MMHU: A LARGE-SCALE MULTIMODAL BENCHMARK FOR HUMAN BEHAVIOR UNDERSTANDING IN AUTONOMOUS DRIVING

Anonymous authors
Paper under double-blind review

Figure 1: We propose **MMHU**, a large-scale dataset for human behavior understanding. We collected 57k human instances with diverse behaviors such as playing with a mobile phone, holding objects, or using mobility devices, from diverse scenes such as in the city, school, park, and alley. We provide rich annotations, including motion and trajectory, text descriptions for human motions, and the labels for behaviors that are critical to driving safety.

ABSTRACT

Humans are integral components of the transportation ecosystem, and understanding their behaviors is crucial to facilitating the development of safe driving systems. Although recent progress has explored various aspects of human behavior—such as motion, trajectories, and intention—a comprehensive benchmark for evaluating human behavior understanding in autonomous driving remains unavailable. In this work, we propose **MMHU**, a large-scale benchmark for human behavior analysis featuring rich annotations, such as human motion and trajectories, text description for human motions, human intention, and critical behavior labels relevant to driving safety. Our dataset encompasses 57k human motion clips and 1.73M frames gathered from diverse sources, including established driving datasets such as Waymo, in-the-wild videos from YouTube, and self-collected data. A human-in-the-loop annotation pipeline is developed to generate rich behavior captions. We provide a thorough dataset analysis and benchmark multiple tasks—ranging from motion prediction to motion generation and human behavior question answering—thereby offering a broad evaluation suite. Our dataset will be released to promote further human-centric research in this vital area of autonomous driving.

1 INTRODUCTION

Humans play an essential role in transportation systems, making the comprehensive understanding of human behaviors—such as motion (Xu et al., 2023), intention (Osman et al., 2023; Xie et al., 2024; Yang et al., 2022), and trajectory (Zhang et al., 2024c; Fang et al., 2024)—critical for developing safe autonomous driving systems. To effectively interact with humans, autonomous vehicles must answer human-centric questions, such as "What is the person doing?", "Is the person going to cross the street?", and "Where does the person intend to go?" Failing to accurately comprehend these behaviors could lead to misinterpretations of human intent, potentially resulting in fatal accidents.

While significant efforts have been devoted to understanding individual aspects of human behaviors by investigating human motion, intention, and trajectory, the absence of a unified dataset limits the comprehensive evaluation of algorithms for human behavior understanding, especially in autonomous driving scenarios that mainly account for human safety. Existing driving datasets are typically designed for general driving tasks, such as depth estimation, 2D or 3D object detection, odometry, and semantic segmentation (Sun et al., 2020; Caesar et al., 2020; Geiger et al., 2013), or narrowly designed human-related tasks, such as intention prediction (Rasouli et al., 2019; Bhattacharyya et al., 2021; Rasouli et al., 2017b), motion and trajectory prediction (Von Marcard et al., 2018), or motion reconstruction (Wang et al.; Kim et al., 2024; Liu et al., 2024c). Moreover, with the emergence of driving-oriented vision-language models (VLMs) (Touvron et al., 2023; Lin et al., 2023a; Chen et al., 2024c; Liu et al., 2024b; 2023b; 2024a; 2023a; Zhang et al., 2024b; Bai et al., 2023) , human behavior understanding tasks can now be approached in a more integrated and flexible manner through images and text queries. However, existing training data for these VLMs are not specifically tailored to human behavior, limiting their effectiveness in capturing critical human-centric details essential for safe driving.

In this work, we aim to answer three core questions regarding human behavior understanding tasks in autonomous driving scenarios: ❶ What aspects of human behavior are critical to autonomous driving? ❷ How effectively do current approaches model human behaviors in autonomous driving contexts? ❸ How can a comprehensive benchmark advance the development of human behavior understanding algorithms? To this end, we propose **MMHU**, a large-scale unified benchmark explicitly designed for comprehensively understanding various human behaviors in driving scenarios. MMHU includes rich annotations generated by a human-in-the-loop annotation pipeline, enabling scalable and precise labeling from diverse data sources using only monocular video inputs. Specifically, we have collected 1.73M frames featuring 57K human instances from source videos obtained from Waymo (Sun et al., 2020), YouTube, and self-collected data. As shown in Fig. 1, the dataset provides detailed annotations covering: **(1)** human motion and trajectory; **(2)** text descriptions of human motions generated using templates and VLMs; **(3)** critical human behaviors extracted via VLMs, along with question-answer (QA) pairs designed to benchmark driving-oriented and generalist VLMs.

Our contributions can be summarized as follows:

- We introduce **MMHU**, a unified, human-centric dataset that provides a comprehensive understanding of humans' behaviors in driving scenarios that can be used as a benchmark for a range of human-centric understanding tasks.
- We develop a scalable, human-in-the-loop annotation pipeline employing multi-source fitting strategies to produce accurate labeling across diverse video sources, ranging from driving videos and general YouTube videos to self-collected streams.
- We evaluated baseline methods of human behavior understanding and analyze their performance, we further demonstrated that our dataset helps these methods achieve better performance.

2 RELATED WORKS

2.1 HUMAN MOTION

Human motion is essential to autonomous driving. We categorize human motion representations into 2D and 3D representations. 2D human motion (Jiang et al., 2024b; Belagiannis & Zisserman, 2017; Luo et al., 2021; Li et al., 2021; Jin et al., 2020) leverages keypoints or heatmaps to mark the local body motion on the image. For 3D representations, the SMPL series (Loper et al., 2015; Romero et al., 2017; Pavlakos et al., 2019) provide compact and expressive representation via learned parameters. While most human motion datasets focus on general human motions (Ionescu et al.,

2013; Von Marcard et al., 2018; Xu et al., 2024; Mahmood et al., 2019; Lin et al., 2023b), there are several datasets are specially designed for driving scenarios (Wang et al.; Liu et al., 2024c; Kim et al., 2024). However, these datasets mainly focus on the human movements and their text description, the behaviors of humans remain unavailable. Based on the representation and datasets, several efforts have been put into human motion reconstructions from temporally aligned multi-view cameras (Huang et al., 2021), unaligned multi-view cameras (Dong et al., 2020), and monocular cameras (Luvizon et al., 2023; Li et al., 2022; Ye et al., 2023; Yuan et al., 2022). Other works have explored human motion generation from action labels (Cervantes et al., 2022) or text (Tevet et al., 2022; Zhang et al., 2024a). However, due to the lack of high-quality data, there is little work that specifically generates human motion in driving scenarios.

2.2 Human Behavior Understanding

Understanding human behavior in driving situations is essential for driving safety. Although there are some datasets and methods to understand human behavior and actions (Zhang et al., 2024c; Rai et al., 2021; Punnakkal et al., 2021; Wang et al., 2012; Li et al., 2010; Soomro et al., 2012; Niebles et al., 2010; Kay et al., 2017; Marszalek et al., 2009), they mainly focus on recognizing human actions in general or sports scenes. While sharing some common behavior that concerns driving safety, they mainly focus on general actions like shaking hands, dancing, or running. The behaviors specifically concerning driving safety remain unexplored. In autonomous driving, besides motion reconstruction, there are some approaches and datasets for understanding several aspect of human behaviors such as (1) human trajectory prediction (Zhang et al., 2024c; Goncalves & Busso, 2022; Medina et al., 2024; Wang et al., 2024c; Guo et al., 2023; Mao et al., 2020), where models are required to predict the future trajectory from previous ones, or (2) human intention prediction (Zhang et al., 2023; Osman et al., 2023; Sharma et al., 2023; Rasouli et al., 2019; Kotseruba et al., 2021; Rasouli et al., 2017a), where pedestrians are simply classified into two states - crossing the street and not crossing the street. Besides the binary classification of crossing the street, there are several datasets Kwak et al. (2017); Quintero et al. (2015); Schneider & Gavrila (2013); Rasouli et al. (2017b) that provide more detailed behavior labels such as stopping, glancing, or running. However, these works still focus on specified aspects of human behavior, such as the posture and action when a pedestrian is crossing the street. Recently, the development of vision language models (VLMs) (Touvron et al., 2023; Lin et al., 2023a; Chen et al., 2024c; Liu et al., 2024b; 2023b; 2024a; 2023a; Zhang et al., 2024b; Bai et al., 2023) enables question-answering based on images or videos, making human behavior understanding more flexible. There have been many specialists driving VLMs (Ma et al., 2024; Sima et al., 2024; Chen et al., 2023a; Shao et al., 2024; Wang et al., 2024b; Yuan et al., 2024; Chen et al., 2024a) and autonomous driving QA datasets (Marcu et al., 2024; Arai et al., 2024; Nie et al., 2024; Chen et al., 2024a; Inoue et al., 2024; Qian et al., 2023). However, these models and the datasets are designed for general VQA tasks for autonomous driving. The comprehensive understanding of human behaviors remains unexplored. We show a comparison of related datasets in Tab. 1.

2.3 Autonomous Driving Datasets

Autonomous driving has been one of the most popular research topics in recent years. There are several datasets that are specially created for developing and evaluating autonomous driving algorithms (Geiger et al., 2013; Sun et al., 2020; Caesar et al., 2020; Maddern et al., 2017). These datasets are typically collected from a vehicle mounted with multiple sensors, such as multi-view cameras, LiDARs, RaDARs, IMU, etc., supporting autonomous driving tasks such as 2D and 3D object detection, semantic segmentation, depth estimation, and planning. Further, some works such as BDD-X (Kim et al., 2018) and Rank2Tell (Sachdeva et al., 2024) have put significant efforts into improving the explainability of decisions made by the automatically driven vehicles. WOMD-Reasoning (Li et al., 2024) leveraged commercial VLM to generate large-scale interaction reasoning data in driving. However, it is built for general autonomous driving, lacking structured labeling of human behaviors. In addition, the dataset is totally labeled by VLMs without human involvement. The labeling quality will be highly restricted by the VLM's ability and inflexible to further incorporate human knowledge. Recently, several works have focused on some specific scenes in autonomous driving, such as the accident (Fang et al., 2021), snowy scenes (Chen et al., 2023b), and foggy scenes (Sakaridis et al., 2018). Some datasets have provided the labeling of several aspects of human behaviors in driving scenarios, such as human motion and trajectory (Kim et al., 2024; Wang et al.; Liu et al., 2024c; Taketsugu et al., 2025; Saadatnejad et al.), intention of crossing the street (Rasouli et al., 2017b; Bhattacharyya et al., 2021; Rasouli et al., 2019), or in the form of general VQA (Arai et al., 2024; Qian et al., 2023; Inoue et al., 2024; Sima et al., 2024). However, these datasets only

investigate some specific human behaviors, and the comprehensive understanding of human behavior remains unexplored.

3 THE MMHU DATASET

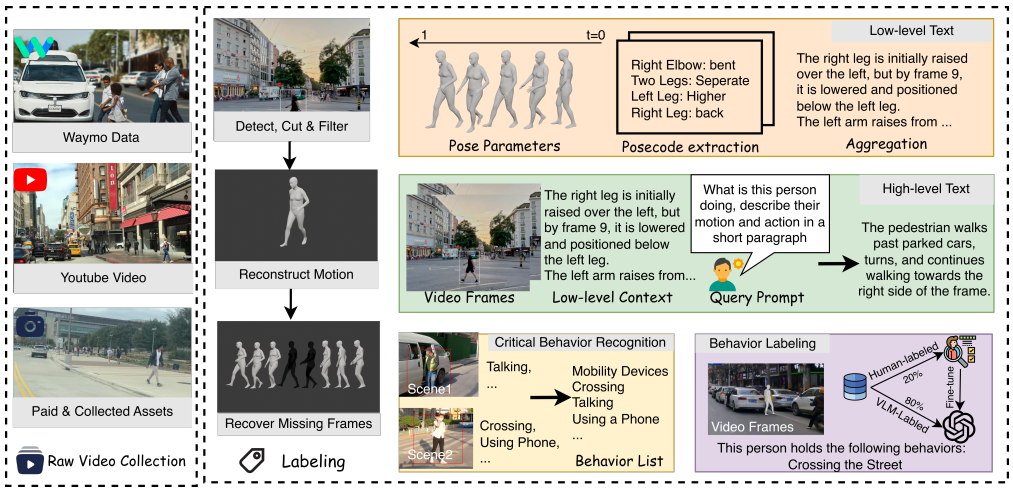


Figure 2: **Data Collection and Annotation.** (Left) We collect data from three sources: the Waymo dataset, the YouTube videos, and the self-collected or paid driving videos. (Right) We demonstrate the annotation pipeline; we first filter and cut the raw videos based on the rough human detection results. Then we reconstruct the SMPL motion for each detected frame. The missing frames are further recovered by an interpolation procedure. For the labeling of text descriptions, we leverage low-level text as a bridge between the SMPL parameters and the semantic label. Then we generate the high-level text from the low-level ones and the video. We recognize the critical behavior lists by leveraging a VLM to go through the Waymo (Sun et al., 2020) data, based on the visual and text information. For behavior labeling, we first employ human annotators to label a small portion of the data, then a VLM is fine-tuned on the human-labeled data and used to label the rest of the instances.

Overview. We propose **MMHU**, a comprehensive human-centric benchmark with rich annotations, emphasizing the criticalness of human behavior understanding in autonomous driving. We built our dataset using high-quality videos from various sources. Then we applied a scalable annotation pipeline that only involves minor human effort to get the rich annotations from the collected videos. The annotation for each video clip includes the 3D motion with trajectory, intention, high- and low-level text description, and critical behavior labels concerning driving safety. We will present our data collection details in Section 3.1. We then introduce the annotation pipelines and data analysis of human motion and trajectory (Sec. 3.2), text description (Sec. 3.3), and intention and critical behaviors (Sec. 3.4). We show the statistics of the dataset in Sec. 3.5. We show some visualization of the data sampled from MMHUin Fig. 5 in the Appendix.

3.1 DATA COLLECTION

Videos Acquisition We collected raw videos of 1.73M frames in total from three kinds of data sources as follows: Autonomous Driving Data contains multi-modality sensor information and rich annotation for generic autonomous driving tasks such as object detection, depth estimation, etc. We collected 1.7 hours videos from Waymo (Sun et al., 2020), which consist of 73K frames. In-the-wild Data includes first-person driving videos that are publicly available on the Internet. We collected 10 hours of YouTube videos, each with a CC license, consisting of 318.25K frames with a resolution ranging from 1080p to 2k. Self-collected Data is driving recordings collected by ourselves or from paid sources. We collected 66.5 hours of videos, consisting of 2393.96 k frames and the resolution varies from 1080p to 4k.

Video Cutting and Filtering Directly applying the annotation pipeline to the entire video can be expensive. We first roughly detect the human presence and filter out the frames that lack human presence. Specifically, we employ Yolo-V8 (Varghese & Sambath, 2024; Jocher et al., 2023) as a human detector to detect the human presence on the raw video at 1 FPS. Then cut the raw video into

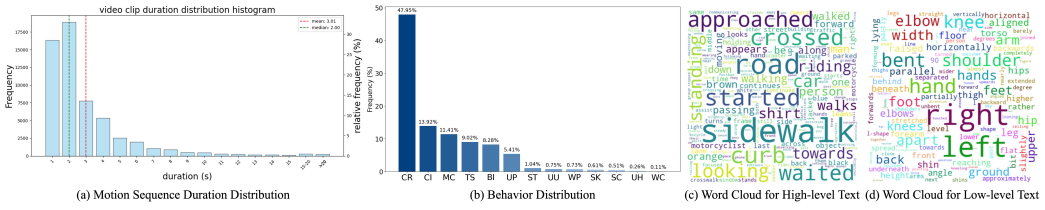
Table 1: **Comparison of Related Datasets.** We compare our dataset with related datasets. From left to right, the columns are the dataset name, total frame count or time duration, human count, providing labels of motion, trajectory, VQA pairs, and text descriptions, and the number of behavior classes. ‡ represents datasets for general purposes. Datasets labeled with * in the "Instance" column capture the motion from real participants. † means upper-bound; § means only consisting the rough direction of the trajectory; ¶ means estimated number. Datasets labeled with "Unstructured" in the last column do not provide explicit human behavior labels but involve them in the QA pairs or captions. Our dataset supports all four tasks and provides the most behavior classes among the driving datasets.

Dataset	Frames / Duration (s)	Instances	Motion	Trajectory	VQA	Text	Behaviors
PIE (Rasouli et al., 2019)	293k / -	1.3k	✗	✗	✗	✗	1
Euro-PVI (Bhattacharyya et al., 2021)	83k / -	7.7k	✗	✗	✗	✗	1
PMR (Wang et al.)	225k / -	54*	✓	✓	✗	✗	✗
CityWalker (Liu et al., 2024c)	- / 110k	120k	✓	✓	✗	✓	✗
BlindWays (Kim et al., 2024)	300k / 10k	11*	✓	✓	✗	✓	Unstructured
3DPW‡ (von Marcard et al., 2018)	51k / 1.7k	7*	✓	✓	✗	✗	✗
Human3.6M‡ (Ionescu et al., 2013)	3.6M / -	11*	✓	✓	✗	✗	15‡
JAAD (Rasouli et al., 2017b)	82k / 33k†	2.2k	✗	✗	✗	✗	11
Drama (Malla et al., 2023)	- / 36k	-	✗	✗	General Driving	✓	Unstructured
CoVLA (Arai et al., 2024)	6M / -	-	✗	✗	General Driving	✓	Unstructured
DriveLM-nuScenes (Sima et al., 2024)	4.8k / -	-	✗	✗	General Driving	✓	Unstructured
nuScenes-QA (Qian et al., 2023)	340k / -	-	✗	✗	General Driving	✓	✗
BDD-X (Kim et al., 2018)	8.4M / 277k	-	✗	✗	General Driving	✓	Unstructured
Rank2Tell (Sachdeva et al., 2024)	23k¶ / 2.3k¶	-	✗	✓§	General Driving	✓	12
MMHU (Ours)	1.73M / 173K	57K	✓	✓	Human-Centric	✓	13

fragments separated by frames without human presence. We then drop the fragments that are less than two seconds, which are of high probability to produce very short human motion sequences.

3.2 Motion and Trajectory

Extraction. Human motion provides rich information about human actions and behaviors. We employed the widely-used SMPL (Loper et al., 2015) to represent the human motion and trajectory. The details of SMPL representation is described in Sec. B.1. After obtaining the roughly filtered fragments as described in Sec. 3.1, we further extract the video clip of each human instance in each fragment. The details are described in Appendix Sec. C.3. Then, given a video clip $V = \{I_0, ..., I_T\}$ for a human instance, we extract a SMPL (Loper et al., 2015) parameter sequence $\mathcal{S} = \{S_t = (\Theta_t, \boldsymbol{r}_t, \boldsymbol{\pi}_t) | t \in \{1, ..., T\}\}$ following (Shin et al., 2024), where S_t is the SMPL parameters for the t^{th} frame, $\Theta_t \in \mathbb{R}^{n\times 3}$ the human motion, $\boldsymbol{r}_t \in \mathbb{R}^3$ the human orientation and $\boldsymbol{\pi}_t \in \mathbb{R}^3$ the location. We put the orientation and location of the human together to represent the human trajectory, denoted as $\Gamma_t = (\boldsymbol{r}_t, \boldsymbol{\pi}_t)$.



(a) Motion Sequence Duration Distribution (b) Behavior Distribution (c) Word Cloud for High-level Text (d) Word Cloud for Low-level Text

Figure 3: **Statistics of MMHU.** The average duration of motion sequences is 3s. The most common behavior is crossing the street, while the rarest behavior is using a wheelchair. For behavior definition, please refer to Sec. 3.4.

Completion. The SMPL sequence of a person can lack a few frames in the middle due to occlusion. To make the motion sequence consistent, we employed a missing motion prediction procedure to complete the missing frames. The human motions Θ_t of the missing frames are completed using angle-based interpolation from the nearest previous and following frames. The missing trajectory parameters are completed using linear interpolation. Specifically, given reconstructed SMPL parameter sequence $\mathcal{S} = \{S_1, S_2, ..., S_{k-1}, ..., S_{k+m+1}, S_{k+m+2}, ..., S_n\}$ with m missing frames $\{S_k, S_{k+1}, ..., S_{k+m}\}$,

the predicted SMPL parameter $S_{k+j} = \{\Theta_{k+j}, \Gamma_{k+j}\}$ at frame $k+j$ is interpolated as follows:

$$\hat{\Theta}_{k+j} = \frac{\sin((1-\frac{j}{m})\theta)}{\sin(\theta)}\hat{\Theta}_{k-1} + \frac{\sin(\frac{j}{m}\theta)}{\sin(\theta)}\hat{\Theta}_{k+m+1} \tag{1}$$

$$\theta = \arccos\langle\Theta_{k-1}, \Theta_{k+m+1}\rangle \tag{2}$$

$$\Gamma_{k+j} = \frac{m+1-j}{m+2}\Gamma_{k-1} + \frac{j+1}{m+2}\Gamma_{k+m+1} \tag{3}$$

3.3 Hierarchical Text Annotation

In addition to the parameterized SMPL motion, the semantic understanding of human behaviors is also critical. We use text as a semantic-level description of the human motion. To narrow the domain gap between the semantic text description and the SMPL parameter, we employ a hierarchical text annotation approach. We first convert the SMPL parameters to an element-level text description for each part of the body at each frame in a rule-based schema. Then we utilize large language models to aggregate the low-level descriptions of each joint over time. Based on the detailed low-level motion description, combined with the video clip, we abstract the high-level descriptions for each motion sequence. The details of hierarchical text annotation are introduced in the supplementary materials. Low-level Text Annotation: As shown in Fig. 2, the low-level text describes the movement of each body part in detail. Following (Delmas et al., 2022), we generate the low-level text description for the SMPL motion S_t at each frame t by calculating the angle, distance, position relation, etc. of different body parts. We then aggregate the movement of each body part over time to get the low-level description of human motion. High-level Text Annotation: High-level captions provide a semantic-level description of human action and motion. We generate one high-level description for the motion sequence of each person, leveraging large vision language models. We provide the VLM with the low-level descriptions and several frames uniformly sampled from the corresponding video clip. The VLM is then prompted to summarize a high-level text description for the video clip.

3.4 Critical Behaviors

Understanding human behaviors — for example, crossing the street — is critical for autonomous driving algorithms. We model the critical behaviors as binary attributes, indicating whether a person is subject to the corresponding behavior.

Critical Behavior Recognition Before we can assign the values to each attribute, we should first answer the question: what behaviors are critical to autonomous driving? One of them might be whether a person is going to cross the street, which is known as the intention prediction task in autonomous driving. As illustrated in Fig. 2, we recognize the critical behaviors by leveraging VLMs. Specifically, we sample n video clips from the dataset. For each video clip, we ask the VLM to recognize the critical behaviors in the scene. Lastly, the answers are collected, and a VLM is instructed to summarize the critical behaviors for autonomous driving. We recognize 13 behaviors, namely walking pets (WP), talking (TS), using a phone (UP), using an umbrella (UU), using headphones (UH), carrying items in hand (CI), crossing the street (CR), using wheelchair (WC), using stroller (ST), riding bike (BI), riding scooter (SC), using skateboard (SK), and riding motorcycle (MC).

Critical Behavior Labling Given the recognized behavior set $\mathcal{B} = \{\boldsymbol{b}_k | k \in \{1, ..., m\}\}$, the label of critical behaviors for a person h is defined as the subset $\mathcal{B}_h \subseteq \mathcal{B}$ in which the behaviors hold for the person. For each instance, we enumerate each element in the behavior set $b_k \in \mathcal{B}$ and construct a corresponding question q_k. Then we provide the corresponding frames to a VLM and ask it about the question q_k. Based on the answer, we append b_k to the behavior set $\mathcal{B}_h$ for the person. Directly applying pre-trained VLMs can suffer from noisy labeling results. To alleviate this, we perform a human-in-the-loop labeling strategy. We randomly selected a small portion of the dataset and employed human annotators to label whether the given person has certain behaviors. Then we use the human-labeled data to fine-tune the annotation VLM, which will further be applied to the rest of the unlabeled data. The details of the human annotation are described in the supplementary materials.

Table 2: **Motion Generation Evaluation.** Given the high FID distance, the generic text-to-motion models cannot properly generate motions in driving scenes.

Model	FID ↓	Multi Modality
Real	0.002	-
MotionDiffuse (Zhang et al., 2024a)	39.275	2.36
MotionGPT (Jiang et al., 2023)	27.059	5.42

Table 3: **Evaluation of Human Behavior Visual QA.** We construct closed-ended questions where the model is asked to select whether the person is subject to certain behaviors. We report the F1 score for each behavior, and then an instance-averaged F1 score is used to evaluate the overall performance. For behavior definition, please refer to Sec. 3.4.

Baseline	WP	TS	UP	UU	UH	CI	CR	WC	ST	BI	SC	SK	MC	Micro-F1
Phi-4-multimodal	42.9	35.6	24.6	90.9	15.4	39.4	26.1	100.0	31.2	56.3	66.7	33.3	77.1	45.5
MiniCPM-o-2_6	75.0	27.4	58.6	100.0	44.4	33.4	26.4	85.7	77.8	88.9	80.0	50.0	73.6	52.2
Dolphins	42.9	2.3	1.0	22.2	40.0	0.3	16.9	40.0	6.5	1.7	40.0	57.1	1.3	3.1
Qwen2-VL-7B	66.7	2.8	63.3	100.0	25.0	30.7	66.7	100.0	76.9	69.3	80.0	66.7	80.7	52.1
Qwen2.5-VL-7B	42.9	16.4	36.4	76.9	15.4	40.6	32.4	85.7	34.3	47.9	66.7	25.0	73.9	44.7
InternVL2-8B	33.3	1.5	17.2	100.0	15.4	11.2	6.3	85.7	25.0	54.7	66.7	28.6	68.8	27.7
InternVL2.5-8B	75.0	14.5	32.6	100.0	15.4	27.3	24.4	100.0	76.9	72.5	80.0	50.0	75.5	42.2
Mantis-8B-SigLIP	80.0	28.1	68.6	100.0	28.6	52.5	22.6	100.0	93.3	91.0	80.0	50.0	87.0	58.4
Aya-Vision-8B	46.2	23.8	43.3	100.0	40.0	40.4	21.3	100.0	75.0	77.6	80.0	50.0	29.1	38.6
Idefics3-8B-Llama3	42.9	18.5	19.7	90.9	15.4	23.5	24.4	85.7	33.3	64.1	80.0	28.6	84.6	40.1
Pixtral-12b	36.4	22.9	18.4	90.9	18.2	25.6	21.7	100.0	37.0	58.5	66.7	33.3	78.7	41.1
Gemma-3-12B-it	53.3	36.6	32.2	90.9	25.0	51.6	12.1	100.0	41.0	63.3	66.7	40.0	87.1	52.9
Deepseek-vl2-small	42.9	19.9	54.1	76.9	15.4	30.6	37.6	85.7	34.3	37.5	66.7	28.6	44.0	34.9
Kimi-VL-A3B	46.2	21.2	39.4	76.9	20.0	35.8	23.0	85.7	46.7	66.7	66.7	33.3	85.7	47.7
GPT4o-mini	85.7	67.7	62.5	100.0	40.0	55.0	58.4	100.0	66.7	90.1	80.0	50.0	65.36	64.8

Table 4: **Evaluation of Motion Prediction Baselines.** We evaluate the motion prediction baselines on MMHU-T. We leverage the pre-trained weights and evaluate them on our dataset without fine-tuning. All of the baselines generate plausible trajectory predictions, and PhysMoP achieves the best performance.

	MPJPE↓							
frame_id	1	3	7	9	13	17	21	24
PhysMoP Zhang et al. (2024c)	0.4	1.7	9.0	14.4	26.3	36.2	45.3	54.3
AuxFormer Xu et al. (2023)	17.0	32.7	47.8	60.0	71.1	79.0	84.3	86.1
CIST-GCN Medina et al. (2024)	18.5	25.3	37.2	40.8	46.2	46.6	46.8	47.4

Table 5: **Benefiting Behavior VQA.** By finetuning on MMHU, the baseline model (QWen2.5-VL) shows a significant improvement on averaged accuracy and F1-score.

Model	Accuracy↑	F1-Score ↑
Baseline	35.31	44.72
Finetuned	67.77	68.54

3.5 STATISTICS

As shown in Fig. 3 (a) and Tab. 1, our dataset provides 48 hours human motion sequences and corresponding video clips in total. The frame rate of both the motion sequence and video clip is 10 Hz, thus the total frames sum up to 1.73M. The durations of motion sequences range from 1 to 12 seconds. The mean and median durations are 3.01 and 2 seconds. The average length of low- and high-level descriptions is 390 and 34 words, respectively. The word clouds for the two kinds of text are illustrated in Fig. 3 (c) and (d). We also show the frequency of different behaviors in Fig. 3 (b).

4 TASKS

MMHU supports multiple human-centric tasks. In this section, we will present the definition of tasks and corresponding evaluation metrics in our experiments.

Motion Prediction. Understanding the historical motion and predicting future ones is essential for the safety of autonomous driving. Following PhysMoP (Zhang et al., 2024c), we leverage the sequence of human motion keypoints as the representation. Specifically, human motion involves n frames $M_1, ...M_n$, each of which represents the global location of human joints at frame I_i. Each motion frame M_i consists the location of m key joint point on human body, i.e. $M_i = \{p_k \in \|k \in \{1, ..., m\}\}$. The human motion prediction task is about predicting future motion frames from historical ones, i.e., predicting $M_{t1+1}, ...M_{t_1+t2}$ from $M_0, ..M_{t1}$. We employed two widely used metrics to evaluate the baseline methods: (1) Mean Per Joint Position Error (MPJPE), which measures the mean 3D Euclidean distance between the predicted and ground truth joint positions after aligning the root joint; (2) and the ACCL metric, measuring the acceleration error averaged over time to measure the physical plausibility of the predicted motion.

Motion Generation. Though motion is an informative and compact representation of human actions, collecting them in the real driving scene is expensive and even dangerous for some behaviors. Motion-from-text generation can be a very efficient and effective way to augment motion data. Specifically, given a text description of motion, the algorithm is supposed to generate motion sequences that are subject to the text description. We thus test the capacity of existing current text-to-motion approaches to generate human motions in a driving scene. We follow MotionDiffuse (Zhang et al., 2024a) to model motion sequences by converting from the SMPL parameters. We leverage the high-level

description as the text prompt to generate the motions. We employ FID (Heusel et al., 2017) and multi-modality as evaluation criteria. The former evaluates the distributional distance between the generated motions and the ground-truth motions, while the latter measures joint position differences among 32 motion sequences generated from the same text description.

Behavior VQA. Humans in the street can have multiple characteristics and behaviors. Unlike previous related tasks, such as intention prediction, where the behavior is simply classified into binary labels — crossing the street or not crossing the street, we propose a new behavior set that provides more comprehensive aspects of human behaviors. To make the tasks flexible and easy to extend, we formulate the task as a visual question-answering (VQA) task. We label humans with 13 behaviors, which are all binary labels, as mentioned in Sec. 3.4. We construct closed-ended questions for all labeled behaviors using a template. An example is "Is the person in the video riding a bike? (A) Yes (B) No." We then evaluated the accuracy of the baseline methods. We employed the accuracy and F1-score as evaluation metrics.

5 Experiments

In this section, we evaluate recent methods for human behavior understanding in our dataset. Specifically, we evaluated methods related to human motion prediction, text-to-motion generation, and human behavior VQAs. We analyze their performance on their corresponding tasks.

5.1 Dataset Splitting

MMHU is split into three subsets, namely MMHU-V, MMHU-H, and MMHU-T, each consisting of 47k, 9.5k, and 840 human instances, representing the VLM-labeled, human-labeled, and testing data. We sample the human-labeling subset (which is MMHU-H + MMHU-T) equally from three strategies. (1) in-video-clip sampling: For some video clips, we randomly sample some individuals from each of them as human-labeling data and the rest as MMHU-V data. (2) in-video sampling: we sample some video clips from each raw video and use all the individuals as human-labeling data, and the rest video clips serve as MMHU-V data. (3) out-of-video sampling: we randomly choose some raw videos, and the entire videos serve as the human-labeling dataset. Similarly, we split the MMHU-H and MMHU-T subsets from the human-labeling data. In the experiments, without specifically mentioning, we use both MMHU-V and MMHU-H as training data, denoted as MMHU. We illustrate the three sampling strategy in Fig. 6 (Appendix).

5.2 Baselines

Motion Prediction: For motion prediction, we employ PhysMoP (Zhang et al., 2024c), CIST-GCN (Medina et al., 2024), and AuxFormer (Goncalves & Busso, 2022) as the baselines. Following the settings in (Zhang et al., 2024c), we randomly select 50 continuous frames from each motion sequence, using the first 25 frames as the input for the baselines and comparing their output with the remaining 25 frames. Motion Generation: We choose MotionDiffuse (Zhang et al., 2024a) and motionGPT (Jiang et al., 2023) as our baseline. For both methods, we provided the high-level descriptions from our dataset as the text prompt. Following motionGPT (Jiang et al., 2023), we only use motion sequences that are within 20 to 196 frames. Behavior VQA: For the 13 critical behaviors, there is no specialist model that is trained to predict all of them. Thus, we employed generalist vision-language models as our baselines. We provided the vision language models 4 to 6 frames sampled from the corresponding video and a closed-ended question constructed from the behavior labels as input, and evaluated whether the VLMs can accurately recognize these essential behaviors and answer the provided question. We employ multi-image querying VLMs, including Phi-4-multimodal (Abouelenin et al., 2025), MiniCPM (Yao et al., 2024), Qwen (Wang et al., 2024a; Bai et al., 2025), InternVL (Chen et al., 2024d), Mantis (Jiang et al., 2024a), Idefics3 (Laurençon et al., 2024), Pixtral (Agrawal et al., 2024), Gemma3 (Team et al., 2025a), Deepseek-VL2 (Wu et al., 2024), Kimi-VL (Team et al., 2025b).

5.3 Baseline Evaluation

Motion Generation: As shown in Tab. 2 and Fig. 4 (upper row), motion generation models pre-trained on generic human motion datasets are not capable of generating plausible motions in driving scenes, due to the domain gap between the general motions and motions in the street. Behavior VQA: We evaluated the ability of mainstream vision-language models to recognize human behaviors in driving scenes. We constructed closed-ended questions based on the behaviors and provided 4 to 6 frames to the VLMs. We evaluated the VLMs based on their correctness in answering these behavior-related

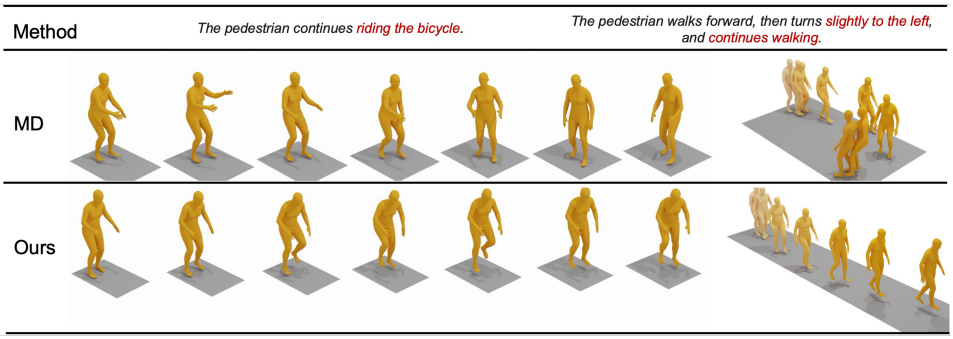


Figure 4: **Qualitative Comparison of Motion Generation.** The baseline model (MotionDiffuse, MD) is not capable of generating proper motions in driving scenes. After fine-tuning on MMHU(second row), the model demonstrates the ability to generate human motions in autonomous driving scenarios.

questions, results shown in Tab. 3. Motion Prediction: We evaluated the pre-trained baselines on MMHU-T, to unify the time argument used in computing the MPJPE metric, we use the frame id to replace it. All baselines are evaluated on the same frames. The results are shown in Tab. 4. All of the approaches generate plausible results, even though they are not specially trained for driving scenes. Among them, PhysMoP (Zhang et al., 2024c) achieves the best performance in our dataset.

5.4 Improving Human Behavior Understanding

In this section, we show that MMHU can facilitate versatile tasks of human behavior understanding. By finetuning different baseline models on our dataset, we observe significant performance gains.

Behavior VQA. We employed QWen2.5-VL (Bai et al., 2023) as our baseline. We fine-tuned the baseline model on our dataset and evaluated both models using the average accuracy and the F1 score on MMHU-T. As shown in Tab. 5, the fine-tuned model achieves a significant performance gain of 32.46% and 23.82% with respect to the average precision and F-1 score.

Motion Prediction. We trained PhysMoP (Zhang et al., 2024c) on the mixed data of 3DPW (von Marcard et al., 2018) and MMHU, and compared it with that trained on 3DPW only. We then evaluated both variants on the original 3DPW dataset, following the original settings of PhysMoP. As shown in Tab. 6, the model training with our MMHU data generalizes to 3DPW and significantly outperforms the baseline model by 9.49 average MPJPE and 1.1 ACCL.

Intention Prediction: Intention prediction answers the question of whether or not a person is going to cross the street, which is a special aspect of behavior QA. We select TrEP (Zhang et al., 2023) as the baseline. We train the baseline model on the JAAD (Kotseruba et al., 2016) dataset, which is a widely used dataset for intention prediction. We then compared the baseline model to that trained on the mixture of JAAD (Kotseruba et al., 2016) and MMHU. From Tab. 7 we observe MMHU significantly contribution to a performance gain of the baseline model. All evaluations are conducted on the JAAD (Jiang et al., 2024a) test set following the original settings of TrEP (Zhang et al., 2023).

Motion Generation. We show that MMHU can narrow the domain gap between generic text-to-motion generation and motions in driving scenes. We fine-tuned MotionDiffuse (Zhang et al., 2024a) on MMHU, and we observe a significant improvement in FID, as shown in Tab. 8. The visualization results in Fig. 4 and the supplementary materials further show the effectiveness of our dataset.

6 Conclusion

In this work, we present **MMHU**, a large-scale dataset for human behavior understanding. The MMHU dataset consists of 57k human instances, each of them are richly annotated with 3D motion sequences, text descriptions, and labeling of 13 critical behaviors. We collected driving videos from various sources and developed a human-in-the-loop annotation pipeline to get high-quality annotations. We evaluated and analyzed the baseline models on the proposed dataset regarding the

Table 6: **Benefiting Motion Prediction.** After adding MMHU to traning set, the baseline model PhysMoP (Zhang et al., 2024c) shows significant performance gain when evaluated on 3DPW (von Marcard et al., 2018).

Train Set	MPJPE-avg↓	ACCL↓
3DPW	47.67	3.8
3DPW+MMHU	38.18	2.7

Table 7: **Benefiting Intention Prediction.** Adding MMHU to the training set of the baseline model TrEP (Zhang et al., 2023), the model achieved more percise intention prediction on the JAAD dataset (Kotseruba et al., 2016).

Train Set	Accuracy ↑	F1-score↑	AuROC↑
JAAD	84.49	84.45	92.98
JAAD+MMHU	91.89	91.89	97.72

Table 8: **Benefiting Motion Generation.** Models fine-tuned on MMHU (labeled with *) can better generate motions in the driving scenes.

Model	FID↓	Multi Modality
Real	0.0020	-
MotionDiffuse	39.27	2.36
MotionDiffuse*	1.86	2.31
MotionGPT	27.06	5.42
MotionGPT*	8.44	3.77

task of motion prediction, text-to-motion generation, and human behavior VQA. We also conduct experiments to show how MMHU can benefit these tasks.

7 Ethics and Reproducibility Statement

Ethics Statement. We follow the previous YouTube dataset (Abu-El-Haija et al., 2016; Xu et al., 2018; Liu et al., 2024c) to protect the privacy of web-sourced videos. All the YouTube videos are with the CC license and are compiled with YouTube's privacy policy and terms of service[1]. We also added mosaics on the copyrighted objects, such as the logo and channel owner information. In addition, we will not provide the video clips directly; instead, we will only provide reference links to the original YouTube videos, which come with all the labels and the bounding boxes to refer to the corresponding individual in the original video. We will provide the processing script to get the corresponding video clips from the video link. For self-collected data, we have removed the identity information by applying mosaics to the faces or license plates, following the approach described in (Grauman et al., 2022). For the VLM-labeled subset, we leveraged the open-source Qwen/Qwen2.5-VL-3B-Instruct model (Bai et al., 2025) for the annotation pipeline. The foundation model is fine-tuned on the human-labeled MMHU-H subset before being used to label the MMHU-V subset. We have followed the Qwen RESEARCH LICENSE AGREEMENT[2] in the data processing. MMHU will be under the CC BY-NC-SA 4.0 license. Additionally, we will implement protective measures on our dataset webpage, including detailed user agreements, encryption protocols for personal data, enforced access restrictions, and monitoring systems to detect misuse or unauthorized access.

Reproducibility Statement. We have provided the details of the dataset in Sec. 3 and Sec. C. The details of the experiments are described in Sec. 4, Sec. 5, and Sec. D. We will release the dataset, with detailed documentation and develop tool kits, upon the acceptence of this paper.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024.

Hidehisa Arai, Keita Miwa, Kento Sasaki, Yu Yamaguchi, Kohei Watanabe, Shunsuke Aoki, and Issei Yamamoto. Covla: Comprehensive vision-language-action dataset for autonomous driving. *arXiv preprint arXiv:2408.10845*, 2024.

[1] https://www.youtube.com/static?template=terms

[2] https://huggingface.co/Qwen/Qwen2.5-VL-3B-Instruct/blob/main/LICENSE

Vishnu Baburajan, João de Abreu e Silva, and Francisco Camara Pereira. Open vs closed-ended questions in attitudinal surveys–comparing, combining, and interpreting using natural language processing. *Transportation research part C: emerging technologies*, 137:103589, 2022.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Vasileios Belagiannis and Andrew Zisserman. Recurrent human pose estimation. In *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*, pp. 468–475. IEEE, 2017.

Apratim Bhattacharyya, Daniel Olmeda Reino, Mario Fritz, and Bernt Schiele. Euro-pvi: Pedestrian vehicle interactions in dense urban centers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6408–6417, 2021.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

Pablo Cervantes, Yusuke Sekikawa, Ikuro Sato, and Koichi Shinoda. Implicit neural representations for variable length human motion generation. In *European Conference on Computer Vision*, pp. 356–372. Springer, 2022.

Guangyi Chen, Xiao Liu, Guangrun Wang, Kun Zhang, Philip HS Torr, Xiao-Ping Zhang, and Yansong Tang. Tem-adapter: Adapting image-text pretraining for video question answer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13945–13955, 2023a.

Haoyu Chen, Jingjing Ren, Jinjin Gu, Hongtao Wu, Xuequan Lu, Haoming Cai, and Lei Zhu. Snow removal in video: A new dataset and a novel method. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023b.

Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14093–14100. IEEE, 2024a.

Weirong Chen, Le Chen, Rui Wang, and Marc Pollefeys. Leap-vo: Long-term effective any point tracking for visual odometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19844–19853, 2024b.

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, 2024c.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024d.

MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. `https://github.com/open-mmlab/mmpose`, 2020.

Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. Posescript: 3d human poses from natural language. In *European Conference on Computer Vision*, pp. 346–362. Springer, 2022.

Junting Dong, Qing Shuai, Yuanqing Zhang, Xian Liu, Xiaowei Zhou, and Hujun Bao. Motion capture from internet videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pp. 210–227. Springer, 2020.

Jianwu Fang, Dingxin Yan, Jiahuan Qiao, Jianru Xue, and Hongkai Yu. Dada: Driver attention prediction in driving accident scenarios. *IEEE transactions on intelligent transportation systems*, 23(6):4959–4971, 2021.

Jianwu Fang, Fan Wang, Jianru Xue, and Tat-Seng Chua. Behavioral intention prediction in driving scenes: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 2024.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

Lucas Goncalves and Carlos Busso. Auxformer: Robust approach to audiovisual emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7357–7361. IEEE, 2022.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Back to mlp: A simple baseline for human motion prediction. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 4809–4819, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Buzhen Huang, Yuan Shu, Tianshu Zhang, and Yangang Wang. Dynamic multi-person mesh recovery from uncalibrated multi-view cameras. In *2021 International Conference on 3D Vision (3DV)*, pp. 710–720. IEEE, 2021.

Yuichi Inoue, Yuki Yada, Kotaro Tanahashi, and Yu Yamaguchi. Nuscenes-mqa: Integrated evaluation of captions and qa for autonomous driving datasets using markup annotations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 930–938, 2024.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024a.

Zhongyu Jiang, Haorui Ji, Cheng-Yen Yang, and Jenq-Neng Hwang. 2d human pose estimation calibration and keypoint visibility classification. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6095–6099. IEEE, 2024b.

Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *European Conference on Computer Vision*, pp. 196–214. Springer, 2020.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. YOLO by Ultralytics, January 2023. URL https://github.com/ultralytics/ultralytics.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Hee Jae Kim, Kathakoli Sengupta, Masaki Kuribayashi, Hernisa Kacorri, and Eshed Ohn-Bar. Text to blind motion. *Advances in Neural Information Processing Systems*, 37:16272–16285, 2024.

Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 563–578, 2018.

Iuliia Kotseruba, Amir Rasouli, and John K Tsotsos. Joint attention in autonomous driving (jaad). *arXiv preprint arXiv:1609.04741*, 2016.

Iuliia Kotseruba, Amir Rasouli, and John K Tsotsos. Benchmark for evaluating pedestrian action prediction. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1258–1268, 2021.

Joon-Young Kwak, Byoung Chul Ko, and Jae-Yeal Nam. Pedestrian intention prediction based on dynamic fuzzy automata for vehicle driving at nighttime. *Infrared Physics & Technology*, 81: 41–51, 2017.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions., 2024.

Jiefeng Li, Siyuan Bian, Chao Xu, Gang Liu, Gang Yu, and Cewu Lu. D &d: Learning human dynamics from dynamic camera. In *European Conference on Computer Vision*, pp. 479–496. Springer, 2022.

Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3d points. In *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*, pp. 9–14. IEEE, 2010.

Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *Proceedings of the IEEE/CVF International conference on computer vision*, pp. 11313–11322, 2021.

Yiheng Li, Cunxin Fan, Chongjian Ge, Zhihao Zhao, Chenran Li, Chenfeng Xu, Huaxiu Yao, Masayoshi Tomizuka, Bolei Zhou, Chen Tang, et al. Womd-reasoning: A large-scale dataset for interaction reasoning in driving. *arXiv preprint arXiv:2407.04281*, 2024.

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, 2023a.

Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36:25268–25280, 2023b.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023b.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024a. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

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, 2024b. URL https://arxiv.org/abs/2412.04468.

Zhizheng Liu, Joe Lin, Wayne Wu, and Bolei Zhou. Learning to generate diverse pedestrian movements from web videos with noisy labels. In *The Thirteenth International Conference on Learning Representations*, 2024c.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6): 248:1–248:16, October 2015.

Zhengxiong Luo, Zhicheng Wang, Yan Huang, Liang Wang, Tieniu Tan, and Erjin Zhou. Rethinking the heatmap regression for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13264–13273, 2021.

Diogo C Luvizon, Marc Habermann, Vladislav Golyanik, Adam Kortylewski, and Christian Theobalt. Scene-aware 3d multi-human motion capture from a single camera. In *Computer Graphics Forum*, volume 42, pp. 371–383. Wiley Online Library, 2023.

Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. Dolphins: Multimodal language model for driving. In *European Conference on Computer Vision*, pp. 403–420. Springer, 2024.

Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000 km: The oxford robotcar dataset. *The International Journal of Robotics Research*, 36(1):3–15, 2017.

Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5442–5451, 2019.

Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. Drama: Joint risk localization and captioning in driving. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1043–1052, 2023.

Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pp. 474–489. Springer, 2020.

Ana-Maria Marcu, Long Chen, Jan Hünermann, 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*, pp. 252–269. Springer, 2024.

Marcin Marszalek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 2929–2936. IEEE, 2009.

Edgar Medina, Leyong Loh, Namrata Gurung, Kyung Hun Oh, and Niels Heller. Context-based interpretable spatio-temporal graph convolutional network for human motion forecasting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3232–3241, 2024.

Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving. In *European Conference on Computer Vision*, pp. 292–308. Springer, 2024.

Juan Carlos Niebles, Chih-Wei Chen, and Li Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *European conference on computer vision*, pp. 392–405. Springer, 2010.

Nada Osman, Guglielmo Camporese, and Lamberto Ballan. Tamformer: Multi-modal transformer with learned attention mask for early intent prediction. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 722–731, 2021.

Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. *arXiv preprint arXiv:2305.14836*, 2023.

Zhijie Qiao, Haowei Li, Zhong Cao, and Henry X Liu. Lightemma: Lightweight end-to-end multimodal model for autonomous driving. *arXiv preprint arXiv:2505.00284*, 2025.

Raúl Quintero, Ignacio Parra, David Fernández Llorca, and MA Sotelo. Pedestrian intention and pose prediction through dynamical models and behaviour classification. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pp. 83–88. IEEE, 2015.

Nishant Rai, Haofeng Chen, Jingwei Ji, Rishi Desai, Kazuki Kozuka, Shun Ishizaka, Ehsan Adeli, and Juan Carlos Niebles. Home action genome: Cooperative compositional action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11184–11193, 2021.

Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Agreeing to cross: How drivers and pedestrians communicate. In *IEEE Intelligent Vehicles Symposium (IV)*, pp. 264–269, 2017a.

Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 206–213, 2017b.

Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John K Tsotsos. Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6262–6271, 2019.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), November 2017.

Saeed Saadatnejad, Yang Gao, Kaouther Messaoud, and Alexandre Alahi. Social-transmotion: Promptable human trajectory prediction. In *The Twelfth International Conference on Learning Representations*.

Enna Sachdeva, Nakul Agarwal, Suhas Chundi, Sean Roelofs, Jiachen Li, Mykel Kochenderfer, Chiho Choi, and Behzad Dariush. Rank2tell: A multimodal driving dataset for joint importance ranking and reasoning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 7513–7522, 2024.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992, 2018.

Nicolas Schneider and Dariu M Gavrila. Pedestrian path prediction with recursive bayesian filters: A comparative study. In *german conference on pattern recognition*, pp. 174–183. Springer, 2013.

Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L Waslander, Yu Liu, and Hongsheng 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.

Neha Sharma, Chhavi Dhiman, and S Indu. Visual–motion–interaction-guided pedestrian intention prediction framework. *IEEE Sensors Journal*, 23(22):27540–27548, 2023.

Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2070–2080, 2024.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European Conference on Computer Vision*, pp. 256–274. Springer, 2024.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2446–2454, 2020.

Hiromu Taketsugu, Takeru Oba, Takahiro Maeda, Shohei Nobuhara, and Norimichi Ukita. Physical plausibility-aware trajectory prediction via locomotion embodiment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12324–12334, 2025.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025a.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025b.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. Label Studio: Data labeling software, 2020-2025. URL `https://github.com/HumanSignal/label-studio`. Open source software available from https://github.com/HumanSignal/label-studio.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Rejin Varghese and M Sambath. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pp. 1–6. IEEE, 2024.

Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, sep 2018.

Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 601–617, 2018.

Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 1290–1297. IEEE, 2012.

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. *arXiv preprint arXiv:2409.12191*, 2024a.

Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M Alvarez. Omnidrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning. *arXiv preprint arXiv:2405.01533*, 2024b.

Xinshun Wang, Qiongjie Cui, Chen Chen, and Mengyuan Liu. Gcnext: Towards the unity of graph convolutions for human motion prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5642–5650, 2024c.

Yichen Wang, Yiyi Zhang, Xinhao Hu, Li Niu, Jianfu Zhang, Yasushi Makihara, Yasushi Yagi, Pai Peng, Wenlong Liao, Tao He, et al. Pedestrian motion reconstruction: A large-scale benchmark via mixed reality rendering with multiple perspectives and modalities. In *The Thirteenth International Conference on Learning Representations*.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

Chen Xie, Ciyun Lin, Xiaoyu Zheng, Bowen Gong, Dayong Wu, and Antonio M López. Gtranspdm: A graph-embedded transformer with positional decoupling for pedestrian crossing intention prediction. *arXiv preprint arXiv:2409.20223*, 2024.

Chenxin Xu, Robby T Tan, Yuhong Tan, Siheng Chen, Xinchao Wang, and Yanfeng Wang. Auxiliary tasks benefit 3d skeleton-based human motion prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9509–9520, 2023.

Liang Xu, Shaoyang Hua, Zili Lin, Yifan Liu, Feipeng Ma, Yichao Yan, Xin Jin, Xiaokang Yang, and Wenjun Zeng. Motionbank: A large-scale video motion benchmark with disentangled rule-based annotations. *arXiv preprint arXiv:2410.13790*, 2024.

Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018.

Dongfang Yang, Haolin Zhang, Ekim Yurtsever, Keith A Redmill, and Ümit Özgüner. Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention. *IEEE Transactions on Intelligent Vehicles*, 7(2):221–230, 2022.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21222–21232, 2023.

Jianhao Yuan, Shuyang Sun, Daniel Omeiza, Bo Zhao, Paul Newman, Lars Kunze, and Matthew Gadd. Rag-driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model. *arXiv preprint arXiv:2402.10828*, 2024.

Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11038–11049, 2022.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE transactions on pattern analysis and machine intelligence*, 46(6):4115–4128, 2024a.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024b. URL https://arxiv.org/abs/2410.02713.

Yufei Zhang, Jeffrey O Kephart, and Qiang Ji. Incorporating physics principles for precise human motion prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6164–6174, 2024c.

Zhengming Zhang, Renran Tian, and Zhengming Ding. Trep: Transformer-based evidential prediction for pedestrian intention with uncertainty. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 3534–3542, 2023.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics. URL http://arxiv.org/abs/2403.13372.

A Appendix Overview

We introduce some background and preliminaries in Sec. B. Then we describe the details of the dataset and its construction in Sec. C. In the last, we describe the details of the experiments we conduct in Sec. D.

B Background and Preliminaries

B.1 The Skinned Multi-Person Linear (SMPL) Representation.

The SMPL model (Loper et al., 2015) is a parametric human body representation. It represents the human body as a triangulated surface and uses linear blend skinning (LBS) to control both pose and body shape. A template mesh in a canonical rest pose is defined as $\mathcal{M}_h = (\mathcal{V}, \mathcal{F})$, where $\mathcal{V} \in \mathbb{R}^{n_v \times 3}$ is the set of n_v vertex positions and $\mathcal{F}$ the set of faces. Given shape parameters β and pose parameters θ, SMPL first produces a shaped mesh by adding learned offsets to the template vertices: $\mathcal{V}_s = \mathcal{V} + B_S(\beta) + B_P(\theta)$, where $B_S(\beta) \in \mathbb{R}^{n_v \times 3}$ are the shape-dependent displacements and $B_P(\theta) \in \mathbb{R}^{n_v \times 3}$ are the pose-dependent displacements, both defined in xyz space. The vertices $\mathcal{V}_s$ describe the body in the shaped space. To obtain the final posed mesh corresponding to a target pose θ', SMPL applies LBS to $\mathcal{V}_s$. It uses fixed skinning weights $W \in \mathbb{R}^{n_k \times n_v}$ and a set of joint transformations $G = \{G_k\}_{k=1}^{n_k}$. Each vertex v_i is transformed according to $v_i' = \left(\sum_{k=1}^{n_k} W_{k,i} G_k\right) v_i$, where n_k is the number of skeletal joints. The transformations G_k are functions of the source pose θ, the target pose θ', and the shape parameters β. The pose vector is usually split into a body pose term $\theta_b \in \mathbb{R}^{23 \times 3 \times 3}$ and a global orientation term $\theta_g \in \mathbb{R}^{3 \times 3}$. The official SMPL repository[3] extends SMPL with global orientation and translation, represented by $\Gamma_t = \{\boldsymbol{r}_t, \boldsymbol{\pi}_t\}$, where $r_t \in \mathbb{R}^3$ and $\pi_t \in \mathbb{R}^3$ denote the root orientation and translation with respect to the camera.

B.2 The Closed- and Open- Ended Questions.

In the VQA task, the term "closed-ended" question generally means the questions with limited possible answers. In contrast, the "open-ended" question refers to a question that allows the respondent to generate any free-form answer. For example, multiple-choice questions are closed-ended questions. And the question asking "What do you think are the paper's strengths and weaknesses?" is an open-ended question. The further explanation of the two terms can be found in (Baburajan et al., 2022).

C Dataset Details

C.1 Data Visualization

We show some examples from our dataset in Fig. 5. In each example, we show the sampled video frames in the first row, the human instance is highlighted using a red bounding box. The corresponding poses are visualized under each frame in the second row, followed by the text description and the behavior labels.

C.2 Data Collection

For the YouTube source, we filtered and downloaded videos with the *Creative Commons Attribution license (reuse allowed)* license using yt-dlp[4]. We also filtered the videos by checking whether they contain the keywords "drive" or "driving" in their title. We collect video clips that sum to over 11 hours, consisting of 10k human instances. We list the details of the used YouTube channels in Tab. 9.

[3] https://github.com/vchoutas/smplx

[4] https://github.com/yt-dlp/yt-dlp

Table 9: **Collected YouTube Raw Data**

Channel Name	License	Collected Minutes	Human Instances
TravelRelaxListen	Creative Commons Attribution	179	2028
planetearthtraveler	Creative Commons Attribution	92	3122
RoamingBrit	Creative Commons Attribution	65	497
Evan-Explores	Creative Commons Attribution	147	62
VietnamSilentRoutes	Creative Commons Attribution	187	4812

C.3 Extracting the Video Clip for Each Human Instance

After roughly cutting the raw video into fragments following Sec. 3.1, the next step is to detect each human instance and extract the corresponding video frames. Following (Shin et al., 2024), we apply the human detector Varghese & Sambath (2024); Jocher et al. (2023) on each frame of the video fragment. For each detected human in each frame, we leverage mmpose Contributors (2020) for 2D keypoint extraction and tracking. Then we collect the frames to form the video clip of each human instance. The Waymo Sun et al. (2020) dataset provides human bounding box labels for each frame. In this case, we leverage the ground-truth human bounding box instead of the detected ones to make it more precise. However, these human instances are labeled based on the 3D point cloud, which may be invisible to the current video point of view. Thus, we drop the bounding boxes that do not intersect with any of the detected bounding boxes with at least 0.2 IoU. In the remaining bounding boxes, if any two of them intersect with each other with at least 0.2 IoU, we drop both of them.

C.4 Details of the Human Annotation for the Behavior Labels

We present the details of the human annotators for the MMHU-H and MMHU-T subsets. We employ 12 annotators. Each annotator is provided with 6 frames of the video clip. The target human is cropped and labeled with a bounding box. We asked the annotators to check for each behavior that the human holds. We pay the annotator 0.15 CNY per video clip, leading to roughly 50 CNY per hour. We use label studio (Tkachenko et al., 2020-2025) to build the annotation environment. The English version of the annotation interface is shown in Fig. 7.

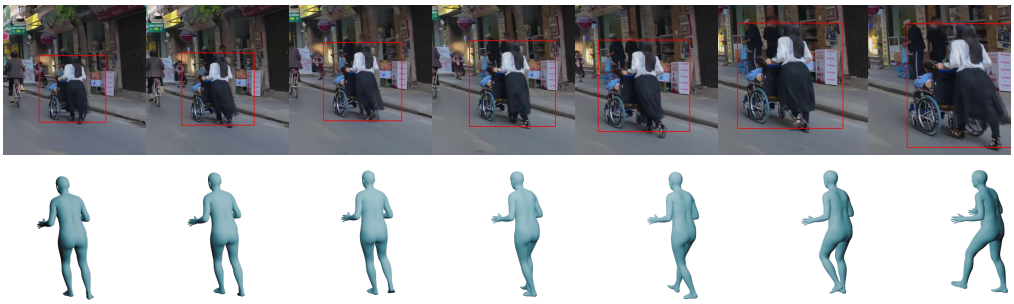
C.5 Details of Hierarchical Text Annotation

To bridge the gap between parameterized SMPL motion and semantic-level understanding, we implement a hierarchical text annotation pipeline that translates joint-level motion into structured language descriptions.

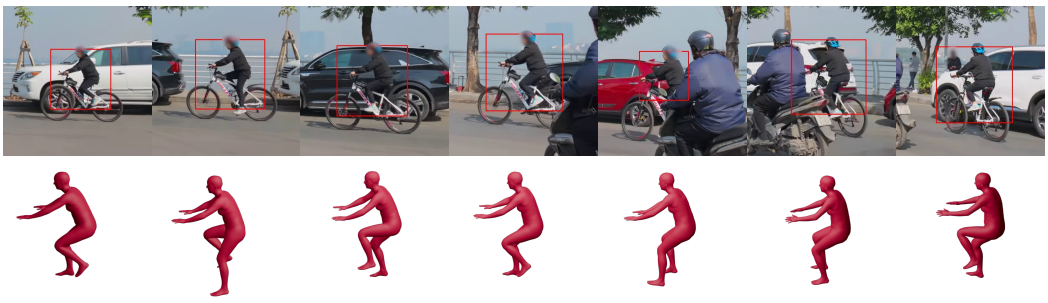
We use PoseScript (Delmas et al., 2022) to extract joint-wise behavioral descriptions (the low-level description) from SMPL pose sequences. Unlike the original method, which outputs a single paragraph per frame summarizing the full-body posture, our approach decomposes captions into joint-specific phrases by isolating text segments corresponding to individual joints. We focus on the most salient joint movements and retain only those joints mentioned in at least 50% of the frames within a sequence. To ensure temporal consistency, we align the low-level descriptions of each selected joint across the sequence. These temporally aligned joint descriptions are aggregated using a large language model to produce concise summaries of each joint's motion over time , as shown in Fig. 8. The low-level descriptions are utilized to support the subsequent aggregation of high-level semantic behavioral descriptions. We show an example of the low-level (joint-wise) descriptions in Box 1. Then, a vision-language model is employed to aggregate the low-level (joint-wise) descriptions with the high-level semantic behavior information from the keyframes. The instructions for this procedure are demonstrated in Box 2.

The pedestrian continues running, maintaining a steady pace with arms swinging and legs alternating in a rhythmic motion.

The pedestrian walks across the street, maintaining a steady pace with arms slightly raised and hands close together. They continue to walk past a police vehicle and a taxi, eventually stopping near a traffic light. [Crossing the Street, Carrying Items By hand]



The pedestrian continues pushing the wheelchair forward, maintaining a steady pace and consistent posture. [Wheelchair, Carrying Items by Hand]



The cyclist maintains a steady pace, with hands gripping the handlebars and knees bent, navigating through traffic. [Bicycle]

Figure 5: **Visualization of the MMHU dataset.** For each human instance, the first line shows the video frames that are sampled from the video clips. The human instance is highlighted using a red bounding box. We crop and zoom in on the human instance for a clearer view. Under each of the frames shows the corresponding mesh-rendered human motion, followed by the text description for the human motion and the behavior labels.

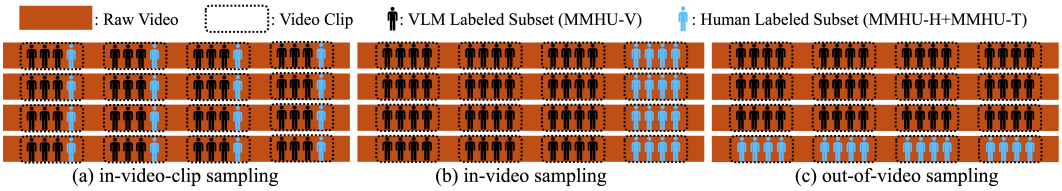


Figure 6: **Example of the Sampling Strategies:** We show an example of the three sampling strategies mentioned in Sec. 5.1. In the example, the ratio of VLM-labeled data to the Human-labeled data is assumed to be 3:1.

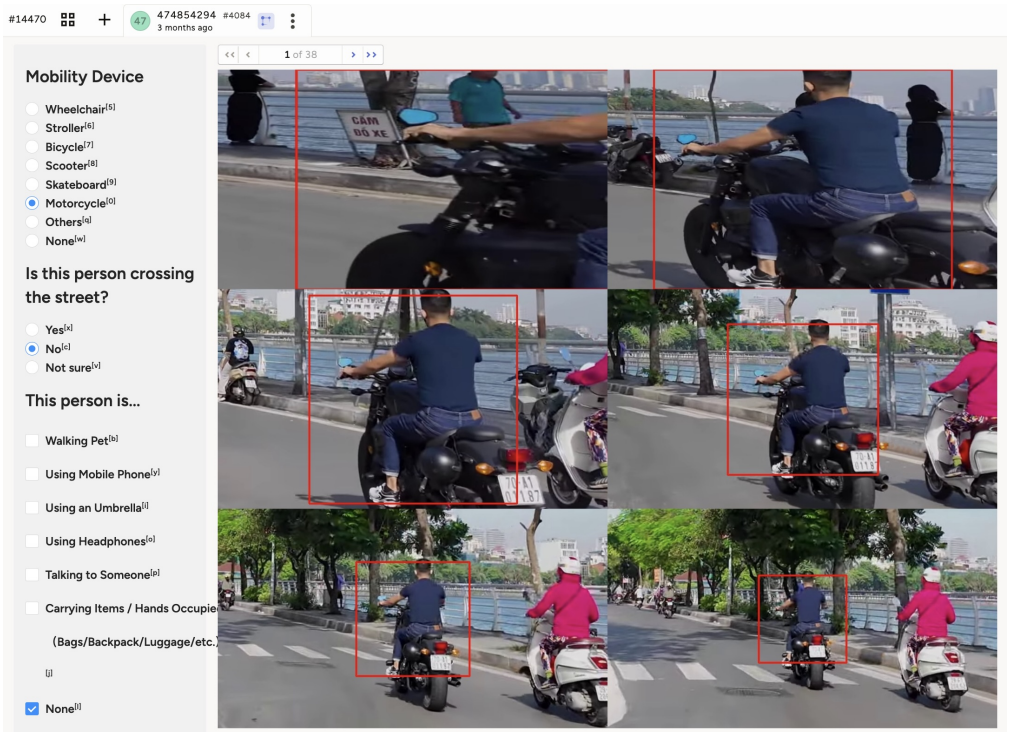


Figure 7: **Example of Annotation Interface.**

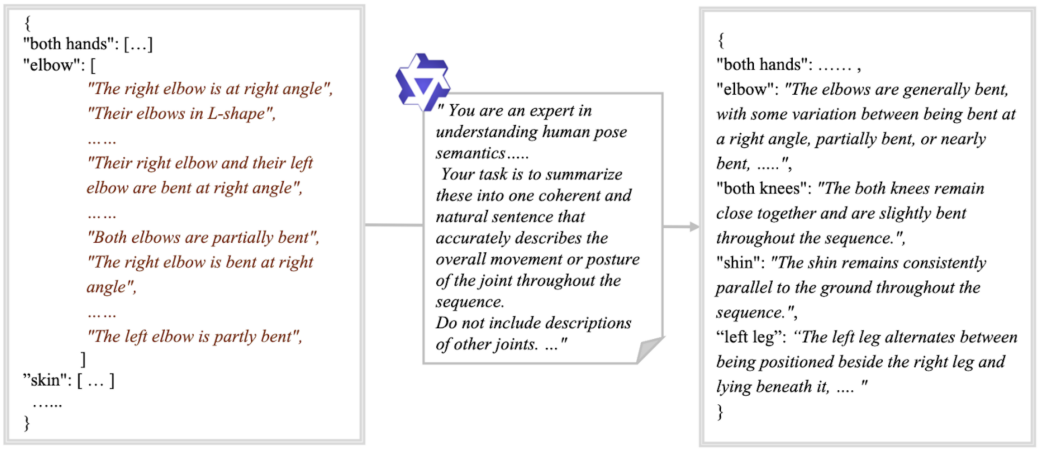


Figure 8: **Example of Low-level Text Aggregation.** Left: per-frame text description for human parts at each frame; Right: aggregated low-level text description.

Box 1. Example of Low-Level Descriptions

Selected Key Frames for Reference:

Low-Level (Joint-wise) Descriptions:

- **Elbow:** The elbows are generally bent throughout the sequence, with some variation in the degree of bending.
- **Forearm:** The forearm remains mostly horizontal with occasional alignment with the thighs and shins, indicating a stable position with some minor adjustments.
- **Knee:** The knees are generally bent with slight variations, occasionally straightening or separating at shoulder width, indicating a dynamic posture.
- **Left Elbow:** The left elbow remains mostly bent throughout the sequence, with occasional variations indicating slight changes in its position.
- ...

Box 2. Instruction Template for Aggregated Text Description

System Prompt: You are an expert in human motion analysis and natural language refinement.
Your task is to reorganize and clarify human pose descriptions derived from SMPL-based 3D models. Ensure all descriptions are structured, concise, and consistent, while strictly preserving the original semantics—do not add, remove, or alter key motion elements.
When provided with low-level pose information (e.g., joint-based summaries or posecodes), use it to enhance the clarity or specificity of the description, especially regarding posture and limb motion. However, maintain overall cohesion and naturalness, with a clear focus on the pedestrian's observable behavior in context.

User Prompt: You are given a sequence of images showing a pedestrian in a green bounding box, captured by a front-facing car camera, along with a dictionary of low-level joint-based pose descriptions.
Your task is to:

- Write a concise, one- to two-sentence summary capturing the pedestrian's overall motion and any key changes in action throughout the sequence.
- Emphasize how the pedestrian's behavior evolves over time (e.g., walking, stopping, turning, crouching, changing direction).
- Use the low-level pose descriptions to refine details when relevant, but avoid excessive mechanical phrasing or redundancy.
- Describe only what is clearly supported by the visual and pose data — do not speculate.

Format your final response using the following tags: `<description>` human pose description here. `<description_end>`
Example: `<description>` The pedestrian stepped forward with arms raised, paused briefly, then lowered their body and turned slightly to the left. `<description_end>`

Here is the low-level pose description: {Input Low-Level Description}

Box 3. Example of Human Behavior VQA: Carrying Items

Frames:

System Prompt: You are an expert pedestrian behavior labeler, specializing in analyzing pedestrians' behavior on the road. You will perform visual analysis on multiple sequential images over time.

Question 1: The images show a pedestrian in a bounding box with surrounding context. The image is a cropped frame from a car's front camera.
Is the pedestrian carrying items?
Please answer y or n, only one letter.

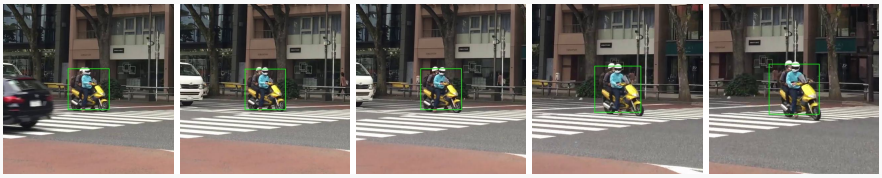
Answer: `y`

Question 2: The images show a pedestrian in a bounding box with surrounding context. The image is a cropped frame from a car's front camera.
The pedestrian is not carrying items. Is this statement correct?
Please answer y or n, only one letter.

Answer: `n`

Box 4. Example of Human Behavior VQA: Using Scooter

Frames:



System Prompt: You are an expert pedestrian behavior labeler, specializing in analyzing pedestrians' behavior on the road. You will perform visual analysis on multiple sequential images over time.

Question 1: The images show a pedestrian in a bounding box with surrounding context. The image is a cropped frame from a car's front camera.
Is the pedestrian using a scooter?
Please answer y or n, only one letter.

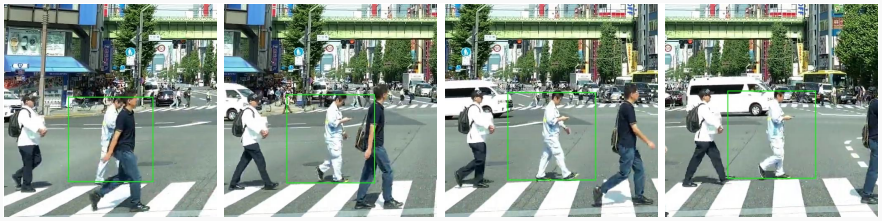
Answer: `y`

Question 2: The images show a pedestrian in a bounding box with surrounding context. The image is a cropped frame from a car's front camera.
The pedestrian is not using a scooter. Is this statement correct?
Please answer y or n, only one letter.

Answer: `n`

Box 5. Example of Human Behavior VQA: Using a Phone

Frames:



System Prompt: You are an expert pedestrian behavior labeler, specializing in analyzing pedestrians' behavior on the road. You will perform visual analysis on multiple sequential images over time.

Question 1: The images show a pedestrian in a bounding box with surrounding context. The image is a cropped frame from a car's front camera.
Is the pedestrian using a phone?
Please answer y or n, only one letter.

Answer: `This person is crossing the street.`

[RollBack] Question 1: The images show a pedestrian in a bounding box with surrounding context. The image is a cropped frame from a car's front camera.
Is the pedestrian using a phone?
Please answer y or n, only one letter.

[RollBack] Answer: `y`

Question 2: The images show a pedestrian in a bounding box with surrounding context. The image is a cropped frame from a car's front camera.
The pedestrian is not using a phone. Is this statement correct?
Please answer y or n, only one letter.

Answer: `n`

D EXPERIMENTS

D.1 HUMAN BEHAVIOR VQA DETAILS.

For each behavior, we construct closed-ended questions based on the template. We provided 4 to 6 frames to the VLMs. The frames are uniformly sampled from the video clips. To reduce randomness, we will ask the VLMs twice, one direct question and one counter question, respectively. The VLM should answer both questions correctly in order to be evaluated as correct for the behavior. We show two examples in Box 3 and Box 4 (both answered correctly). We find that some VLMs have difficulty following the instructions; thus, if the answer has the wrong format, we will roll back at most three times and repeat the same question. An example is shown in Box 5.

Table 10: **Generalizability on Motion Prediction.** The baseline model trained on MMHUonly produces plausible results and generalizes to the 3DPW (von Marcard et al., 2018) testset.

TrainSet	TestSet	MPJPE-avg↓	ACCL
3DPW	3DPW	47.67	3.8
MMHU	3DPW	49.89	2.4

D.2 GENERALIZING ON MOTION PREDICTION

We train the PhysMoP (Zhang et al., 2024c) model on MMHUalone, without any training on the original training set. We then evaluate the model on the original test set of 3DPW (von Marcard et al.,

2018). The results are listed in Tab. 10, from which we observe that even trained on MMHUalone, the baseline model still produces plausible results, demonstrating the generalizability of MMHU.

D.3 Trajectory Recovery

We employed two recent trajectory recovery approaches (Shen et al., 2024; Chen et al., 2024b) and evaluated them on the MMHU-T subset. As shown in Tab. 11, the recent approaches all produce plausible results on MMHU-T, indicating the correctness of the human trajectory in the dataset.

D.4 Additional Results on Motion Generation

Table 11: **Evaluation of Trajectory Recovery.**

Method	MPJPE(mm)	PA-MPJPE(mm)
Shen et al. (Shen et al., 2024)	24.24	45.44
LEAP-VO (Chen et al., 2024b)	27.86	42.93

To better show the performance improvement of motion generation tasks after finetuning on MMHU dataset, we provide some videos of generated motions along with the supplementary materials. We provided three extra cases to show that the fine-tuned model is more capable of generating human motion in the driving scene. The video is named following "{casenumber}-{Baseline/Finetune}-{behavior_label}-{Mesh/Skeleton}". The first case shows that the baseline is not capable of generating feasible motion of riding a bike. In the second case, the baseline model only captured the "walking" keyword and ignores the behavior of "talking to someone", while the fine-tuned model can generate a more plausible motion with hand postures. In the Third case, we add a more detailed description of human pose, and the fine-tuned model shows the ability to follow the more detailed instructions.

D.5 Benefiting End-to-End Driving Model

We show that understanding human behavior can be beneficial to end-to-end autonomous driving models. We employ LightEmma (Qiao et al., 2025) as a representative. As shown in Fig. 9, the end-to-end driving model produces more plausible future trajectory after being given additional specific human behavior in the prompt, or generally asked to pay extra attention to the human behavior.

D.6 Validation of Human Motion Labels

We show the correctness of our human motion labels by comparing the SMPL keypoints extracted by our pipeline with the ground-truth human keypoints provided in the official Waymo toolkit[5]. Due to the difference in human keypoint representation between the Waymo (9 points) and the SMPL (24 points), we conduct visual validation on the rendered keypoints. Specifically, we back-project and render the GT keypoints and our SMPL keypoints to the original image frames. As shown in Fig. 10. The SMPL keypoints in MMHUare generally aligned with the GT keypoints and the human body in the image.

D.7 Setting Details in Finetuning Models

We provide the details of the fine-tuning procedure used in Sec. 5.4 of the main paper.

Motion Prediction. The training data consists of the entire 3DPW dataset (training set) (von Marcard et al., 2018) and 30k randomly selected human sequences from MMHU. All training settings follow the default configurations described in these papers. Since the frame rate of 3DPW is 25 FPS, while ours is 10 FPS, we upsample all sequences in our dataset to 50 FPS and then downsample them to 25 FPS. Training is conducted on an NVIDIA RTX 4090 GPU.

Motion Generation. We select the test set from our MMHU dataset and split it into training, validation, and test subsets with a 7:1:2 ratio for this experiment. The main training settings and model architecture follow the default configurations in MotionDiffuse (Zhang et al., 2024a). We use the Adam optimizer with a learning rate of 0.0002. For fine-tuning, we set the batch size to 192 and train for 20 epochs using a single NVIDIA RTX 6000 Ada GPU.

[5]https://github.com/waymo-research/waymo-open-dataset/blob/master/tutorial/tutorial_keypoints.ipynb

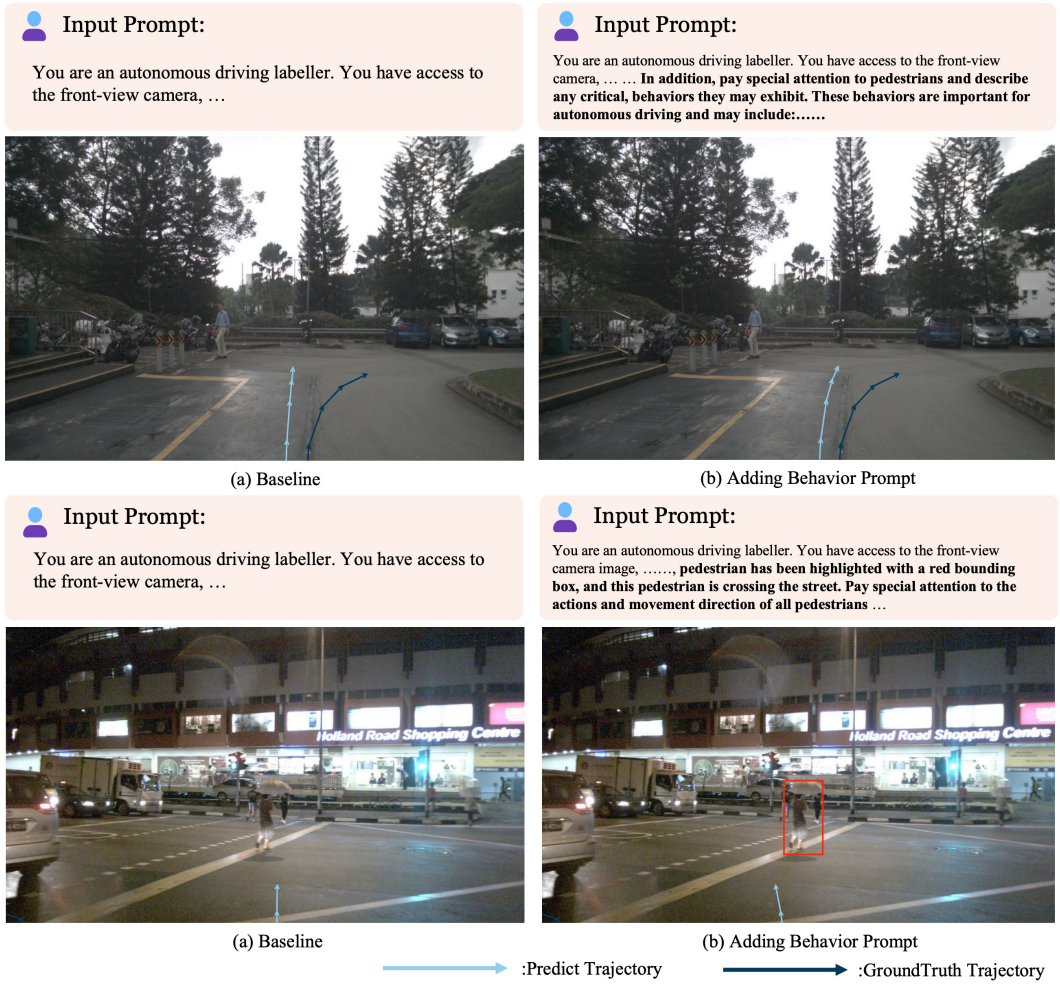


Figure 9: **Benefiting End-to-End Driving Models.** LightEmma (Qiao et al., 2025) generates a more plausible future trajectory when additional human behavior information is added to the text prompt.

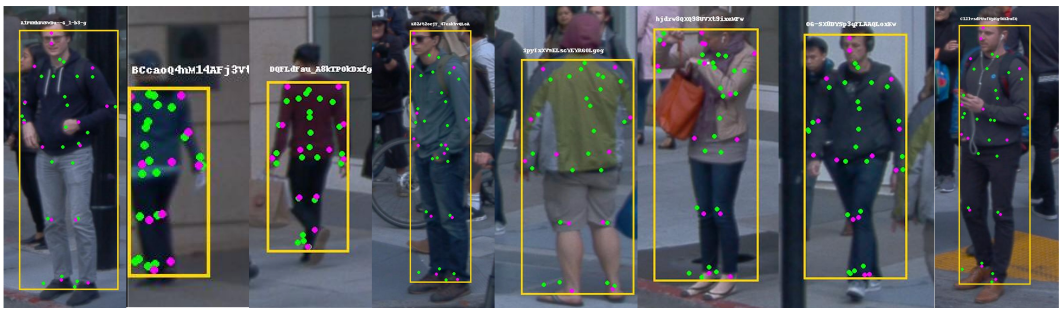


Figure 10: **Validating Motion Labels.** We render the ground-truth keypoints and the SMPL keypoints provided in MMHUtogether with the original human image. The SMPL keypoints are aligned with both the ground-truth keypoints and the human body.

Human Behavior VQA. We apply LoRA (Hu et al., 2022) fine-tuning to Qwen2.5-3B-Instruct by using LLaMA-Factory Framework (Zheng et al., 2024). The visual branch is frozen, and LoRA is applied to all other MLP layers of the model. Our LoRA settings are: lora_rank = 8 and lora_alpha = 16. The model is trained for one epoch on a subset of the MMHU dataset and evaluated on the MMHU-T dataset. Due to class imbalance, we employed resampling to ensure the same training data

from each behavior class. The training is conducted on 4 NVIDIA RTX 6000 Ada GPUs with a total batch size of 128. The learning rate is set to 1e-4, with a warmup ratio of 0.1, and the learning rate scheduler is CosineAnnealingLR. All input images are cropped to 256×256 around the bounding box, or resized to 256×256 if the bounding box is larger than 256×256.

Intention Prediction. The training data consists of the JAAD dataset and MMHU training set. All training settings primarily follow the default configurations in Trep (Zhang et al., 2023) on the JAAD dataset. All models are implemented using the Adam optimizer with a learning rate of 0.005. We adopt an early stopping strategy with a patience of 5, and experiments are conducted on a single NVIDIA RTX 6000 Ada GPU.

E THE USE OF LARGE LANGUAGE MODELS (LLMS)

Polishing Paper Writing. We leverage ChatGPT-4o (Achiam et al., 2023) to help paper writing. Specifically, we provided draft sentences to the LLM, and asked it for advice regarding the word choice and sentence structure.