making). Additionally, to the point of subtask analysis, our prompts seek to extract either an ‘Explanation’ of the agent’s physical motion involved in the gesture, for an LLM to interpret downstream, or a direct interpretation of the gesture from the VLM itself. This is useful in cases when the VLM does not grasp the concept of the meaning of the gestures, but could explain the agent’s movements well enough for an LLM to interpret.

To give the models a broader chance to successfully caption the gesture correctly, varying text prompts were used to evaluate each model. The five varying prompts used are referred to as ‘Blank’ (*The prompt is left empty; the images alone serve as prompt*), ‘Determine’ (*“Determine what gesture the pedestrian is making.”*), ‘Body’ (*“Provide a detailed explanation of the pedestrian’s body posture and movements.”*), ‘Context’ (*“You are an autonomous vehicle navigating a road. Determine what gesture the pedestrian is making.”*), and ‘Objective’ (*“You are an autonomous vehicle navigating a road. Determine what gesture the pedestrian is making. Your response will be used by an AI system to make real-time driving decisions.”*), which vary in understanding and focus. While ideally the accurate interpretation of motion should be enough to inform the meaning of the gesture, in some prompts we suggest that the VLM consider the context of the 3D driving scene.

4. Evaluation

The evaluation section consists of the three evaluation methods: ‘Embedded Similarity’, ‘Classification’, and ‘Reconstruction’. We approach this topic from multiple angles, in different degrees of abstractions and focus points. The evaluations are conducted using the ATG dataset. Each section provides a description of the method and the results. Discussions are combined in Section 5.

4.1. Embedded Similarity

This evaluation method seeks to understand the similarity between the generated caption and the ground truth caption. The generated captions’ similarity was evaluated by embedding them using the SBERT encoder (all-MiniLM-L6-v2) [32] and measuring their similarity to the ground truth embeddings using *Cosine Similarity* [33]. The cosine similarity scores of the expert-generated captions serve as a baseline for assessing the overall effectiveness and accuracy of the evaluation method. To ensure that the results are not biased by a particular gesture or prompt, the outcomes are analyzed separately by prompt type in Fig. 6 and by gesture in Fig. 7.

To argue for the selected metric, we validated multiple metrics on decreasing similarity rephrasing of a target caption. We expect to see a trend of metric values decreasing as caption quality decreases (relative to the original target caption). We illustrate the results of this experiment in Fig. 4, with ‘Ideal’ as the optimal metric trend, which is hardcoded. The sampled metrics are BERT score [34] (not to confuse with the SBERT encoder), BLEU [35], Cosine Similarity [33], Semantic Textual Similarity (STS) [32], Jaccard [36], METEOR [37], and ROUGE [38].

SBERT was selected as the implemented encoder, due to its 74% difference in cosine similarity between ‘Equivalent’ and ‘Unrelated’ validation captions, in contrast to Vanilla BERT [39], showing only an 11% difference. This makes it easier to distinguish. Other evaluation metrics remained consistent regardless of the encoder used.

We do not expect a similarity score of 1.0, as the wording has to be exact for this to happen. That is neither likely nor intended. The similarity score is considered highly accurate at around 0.80, the score of the ‘Equivalent’ validation captions, by looking at the validation at Fig 4. Scores near 0.75 are considered moderately accurate as ‘Extended’ or ‘Partial’. Scores near and below 0.55 are considered low accuracy as ‘Slight’ similarity.

4.2. Classification

Arguing that the semantic sentence comparison is complex to generate and evaluate, we reduce the model task to predicting a precise answer with complete interpretation. The possible classes are the same as the videos provided, plus additional common traffic gestures, making a total of 9 classes: ‘Follow’, ‘Hail’, ‘Forward’, ‘Right’, ‘Left’, ‘Idle’, ‘Reverse’, ‘Stop’, and ‘Other’. The provided videos include more complex gestures like ‘Stop, pass’, but to simplify it, this evaluation method only focuses on the gesture towards the ego driver. The prompt was formulated with the context, steps to identify the matter, output format, and possible classes, each with a short description.

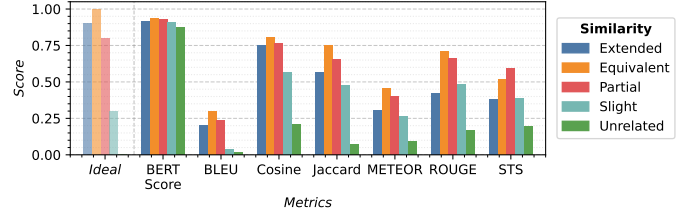


Figure 4. We validate the selected encoder and metric by testing their ability to illustrate the expected similarity trend. The ‘Ideal’ ratio trend is illustrated on the left, used to validate the most suitable metric. This involves comparing a target sentence to a series of rephrased versions with progressively decreasing similarity levels. Two target captions were formulated about the same hypothetical scenario (e.g., “A person signals the ego driver to stop, by putting their hand towards the ego driver.”). Each similarity level contains two rephrases cross-validated against both target captions to reduce sentence noise. The rephrases span five levels of similarity: ‘Extended’ with additional information, which can be difficult to know it is irrelevant (noise) or incorrect (false positive) information (e.g., “A pedestrian raises their hand towards the ego driver to stop traffic. They are looking scared and in need of help.”), ‘Equivalent’ with the same information (e.g., “A pedestrian raises their hand towards the ego driver to stop traffic.”), ‘Partial’ with partially equivalent information (e.g., “A person raises their hand towards the ego driver.”), ‘Slight’ which is missing important details (e.g., “A human gestures to the ego driver.”), and ‘Unrelated’ information (e.g., “The sky is blue and the sun is shining.”).

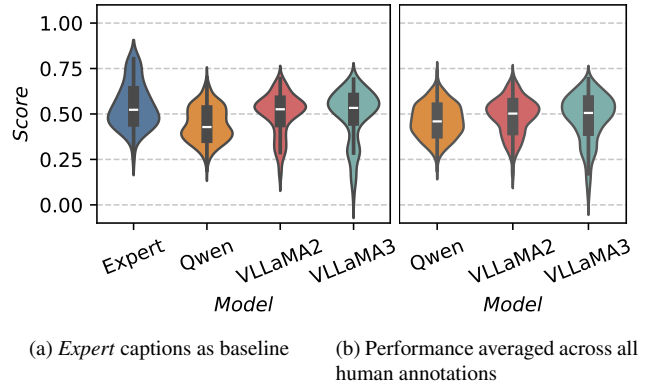


Figure 5. Cosine similarity with the generated captions relative to only the ground truth captions (5a, left) and both the ground truth and expert captions (5b, right). They are visualized together to illustrate the decreasing standard deviation due to the denoising. In 5a VideoLLaMA2 and VideoLLaMA3 show a slightly higher Q1 than the expert-generated captions. However, experts’ mean is 0.54, while the mean of VideoLLaMA2 is 0.50, and VideoLLaMA3 is 0.49. The mean of Qwen is 0.44. In 5b Qwen is increased to 0.46, but VideoLLaMA2 and VideoLLaMA3 are decreased to 0.48 and 0.47. This again shows a diversity in the non-VLM captions, and there are multiple ways to describe a certain gesture.

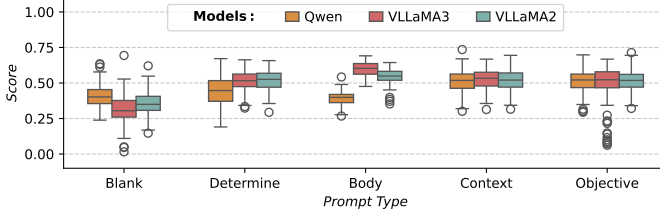


Figure 6. Cosine similarity between *expert* and *ground truth* captions, across prompts. The varying prompts do not score higher than 0.75 in any samples, with means from 0.31 in VideoLLaMA3 ‘Blank’ to 0.59 in VideoLLaMA3 ‘Body’. This indicates that specific prompts output more accurate captions than non-VLM captions. ‘Context’ and ‘Objective’ are around 0.51 across all models, while ‘Body’ is only more accurate when used with the VideoLLaMA models.

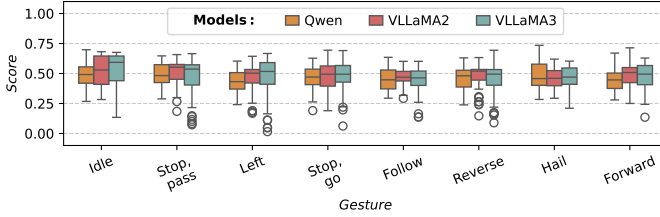


Figure 7. Cosine similarity between *expert* and *ground truth* captions, across gestures. No gesture interpreted by a VLM reaches a score of 0.75 in any sample. The means vary from 0.43 in Qwen ‘Left’ to 0.52 in VideoLLaMA3 ‘Idle’. Further, no specific gesture stands out as most readily interpretable.

Model	Caption
Ground Truth	<i>“The pedestrian is standing in front of the ego driver. They are facing their torso and head towards the ego driver. They are moving their flat palms back and forth towards the ego driver, gesturing it to reverse. They are moving slightly to the side.”</i>
Expert	<i>“The person pushes their hands that are facing the camera outwards and brings them back in almost as if their signaling to stop or slow down”</i>
Qwen	(abr.) <i>“The person in the image appears to be making a stop or “no” gesture with both hands extended forward and fingers spread apart, palm facing outward ...”</i>
VLLaMA2	<i>“The pedestrian is making a stop gesture.”</i>
VLLaMA3	<i>“The pedestrian is making a gesture with his hands.”</i>

Table 1. Examples of captions from the video sequence “Reverse” from frame 16 - 24 given the “Determine” prompt. We manually analyze the captions focusing on body movement, direction description, biases, and common traits. We see that the outputs from VideoLLaMA are short, with only VideoLLaMA2 giving a gesture response, albeit inaccurate. The example from Qwen is only a snippet, since its response is quite long, and it also responds inaccurately with ‘Stop’. Examples of generated captions from whole videos are shown in Table 2.

Model	Caption
Ground Truth	(abr.) <i>“..They look up, and move their hands back and forth towards me at their chest, indicating me to reverse..”</i>
Qwen	(abr.) <i>“The pedestrian appears to be waving or gesturing with both hands as they walk past the camera. The movement suggests that they might be saying hello, goodbye, or simply acknowledging someone..”</i>
VLLaMA2	<i>“The pedestrian is making a stop gesture.”</i>
VLLaMA3	<i>“The pedestrian is making a gesture with his hands.”</i>

Table 2. Examples of captions from the whole ‘Reverse’ video parsed the ‘Determine’ prompt (without expert-generated caption). The VideoLLaMA models output the same responses even when given more video information. Qwen outputs another inaccurate gesture interpretation. Examples of captions per 8 frames are shown in Table 1.

Model	Accuracy ↑	Precision ↑	Recall ↑	F1-Score ↑
Expert	0.72	0.71	0.72	0.70
Qwen	0.33	0.11	0.33	0.17
Vllama2	0.15	0.28	0.15	0.14
Vllama3	0.52	0.32	0.52	0.39

Table 3. Classification accuracy with 9 classes on ATG with 8-frame interval. All VLMs have difficulty interpreting gestures correctly, even toward this reduced task, with a best F1 score lower than 0.40. This shows they can perceive and interpret traffic gestures to some extent, but are unreliable toward autonomous driving in their current form. Processed expert annotations successfully classify ‘Left’, ‘Reverse’, and ‘Stop’ with F1 above 0.80. Qwen captions predict ‘Stop’ 56 times and ‘Hail’ 6 times out of 62. VideoLLaMA2 predicts ‘Hail’ 53 times out of 62. VideoLLaMA3 confuses ‘Left’ with ‘Forward’ and ‘Reverse’ with ‘Stop’, and has ‘Hail’ and ‘Stop’ with an F1 above 0.70.

4.3. Reconstruction

We design one additional method of evaluation of machine-generated captions, built around the premise that precise and detailed language should carry sufficient information for a movement to be reenacted accurately. For example, “raise one arm” can have many meanings, while “raise your left arm right above your head” is much more precise and less likely to be misunderstood. An example is visualized in Fig. 8.

This evaluation was constructed by reading the captions aloud to a participant, who would reenact the described movements as informed in the caption with their interpretation. This reenacted scene was compared with the original video to compute a coarse evaluation metric, using pose estimation and MSE upon the equivalent pose points. The reconstructed videos were cut only to contain the movement and sped up to match the length of the original video. Results of this analysis are shown in Fig. 9.

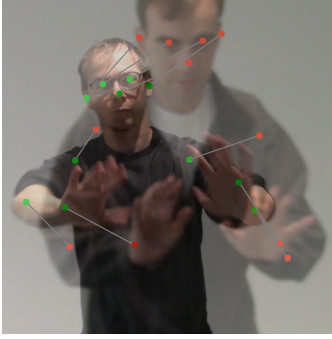


Figure 8. Reconstructed gesture using the ground truth caption, overlay upon the original ‘Reverse’ video (timestamp 00:03).

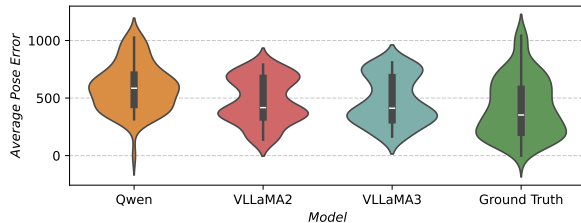


Figure 9. The results show MSE for every frame with the caption from the whole video ‘Reverse’. The participant is not the same height (Fig. 8) as in the original video, which makes it more challenging to be accurate, but leaves an opportunity for future research as an evaluation method. As a further complication for future consideration, it was found that it was either not temporally correct to reenact the gesture as it was read out loud, or could miss minor details if reenacted after reading the whole caption. However, the ground truth scored better than the VLM caption reenactments. The Qwen model had a mean score around 600, VideoLLaMA models around 400, and ground truth around 300. A few frames scored around 0, an exact pose in the same frame. However, considering the difficulty of complete synchronous pose matching, we consider this largely a coincidence with the large number of frames evaluated.

5. Discussion

Due to the desired application of the VLMs, we explored prompting for detailed captions and closed-set interpretation classes. Some of these models vary considerably in their capabilities of captioning human gestures. Occasionally, some captions would be spot-on, or more accurate to certain gestures like ‘Stop’ and ‘Hail’, but in all three evaluation methods, the VLMs had a lower score than the expert baseline. As the expert baseline was also considerably low, ambiguity and evaluation strategies merit future research for reliable interpretation of human traffic gestures.

For data collection and experimental design, we note that the expert-generated captions are biased in knowing the whole video and the name of the video, which gives more information about the gesture than the individual se-

quences. To increase fidelity, each expert should too have varying prompts like the VLMs and the ground truth captions, but this would require more participants to avoid ‘preknowledge’-biases. This further begs the additional question of bias: will pedestrians use the same gestures and behave the same, knowing the vehicle they are gesturing to is an autonomous vehicle and not a human driver? And what gestures will still be relevant? To ‘Hail’ an autonomous taxi, one may order only through their phone. For now, we assume pedestrians behave and use gestures towards the AV like a human driver.

5.1. Gesture Annotation

Annotating the videos could be done in many ways, especially regarding direction, which could also be challenging for the VLM or an LLM to understand. Directions could be explained from both the ego driver’s and the pedestrian’s perspective using the road.

As the captions are supposed to be combined as a prompt for an LLM to decide on an action, the captions should include enough information to assist with this decision. Assuming the caption is accurate, interpretation of the gesture should be enough for the LLM to avoid including a physical description of the pedestrian, which is less quantizable to discrete control decisions for the autonomous agent. However, if the model does not understand the complexity of the real-life scenario and misinterprets, the LLM can still interpret the gesture using the body analysis, depending on the accuracy of the VLM. Our study also highlights that sentence comparison is difficult to use as a proper evaluation, as it can be challenging to find a ‘language’ that expresses all important properties of the body in the context to get the whole picture. Annotating the captions of the pedestrian can be tricky and must follow specific rules, as direction alone can have meaning from multiple perspectives, describing it as first- or third-person. Additionally, ‘Left’ and ‘Right’ are not always enough to be precise, and can easily be misunderstood as another perspective or direction, and can be especially confounding in multiview input situations [40]. Using clock time or compass directions allows a broader scope of precision. This precision, however, may not be required and would be inefficient token utilization. In formal settings, directions do not require as much precision, whereas in informal settings, especially in emergencies, they require high precision.

We highlight some common characteristics here: P1 (‘Forward’, ‘Follow’, ‘Hail’) mentions the specific arm, uses “me” as an ego driver, and does not always include interpretation. P2 (‘Idle’, ‘Left’) changes from “person”, “guy”, and “subject” when talking about the pedestrian, states perspective of left and right, (*the driver’s left side*). P3 (‘Reverse’, ‘Stop + pass’) uses ‘person’ as the pedestrian. P4 (‘Stop, go’) states both perspectives from the driver and

the pedestrian. We highlight these differences to illustrate the inherent ambiguity in interpretation and language describing an instructive gesture.

5.2. Embedded Similarity - Method 1

The cosine similarity had some issues in terms of ‘Extended’ information. Especially Qwen habitually expanded the caption with additional and sometimes hallucinated information. Indeed, it would not be directly similar to the ground truth, but as an LLM would interpret the correct gesture even with additional information, it should be evaluated as similar. However, this would not be possible if the caption contained too many tokens for the LLM to gather all the information, and it is also inefficient; the goal is a complete and compact representation of the relevant information.

Additionally, to validate the metric selection further, the number of captions could be extended to reduce noise even more. Adding one more caption to each level found a slight improvement towards the ideal metric. With more samples, utilizing, e.g., a boxplot instead of a barplot would enable more accurate statistical analysis. Specifically, ‘Extended’ should be punished according to the length of the sentence.

5.3. Classification - Method 2

Requesting the VLM to respond with a single classification from possible classes may eventually be the most reliable way of evaluating task-specific capabilities. At least before continuing with more complex evaluation methods. However, our study shows that the VLMs are incapable of accurate classification. The models had a limited understanding, with output generalizing mainly to ‘Stop’ gestures.

This evaluation is only 0.70 accurate even with human expert captions. This was possibly due to the low number of 8 frames, making it difficult to interpret. Also, some classes were similar, like ‘Forward’ and ‘Follow’, which can be challenging to distinguish.

5.4. Reconstruction - Method 3

Using MSE to compute the error for the movement has some limitations left to future research regarding hand gestures that can change the intent of a gesture significantly, especially in relation to arm position and with regard to the movement of various pose keypoints. This would be accommodated by weighting the smaller body parts, but an open question is with what ratios? Further, temporal alignment of most-similar poses is another important area for future development of such a reconstruction metric. Forcing both to stand on a specific mark would eliminate the distance and location variable, to focus more on the gesture itself.

This evaluation method computes the pedestrian’s exact position, build, and movement, making an exact reenactment very difficult. Also, to a point where the details to reenact it correctly are not necessary to interpret the ges-

ture. Reconstructing the human movements could also be done using Video Generative Models (VGMs), to enable efficient data generation. This would also show how well an LLM interprets information and avoids human unconscious or cultural interpretation.

6. Concluding Remarks

In conclusion, this study evaluated VLM’s capability to recognize and caption human traffic gestures in the format of longer descriptions and interpretations. This was evaluated across three evaluation methods, each varying in abstraction, reasoning, and response details. Throughout the evaluation, the method results show that currently-trained VLMs are unreliable in capturing human traffic gestures with one individual participant in the frame.

The evaluation methods provide a range of evaluation foci. They can be applied in multiple evaluation studies containing information comparison in sentences, abstract human movement description forwarding, and concrete categorical human gesture classification.

6.1. Future Research

Expanding this study would include varying static and temporal inputs, additional models, and videos with varying but realistic pedestrians and scenes. Separate studies would look into varying window sizes and frame rates, and how the VLMs would behave on only crops of the individual pedestrian. Alternative methods to evaluate captions should be explored, such as utilizing an LLM to compare generated captions with ground truth or identifying missing information within the captions as a basis for evaluation.

Advancing VLMs in zero-shot to caption human movement and traffic gestures could be enforced by using pose models to help explain poses to the VLM or by augmenting the video by projecting poses upon the video.

Overall, this research highlights the relevance but difficulty of the task of gesture understanding for autonomous systems to safely navigate in driving environments where human interpretation and interaction are necessary. Future research in both model performance and evaluation can drive the development of interpretable and robust human-cooperative autonomous driving systems.

References

- [1] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, “Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data,” in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pp. 683–700, Springer, 2020. 1
- [2] A. Prakash, K. Chitta, and A. Geiger, “Multi-modal fusion transformer for end-to-end autonomous driving,” in *Proceed-*


- ings of the *IEEE/CVF conference on computer vision and pattern recognition*, pp. 7077–7087, 2021.
- [3] R. Greer, N. Deo, and M. Trivedi, “Trajectory prediction in autonomous driving with a lane heading auxiliary loss,” *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4907–4914, 2021. 1
 - [4] A. Alofi, R. Greer, A. Gopalkrishnan, and M. Trivedi, “Pedestrian safety by intent prediction: A lightweight lstm-attention architecture and experimental evaluations with real-world datasets,” in *2024 IEEE Intelligent Vehicles Symposium (IV)*, pp. 77–84, IEEE, 2024. 1
 - [5] P. Roy, S. Perisetla, S. Shriram, H. Krishnaswamy, A. Keskar, and R. Greer, “doscenos: An autonomous driving dataset with natural language instruction for human interaction and vision-language navigation,” *arXiv preprint arXiv:2412.05893*, 2024. 1
 - [6] D. Dey and J. Terken, “Pedestrian interaction with vehicles: roles of explicit and implicit communication,” in *Proceedings of the 9th international conference on automotive user interfaces and interactive vehicular applications*, pp. 109–113, 2017. 2
 - [7] A. Rasouli and J. K. Tsotsos, “Autonomous vehicles that interact with pedestrians: A survey of theory and practice,” *IEEE transactions on intelligent transportation systems*, vol. 21, no. 3, pp. 900–918, 2019. 2
 - [8] A. K. AlShami, A. Kalita, R. Rabinowitz, K. Lam, R. Bezbaruha, T. Boulton, and J. Kalita, “Cool: Challenge of out-of-label a novel benchmark for autonomous driving,” *arXiv preprint arXiv:2412.05462*, 2024. 2, 3
 - [9] S. Atakishiyev, M. Salameh, H. Babiker, and R. Goebel, “Explaining autonomous driving actions with visual question answering,” in *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1207–1214, IEEE, 2023. 2
 - [10] A. Keskar, S. Perisetla, and R. Greer, “Evaluating multi-modal vision-language model prompting strategies for visual question answering in road scene understanding,” in *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 1027–1036, 2025. 2
 - [11] H. Shao, Y. Hu, L. Wang, G. Song, S. L. Waslander, Y. Liu, and H. Li, “Lmdrive: Closed-loop end-to-end driving with large language models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15120–15130, 2024. 2
 - [12] R. Greer and M. Trivedi, “Towards explainable, safe autonomous driving with language embeddings for novelty identification and active learning: Framework and experimental analysis with real-world data sets,” *arXiv preprint arXiv:2402.07320*, 2024. 2
 - [13] R. Greer, B. Antoniusen, A. Møgelmoose, and M. Trivedi, “Language-driven active learning for diverse open-set 3d object detection,” in *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 980–988, 2025. 2
 - [14] Z. Xu, Y. Zhang, E. Xie, Z. Zhao, Y. Guo, K.-Y. K. Wong, Z. Li, and H. Zhao, “Drivept4: Interpretable end-to-end autonomous driving via large language model,” *IEEE Robotics and Automation Letters*, 2024. 2
 - [15] L. Chen, P. Wu, K. Chitta, B. Jaeger, A. Geiger, and H. Li, “End-to-end autonomous driving: Challenges and frontiers,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2
 - [16] N. Madan, A. Møgelmoose, R. Modi, Y. S. Rawat, and T. B. Moeslund, “Foundation models for video understanding: A survey,” *Authorea Preprints*, 2024. 2
 - [17] H. Arai, K. Miwa, K. Sasaki, Y. Yamaguchi, K. Watanabe, S. Aoki, and I. Yamamoto, “Covla: Comprehensive vision-language-action dataset for autonomous driving,” *arXiv preprint arXiv:2408.10845*, 2024. 2, 4
 - [18] R. Greer, A. Gopalkrishnan, J. Landgren, L. Rakla, A. Gopalan, and M. Trivedi, “Robust traffic light detection using salience-sensitive loss: Computational framework and evaluations,” in *2023 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1–7, IEEE, 2023. 3
 - [19] R. Greer, J. Isa, N. Deo, A. Rangesh, and M. M. Trivedi, “On salience-sensitive sign classification in autonomous vehicle path planning: Experimental explorations with a novel dataset,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 636–644, 2022.
 - [20] R. Greer, A. Gopalkrishnan, N. Deo, A. Rangesh, and M. Trivedi, “Salient sign detection in safe autonomous driving: Ai which reasons over full visual context,” in *27th International Technical Conference on the Enhanced Safety of Vehicles (ESV) National Highway Traffic Safety Administration*, no. 23-0333, 2023. 3
 - [21] R. Zhao, Q. Yuan, J. Li, Y. Fan, Y. Li, and F. Gao, “Drivellava: Human-level behavior decisions via vision language model,” *Sensors (Basel, Switzerland)*, vol. 24, no. 13, p. 4113, 2024. 3
 - [22] A. Amir, B. Taba, D. Berg, T. Melano, J. McKinstry, C. Di Nolfo, T. Nayak, A. Andreopoulos, G. Garreau, M. Mendoza, et al., “A low power, fully event-based gesture recognition system,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7243–7252, 2017. 3
 - [23] S. Escalera, J. González, X. Baró, M. Reyes, O. Lopes, I. Guyon, V. Athitsos, and H. Escalante, “Multi-modal gesture recognition challenge 2013: Dataset and results,” in *Proceedings of the 15th ACM on International conference on multimodal interaction*, pp. 445–452, 2013.
 - [24] B. Hamner, Isabelle, LoPoal, sescalera, and xbaro, “Multi-modal gesture recognition.” <https://kaggle.com/competitions/multi-modal-gesture-recognition>, 2013. Kaggle.
 - [25] S. Ruffieux, D. Lalanne, E. Mugellini, and O. Abou Khaled, “A survey of datasets for human gesture recognition,” in *Human-Computer Interaction. Advanced Interaction Modalities and Techniques: 16th International Conference, HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014, Proceedings, Part II 16*, pp. 337–348, Springer, 2014. 3
 - [26] J. Wiederer, A. Bouazizi, U. Kressel, and V. Belagiannis, “Traffic control gesture recognition for autonomous vehicles. in 2020 ieeec,” in *RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10676–10683. 3

- [27] O. Kobzarev, A. Lykov, and D. Tsetserukou, “Gestllm: Advanced hand gesture interpretation via large language models for human-robot interaction,” *arXiv preprint arXiv:2501.07295*, 2025. 3
- [28] Z. Cheng, S. Leng, H. Zhang, Y. Xin, X. Li, G. Chen, Y. Zhu, W. Zhang, Z. Luo, D. Zhao, *et al.*, “Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms,” *arXiv preprint arXiv:2406.07476*, 2024. 4
- [29] B. Zhang, K. Li, Z. Cheng, Z. Hu, Y. Yuan, G. Chen, S. Leng, Y. Jiang, H. Zhang, X. Li, *et al.*, “Videollama 3: Frontier multimodal foundation models for image and video understanding,” *arXiv preprint arXiv:2501.13106*, 2025. 4
- [30] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, *et al.*, “Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution,” *arXiv preprint arXiv:2409.12191*, 2024. 4
- [31] T. Dao, “Flashattention-2: Faster attention with better parallelism and work partitioning,” *arXiv preprint arXiv:2307.08691*, 2023. 4
- [32] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” *arXiv preprint arXiv:1908.10084*, 2019. 5
- [33] G. Salton, “Modern information retrieval,” (*No Title*), 1983. 5
- [34] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “Bertscore: Evaluating text generation with bert,” *arXiv preprint arXiv:1904.09675*, 2019. 5
- [35] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002. 5
- [36] P. Jaccard, “Étude comparative de la distribution florale dans une portion des alpes et des jura,” *Bull Soc Vaudoise Sci Nat*, vol. 37, pp. 547–579, 1901. 5
- [37] S. Banerjee and A. Lavie, “Meteor: An automatic metric for mt evaluation with improved correlation with human judgments,” in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005. 5
- [38] L. Chin-Yew, “Rouge: A package for automatic evaluation of summaries,” in *Proceedings of the Workshop on Text Summarization Branches Out, 2004*, 2004. 5
- [39] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, *et al.*, “Transformers: State-of-the-art natural language processing,” in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45, 2020. 5
- [40] R. Greer and M. Trivedi, “Ensemble learning for fusion of multiview vision with occlusion and missing information: Framework and evaluations with real-world data and applications in driver hand activity recognition,” *arXiv preprint arXiv:2301.12592*, 2023. 7

B Co-author Statement and Signature for [4]

Co-author Statement and Signatures

The undersigned certifies that they had contributed to the included paper "*Can Vision-Language Models Understand and Interpret Dynamic Gestures from Pedestrians? Pilot Datasets and Exploration Towards Instructive Nonverbal Commands for Cooperative Autonomous Vehicles*". Their contribution consisted of supervision, co-writing, revising, and refining.

Name:	<u>Ross Greer</u>
Email:	<u>rossgreer@ucmerced.edu</u>
Signature:	<u></u>
Date:	<u>3 June 2025</u>

C Review Analysis of the “VLM Evaluation” Paper

This section summarizes and analyses the received reviews for the first version of the “VLM Evaluation” [4] paper. This helps determine proper changes, additions, and rewriting. We describe a brief understanding of the review, a discussion, and a camera-ready revised version.

Strengths/Pros

1. Relatively novel, pioneering, and relevant research area.
2. Figures and tables offer a clear and intuitive representation.
3. Datasets fill a gap in gesture-based interaction (*Very strong*).
4. Evaluation from varying angles.
5. Prompt design insights illustrate influence on VLMs.
6. Similar results demonstrate a lack in VLMs.
7. Discussion section about limitations

Weaknesses/Cons

1. General

- (a) Title is excessively long.
- (b) The Length of sentences is too long.

2. Related work

- (a) Figure 2, clarity in caption, as the girl is difficult to understand (*Note: Maybe add full image. I have maybe seen myself blind upon it*)
- (b) “Related work” section should be broader. VLM outside of AV. And influence to study. More than two references. At least ten.

3. Dataset

- (a) Add more variation to the acted dataset (distance, people, gestures) (*Note: Wanted to but didn’t have time..*)
- (b) Add camera specs and parameters
- (c) Scale of datasets is unclear. More specific data and labels. Greater details

4. Models

- (a) Include more VLMs (CLIP, ViLA, etc.) (*This was tried, without success*)

5. ‘Baseline’

- (a) Lack of a non-VLM baseline, such as pose estimation. Non-contextualization of the capabilities of VLMs.

(b) Include VLMs capability upon general gesture recognition datasets!

6. Semantic

(a) Semantic bias in similarity. May not work using cosine similarity. More claims this. More metrics.

(This should be looked into)

(b) Clarify the selection of similarity metrics

(c) The “Extended” caption seems unnecessary. Clarify

7. Classification

(a) Classification using ChatGPT introduces additional bias. *(Note: This was only done with the export annotators. Clarify or redo)*

8. Reconstruction

(a) Reconstruction is very subjective. Interpretation, embodiment, and timing. *(Note: Clarify, that is the purpose - to illustrate this.)*

9. Evaluation

(a) Clear explanation of which dataset is used in what evaluation

10. Format

(a) Figure, table, and colors format according to CVPR guidelines

(b) Reference to original published papers *(I did that?)*

11. Discussion

(a) Add more to limitations, like real-world deployment,

(b) Discuss the reason for the desired caption output, not simply classification.

12. Other

(a) Lack of proposing innovation

Forward Plan

The research of the paper leads to a lot of changes, and additions in the paper (e.g., context window, additional non-VLM baseline, VLM advancing), but instead of changing them in this paper, it is discussed and used for future development of the dataset and other papers referring to this paper’s flaws. *(Trial and error)*. This concludes the forward plan of changes and additions to this paper following the reviews. A short peek into what the paper leads up to is mentioned below.

Plan for camera-ready version:

1. ✓ Formulate parts of the paper that may seem unspecific, unclear, and lacking in detail
2. ✓ Format CVPR guidelines references, figures, tables, etc.
3. ✓ Add an ‘Ideal’ representation of the metric validation similarity trend
4. ✓ Expand Related work with datasets that don’t compile, etc.

5. ✓ Conclude the property of only eight frames, and future work will be across whole videos, as each video is a single actor and gesture.
6. ✓ Expand the acted dataset, with a single caption and classification for each video (can be done after)
7. (Opt. include CLIP, ViLA, etc. (if works))

Plan after and maybe for the presentation:

1. Enhancing methods: Augmented, described body (fine-tuning)
2. (**‘international ground-to-air emergency signaling system’**)

D Post ‘VLM Evaluation Paper’ Planning

1. Evolve VLM evaluation paper
 - Record and add more videos, even of me or other people.
 - Add the instructive real-life data, with annotations
 - General clean up
2. Advance VLM models in traffic gesture detection
 - I found a paper advancing VLM in hand gestures enhanced by hand pose
 - Use the classification method to evaluate first, and the sentence comparison later
 - After VLMs didn’t work, my original idea was to train a classifier from a traffic gesture dataset I would make myself. That’s also an option, to make that dataset, and separate the VLM and gesture classification. More or less
3. Evolve reconstruction method
 - Try more aspects, such as single frame, fixed position, etc.
 - Utilize GenAI
4. Zero-shot VLM and LLM in pedestrian classification and hierarchy ranking evaluation
 - Clean up the acted multi-pedestrian dataset (maybe write a publishable paper)
 - Make a similar evaluation as we just did

1 makes sense to make a more reliable point, but to do more diverse development too, 2 would be good. I’ve started a bit on 2 before WACV, and this should hopefully prove the evaluation methods work. 3 could be fun, but it’s a bit of an entirely new topic. 4 is a part of the bigger picture, and the cleaning, sorting, and annotating of the dataset at least needs to be completed with the help of the undergrads, which I’ll be in charge of.