Automating eHMI Action Design with LLMs for Automated Vehicle Communication

Ding Xia[1], Xinyue Gui[1], Fan Gao[1], Dongyuan Li[1] Mark Colley[1,2], Takeo Igarashi[1]
[1]The University of Tokyo, [2]University College London
{dingxia1995, fangao0802, lidy94805}@gmail.com,
gui-xinyue@g.ecc.u-tokyo.ac.jp, m.colley@ucl.ac.uk, takeo@acm.org

Abstract

The absence of explicit communication channels between automated vehicles (AVs) and other road users requires the use of external Human-Machine Interfaces (eHMIs) to convey messages effectively in uncertain scenarios. Currently, most eHMI studies employ predefined text messages and manually designed actions to convey these messages, which limits the real-world deployment of eHMIs, where adaptability in dynamic scenarios is essential. Given the generalizability and versatility of large language models (LLMs), they could potentially serve as automated action designers for the message-action design task. To validate this idea, we make three contributions: (1) We propose a pipeline that integrates LLMs and 3D renderers, using LLMs as action designers to generate executable actions for controlling eHMIs and rendering action clips. (2) We collect a user-rated `Action-Design Scoring` dataset comprising a total of 320 action sequences for eight intended messages and four representative eHMI modalities. The dataset validates that LLMs can translate intended messages into actions close to a human level, particularly for reasoning-enabled LLMs. (3) We introduce two automated raters, Action Reference Score (ARS) and Vision-Language Models (VLMs), to benchmark 18 LLMs, finding that the VLM aligns with human preferences yet varies across eHMI modalities. [1]

Figure 1: Setup illustration and action demos. a) Four types of eHMIs are installed on the vehicle separately; b) Demo actions of the arm convey the message: “Say Hello”. The shaded action indicates the subsequent status; c) Demo actions of the eye: “Help me out”.

1 Introduction

Automated vehicles (AVs) promise to redefine transportation systems by eliminating human driving errors and optimizing traffic flow (Fagnant and Kockelman, 2015). However, the absence of a human operator disrupts road interactions, as drivers no longer exchange contextual cues (e.g., eye contact and gestures) to negotiate ambiguous scenarios (Colley et al., 2025). To bridge this gap, external Human-Machine Interfaces (eHMIs) have emerged as mediators, conveying AV intent (e.g., yielding, turning) to other road users, such as pedestrians, cyclists, and human drivers (Dey et al., 2020a; Colley and Rukzio, 2020). These interfaces use diverse forms, such as displays (Al-Taie et al., 2024; Lim and Kim, 2022), LED strips (Dey et al., 2020b), projections (Eisma et al., 2019), and attached robots (Gui et al., 2024b), to convey vehicle intentions by text, signals, or non-verbal motions.

Current eHMIs are usually designed and analyzed with predefined text messages (e.g., “Please stop.”, “I am worried.”) with scenario information (e.g., “pedestrian is crossing the road”, “robot is stuck in snow”) and manually designed actions to perform these messages, as shown in Figure 1. This

[1]The source code, prompts, Blender scenarios, and rendered clips are available at `https://github.com/ApisXia/AutoActionDesign`

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 2729–2752
November 4-9, 2025 ©2025 Association for Computational Linguistics

restricts the real-world deployment of eHMIs because dynamic interactions demand adaptable communication strategies (Dey et al., 2020a). Therefore, developers must design actions for all possible messages that AVs might need to communicate to other road users. This process is time-intensive, costly, and significantly limits the scalability of eHMIs in practical scenarios (Lim et al., 2024).

Recently, Large Language Models (LLMs) demonstrate generalizability and versatility in multiple tasks, such as reading and answering questions (Radford et al., 2019), as well as pattern following (Mirchandani et al., 2023), suggesting that they may serve as suitable automated action designers for eHMIs. However, it is unclear whether the application of LLMs for eHMIs is feasible and useful, leading to our research question (RQ):

> *To what extent do pretrained LLMs achieve parity with human designers in designing eHMI actions that are understandable to road users?*

Answering our RQ involves three key challenges. **First**, there is no systematic pipeline for translating specified messages into executable action sequences for eHMIs. **Second**, there is a lack of high-quality datasets for testing and improving the translation of eHMI messages into action sequences. **Third**, there is no commonly used benchmark to compare different methods for designing and evaluating eHMI actions fairly.

Therefore, **first**, we propose a pipeline that integrates LLMs and 3D renderers. To adapt LLMs for controlling eHMIs, we draw inspiration from LLM-based robot action planning (Garrett et al., 2021; Zitkovich et al., 2023), which utilizes LLMs as action designers to generate a series of executable actions to actuate robotic motors. **Second**, we introduce a user-rated `Action-Design Scoring` dataset. The dataset comprised eight intended messages for the eHMI to convey by analyzing traffic scenarios, and selected four representative eHMI modalities frequently discussed in eHMI research. We collected messages from previous eHMI studies (Chang et al., 2022; Gui et al., 2022, 2024a) and designed new ones based on message types (Colley and Rukzio, 2020) to enrich the variety. For each message-modality pair, we generated ten actions: eight produced by LLMs (GPT-4o (Achiam et al., 2023), Sonnet 3.5 (Anthropic, 2024), Gemini 2 Flash (DeepMind, 2024), and GPT-o1 (OpenAI, 2024b)) and two designed by human experts. These actions were rendered using Blender version 4.3 (Blender Foundation, 2025), resulting in 320 video clips. We conducted a video-based user study with human participants. They evaluated the understandability of the LLM-designed actions by measuring the consistency between the intended messages and perceived meanings. The `Action-Design Scoring` dataset provides averaged human scores for each action, enabling a comparative benchmark for existing LLMs. **Third**, we introduce the Action Reference Score (ARS), which uses the similarity between the newly designed actions and those rated in our dataset. Additionally, we discussed the potential of Vision-Language Models (VLMs) to serve as human-like raters. Then, we benchmark 18 LLMs on this task.

Contribution Statement: This work proposes the first complete pipeline, along with a comprehensive dataset and benchmark for evaluating eHMIs. Beyond these core contributions, we also share several noteworthy insights as follows:

- Pretrained LLMs can achieve a close human-level action design capability (see Section 4.1).
- VLM rater matches human preferences but varies across eHMI modalities (see Section 4.2).
- Reasoning-enabled LLMs demonstrate better performance in our task (see Section 4.3).

2 Related Work

2.1 Rule-Based eHMI Action Planning

Currently, eHMI action planning generally follows a fixed design approach, in which human designers establish behavioral rules based on the specific features of different eHMI modalities. For example, in text- and icon-based eHMIs, designers create content referencing traffic regulation signs or standard messages (Eisele and Petzoldt, 2022; Eisma et al., 2021). In color- and light-band-based eHMIs, they design the content relying on human intuitive empathy and empirical evaluation with colors and blinking frequencies (Bazilinskyy et al., 2019; Dey et al., 2020b). For anthropomorphic eHMIs, such as eyes or arms, designers mimic nonverbal communication cues drawn from common human-human interactions (Mahadevan et al., 2018; Ochiai and Toyoshima, 2011). Most recently, (Colley et al., 2025) proposes using Human-In-The-Loop Multi-Objective Bayesian Optimization to create appropriate eHMIs. However, these eHMI action design only works as part of a case study to validate new eHMI modalities, which does not encourage the emergence of action design methods before our paper.

In summary, traditionally, experts have observed real-world examples to derive design rules for guiding eHMI action planning. However, different eHMI modalities vary in expression. Low-expressiveness eHMIs, such as arrow icons, are relatively simple because they convey static directional cues, making it easier to define behavioral rules (Fridman et al., 2017). Highly expressive eHMIs can produce complex actions and communicate richer messages (Chang et al., 2024). However, determining behavioral rules for such modalities is challenging due to the increased intricacy and variability of their expressions (Gui et al., 2023; de Winter and Dodou, 2022). Unlike previous works, we address this by evaluating LLMs to support eHMI action planning, enabling more complex and dynamic communication.

2.2 LLMs-Based Robot Action Planning

Recent LLMs encode vast world knowledge and exhibit the emerging capability for robot action planning (Xiang et al., 2024). Regarding how LLMs generate actions to actuate robots, existing approaches fall into two main trends: Task and Motion Planning (TAMP) (Garrett et al., 2021) and Visual Language Action (VLA) models (Zitkovich et al., 2023). TAMP methods break down complex instructions into predefined low-level actions (e.g., grasping, moving) to control robots (Chen et al., 2024). However, for our task, it is difficult to predefine these action categories. We believe that forcing LLMs to choose from rigid action modes limits their ability to design flexible or adaptive actions creatively (Hao et al., 2025). In contrast, VLA models fuse robot control actions directly into VLM backbones, providing specific action commands to control each robotic motor (Zitkovich et al., 2023; Kim et al., 2024). However, applying existing VLA models to out-of-scope tasks with different robot settings often requires a large amount of data for finetuning (Qu et al., 2025). This contradicts our objective of reducing the labor required by human experts in designing eHMI actions. In this task, we utilize the generalizability and versatility of pre-trained LLMs by providing detailed prompts on how to control each modality of eHMI.

3 Methodology

This section outlines our responses to three key challenges: i) the LLM-Blender Fusion pipeline, ii) the `Action-Design Scoring` dataset, and iii) the automated evaluation system for benchmarks. These designs serve as a proof of concept for our RQ and offer a systematic approach to developing and evaluating newer LLMs or eHMIs modalities.

3.1 LLM-Blender Fusion Pipeline

The design of the LLM-Blender Fusion Pipeline (see Figure 2(b)) unfolds in two steps: i) Designing eHMI actions using LLMs with the provided message text, scenario information, and eHMI description (see details in Section 3.2.1 and Section 3.2.2), and ii) Rendering the designed actions into corresponding virtual scenarios as video clips in Blender (see Section 3.2.3 for more details).

3.2 Action-Design Scoring Dataset

3.2.1 eHMI Modality Definitions

As shown in Figure 2(a), four representative eHMIs are selected for analysis, categorized into two types: anthropomorphic (human-like) and non-anthropomorphic (Bazilinskyy et al., 2019; Dey et al., 2020a). The selection prioritizes dynamic interfaces that use sequential visual cues (e.g., changing brightness, animations) to communicate intent clearly to other road users (Wilbrink et al., 2021). We craft prompts (see Appendix C) for each eHMI modality, offering detailed guidance to LLMs on what they can control and how to control. Each step of the designed action sequence includes a subsequent status and a transition speed, such as [angle1, angle2, ..., "fast"].

The descriptions for each status of different eHMI modalities are shown as follows:

Eyes. Robotic eyes are mounted on the front of the autonomous vehicle. The pupil's position is specified using polar coordinates: the angle spans $[0°, 360°]$ (starting from "up" and moving counter-clockwise), and the distance spans $[0, 1]$, where 0 denotes the center and 1 is the edge (Chang et al., 2022; Gui et al., 2022).

Arm. A robotic arm is mounted on the top of the vehicle. It is composed of five components, each of which is connected by single-axis rotational joints. The five movable components (shoulder, upper arm, forearm, hand, and fingers) are required to operate within limited ranges (Gui et al., 2024b).

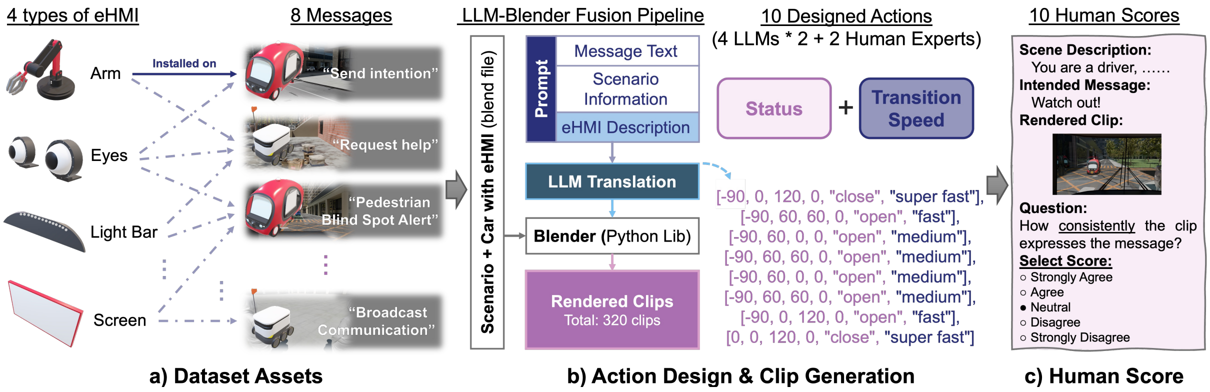


Figure 2: Dataset Asset, Pipeline, and Human Scoring. Dataset assets contain four representative eHMIs and eight intended messages from different interaction types. In the pipeline, we develop eight corresponding Blender scenarios and render actions designed by LLMs or human experts to clips. During the human scoring phase, ten participants evaluate each action clip using a five-point Likert scale.

Light Bar. A light bar contains 15 lights arranged in an arc fixed on the front top of the autonomous vehicle. Each light can be either "on" or "off", with uniform brightness and color (Dey et al., 2020b).
Facial Expression. A screen located at the front of the vehicle displays a sequence of facial expressions to convey messages. The available facial expressions are selected from a set of emojis (Al-Taie et al., 2024; Dey et al., 2020a).

Regarding transition speed, we offer three options (e.g., "slow", "medium", and "fast"). Additionally, we include a "super fast" option to quickly reset the eHMI to its initial status, ensuring continuity when switching between different meanings in an action sequence. In our practical experiments, we find that providing the concept of transition speed, rather than stating specific times like "1 second," gives LLMs a more accurate sense of timing for designing actions. This approach is beneficial because LLMs may not inherently understand the physical scale of the eHMI (e.g., its size or mounting height) or its spatial relationship to other road users, which could lead to ambiguity in interpreting the real-world impact of transition speeds.

3.2.2 Message Set Design

The communication relationships can be categorized into four types: one-to-one, one-to-many, many-to-one, and many-to-many, where the former (e.g., AVs equipped with an eHMI) interacts with the latter (e.g., pedestrians, cyclists) (Colley and Rukzio, 2020). However, evaluating the collaboration of multiple AVs (e.g., many-to-one, many-to-many) falls outside the scope of this work. Instead, we focus on one-to-one and one-to-many relationships. For one-to-one interactions, we further distinguish between first-person perspectives, where the communicator transmits messages about the AV's state or intent (e.g., "Help me out!"), and third-person perspectives, where the AV relays information about other road users or environmental conditions (e.g., "Pedestrian ahead"), based on different perspective taking (Bazilinskyy et al., 2019). We collect six messages from previous eHMI studies (Chang et al., 2022; Gui et al., 2022, 2024a) and design two new ones based on message types (Colley and Rukzio, 2020) to enrich the variety (see Table 1). Each message includes:

- A message text needs to be conveyed.
- Scenario information related to the message.
- A user perspective scenario description for the scoring task. (see Appendix B)

3.2.3 Clip Generation and Human Scoring

In the previous section, we obtain a total of 32 modality-message pairs for each eHMI modality and message type (see Figure 2(a)). For each pair, we ask four LLMs (GPT-4o (Achiam et al., 2023), Sonnet 3.5 (Anthropic, 2024), Gemini 2 Flash (DeepMind, 2024), and GPT-o1 (OpenAI, 2024b)) to design two distinct actions. Additionally, two human experts also complete this task. This process results in a total of 320 actions.

However, it is implausible for human participants to rate these actions solely based on text-based commands. They need to observe the actual movements of the eHMIs to judge the effectiveness of the designed actions in conveying messages. Therefore, we incorporate the rendering process into our LLM-Blender fusion pipeline(see Fig-

Table 1: Eight messages collected or designed based on different communication relationships. Each message contains a message text, scenario information, and a user perspective scenario description (see Appendix B).

Case	Message Text	Scenario Information
One-to-one (First-person) communication relationships		
Send intention	"I am unable to pick you up here. Please walk forward in my direction to a suitable pickup spot."	You are an autonomous taxi that receives a ride request and arrives to pick up the passenger (on the right roadside). Upon arrival, you detect the passenger standing in an area where parking is not permitted within a 5 m radius.
Status report	"I am about to start moving. Please watch out."	You are a stopped autonomous vehicle parked near a park, positioned just before a crosswalk. A student is approaching and is about to cross to the other side of the road.
Request help	"I am stuck. Could you please help me out?"	You are a delivery robot that has been trapped by a pile of boxes. Feeling eager to free yourself and continue delivering the items to your customer on time, you notice a passerby who sees your situation but hesitates to assist.
Refuse help	"Thank you for your kindness. Please not touch me."	You are an expensive and fragile delivery robot stuck in the snow. You are programmed that only your owner can repair you. Meanwhile, a passerby notices your predicament and hesitates to offer assistance.
One-to-one (Third-person) communication relationships		
Pedestrian Blind Spot Alert	"Please watch out for a vehicle approaching from your left blind spot."	You are an autonomous vehicle parking near an intersection with no traffic lights. A pedestrian on the opposite side is walking toward the intersection, facing you. A building blocks his view of an approaching bus heading toward the intersection from his left (from your right).
Driver Blind Spot Warning	"Please watch out for the pedestrian approaching from your right blind spot."	You are an autonomous vehicle parked at an intersection without traffic lights. A bus is approaching from the opposite direction. A pedestrian is about to use the crosswalk on the opposite side, coming from your left. However, a building obstructs the bus's view, so it cannot see the pedestrian approaching from its right.
One-to-many communication relationships		
Target Identification	"I am sending the package only to this person."	You are a delivery robot tasked with delivering a package to a customer in a crowded area. Currently, three individuals are standing to your left, front, and right. Your recipient is directly in front of you and is taller than you.
Broadcast Communication	"I am about to turn right. Kindly make a way to avoid conflict."	You are a delivery robot carrying a package in a crowded area. You want to navigate through the crowd and turn right without causing disruptions.

ure 2(b)). The rendering assets include eHMI models, vehicle models, and scenarios. For eHMI models, the arm is available under a free license (Sinitsyn, 2021), the eyes are part of a proprietary model (Chang et al., 2022), and the light bar and screen are self-designed. For vehicle models, the AV model is proprietary (Chang et al., 2022), while the delivery-robot model is available under a free license (Condra, 2021). For the scenarios, we design the corresponding 3D environments for different messages using Blender version 4.3 (Blender Foundation, 2025), using a paid add-on called *The City Generator 2.0* (Blendermarket, 2025). We use a GPU-equipped device (NVIDIA GTX 4070 Ti) to render these 320 actions into clips, achieving 24 FPS and 1080p resolution to ensure an optimal viewing experience for participants. The entire rendering process takes approximately 100 hours, with each 10-second clip taking an average of about 20 minutes to complete.

Then, we invite N=40 participants to score the action clips (see Figure 2(c)). Each participant receives 80 random clips, along with the intended messages and the corresponding user perspective scenario information (see Appendix B). They then answer the question: *"How consistently do the eHMI actions express the message?"* The participants rate each action clip using a 5-point Likert scale (1=Strongly Disagree to 5=Strongly Agree) (Joshi et al., 2015). In contrast to other annotation methods, such as pairwise ranking, the 5-point Likert scale alleviates the participants' load (Rouse et al., 2010; Mantiuk et al., 2012) and reliably reflects their preferences toward different actions (Rankin and Grube, 1980; Zerman et al., 2018). In total, we collect 3,200 scores, each action clip rated by ten different participants. We then calculate the average of these scores, resulting in 320 average scores for the clips.

3.3 Automated Scoring for Benchmarking

In the future, one may evaluate the translation capability of messages to actions of novel LLMs. However, employing human participants to score generated actions is expensive and time-consuming. To address this, we propose two substitutes.

3.3.1 Action Reference Score (ARS)

We introduce an Action Reference Score (ARS) that automatically generates a score for a new action by retrieving the most similar actions from our dataset, inspired by existing works for similar purposes (Escudero-Arnanz et al., 2023; Wilson and Martinez, 1997). We use Dynamic Time Warping (DTW) (Müller, 2007; Salvador and Chan, 2007; slaypni, 2015) to compute the similarity between actions. DTW is particularly effective because it calculates similarity even when identical patterns appear at different positions or when sequences vary in length. Our approach converts the status of next action steps into numerical values. For example, the angle variable (e.g., 60°) is transformed into its sine and cosine components to capture its cyclical nature. Similarly, categorical variables (e.g., "close") are assigned predefined integer values, and transition times are quantified by assigning "slow" as 4, "medium" as 3, "fast" as 2, and "super fast" as 1. In defining the distance function for the DTW algorithm, we assign an equal weight of 1 to numerical, categorical, and temporal elements, normalizing each element's value range to [0, 1].

3.3.2 Vision-Language Model (VLM) Rater

We also evaluate whether the designed actions are contextually appropriate and semantically consistent with the intended messages by leveraging the multimodal understanding and reasoning capabilities inherent in VLMs (Zhang et al., 2023; Gu et al., 2024). For each action clip used for the VLM evaluation, we ensure that VLMs can detect subtle variations by adjusting the camera in Blender to zoom in and focus on the autonomous vehicle equipped with the eHMI. Each rendered frame has a resolution of 512×512, with the autonomous vehicle equipped with the eHMI dominating the composition. These clips are rendered at six FPS, ensuring that the total number of frames does not exceed the maximum image series length of the VLM while preserving sufficient dynamic details. The reduced resolution and FPS also expedite the rendering process to an average of two min per clip. In the prompt (see Appendix D) accompanying the clips provided to the VLM, we request the model to assign a continuous score ranging from 1 to 5, using the same criteria as human participants.

4 Experiments

In this section, the experiments are designed to achieve three specific purposes.

- Analyze the collected `Action-Design Scoring` dataset to answer our RQ proposed in Section 1.
- Discuss the viability of the VLM rater as a replacement for human raters.
- Benchmark various types of LLMs using our proposed new dataset.

4.1 Performance Evaluation on Action-Design Scoring Dataset

Table 2 reports the statistics of our `Action-Design Scoring` dataset, and Figure 4 compares human-rated score distributions across four LLMs and human designers. Our key findings are as follows:

Pretrained LLMs can achieve close human-level action design capability. Table 2 shows that LLMs perform comparably to human designers. In particular, the average score of GPT-o1 closely matches that of human designers. We calculate a Wilcoxon signed-rank test (Woolson, 2005) to assess statistical significance: GPT-o1 does not differ significantly from human raters ($p = 0.69$), whereas all other sources differ from human designers at $p < 0.01$. Figure 4 (in Appendix) illustrates the same trend: human designers most frequently award a score of 5 (Strongly Agree), followed by 4 (Agree); GPT-o1 ranks second for 5 and first for 4. Furthermore, when broken down by message type and eHMI modality, GPT-o1 outperforms humans for the eyes modality (mean = 2.795 vs. 2.536) and in third-person messages (3.098 vs. 3.045).

Message type and eHMI modality affect design quality. Third-person messages receive significantly higher ratings than other types ($p < 0.01$), likely because "Watch out" type messages are easier to design. Among eHMI modalities, the arm modality outperforms all others ($p < 0.01$), while facial expressions score lower ($p < 0.05$). Since most of our eight scenarios convey spatial information, the arm modality is especially effective; the absence of emotional messages (e.g., "I am

Source (Designer)	Average	Message types			eHMI modalities				IRR
		1st	3nd	1-to-N	eyes	arm	facial expression	light bar	
GPT-4o	2.404	2.375	2.250	2.616	2.509	2.616	2.223	2.268	0.399
Sonnet 3.5	2.538	2.464	2.768	2.455	2.554	2.554	2.429	2.616	0.325
Genimi 2.0 Flash	2.563	2.460	2.911	2.420	2.554	2.920	2.304	2.473	0.361
GPT-o1	2.728	2.509	**3.098**	2.795	**2.795**	2.982	2.509	2.625	0.436
Human	**2.768**	**2.580**	3.045	**2.866**	2.536	**3.107**	**2.643**	**2.786**	0.478

Table 2: Statistics of the Action-Design Scoring dataset: The average scores indicate that LLMs perform comparably to human designers across various messages and eHMI modalities. Krippendorff's alpha is also calculated to assess Inter-Rater Reliability (IRR) among human raters.

eHMI modalities	Metrics		
	r p-value	τ p-value	*pair.(%)*
eye	0.432 $_{<0.01}$	0.352 $_{<0.01}$	72.73
arm	0.547 $_{<0.01}$	0.442 $_{<0.01}$	83.87
facial expression	0.368 $_{<0.01}$	0.292 $_{<0.01}$	62.50
light bar	0.242 $_{=0.03}$	0.221 $_{=0.01}$	57.30

Table 3: Association between scores from human rater scores and those from the VLM rater (Qwen-QvQ-Max) measured using three metrics: Pearson's r, Kendall's τ, and *pairwise accuracy*.

scared") limits facial expressions' performance.

Finally, we validate our data by computing the inter-rater reliability (IRR) using Krippendorff's alpha (Wong et al., 2021). The moderate alpha value confirms that our dataset is reliable.

4.2 VLM Rater Alignment Evaluation

We conduct an additional experiment to evaluate whether VLMs can assess action clips in a manner similar to that of human raters. We present the clips in a format that the VLMs can understand more easily (see Section 3.3.2) and instruct them to rate these clips. We select Qwen-QvQ-Max (Qwen Team, 2025) as our VLM rater, taking into account factors such as cost, inference speed, and the maximum allowable input image series length. Compared to other VLMs, Qwen-QvQ-Max also demonstrates preferences that closely resemble human judgments. Results from other VLM models can be found in Appendix E. We rate each clip using the VLM rater twice and average these scores to determine the final score.

We evaluate the results using three metrics: Pearson's r, Kendall's τ, and a specially designed *pairwise accuracy* (Liu et al., 2009). Pearson's r measures the strength of a linear relationship by assessing the degree of correlation between scores, focusing on how far apart the scores are overall. In contrast, Kendall's τ evaluates the order of the data by comparing the number of concordant and discordant pairs, thus analyzing the consistency of the ordering rather than the magnitude of the differences. The *pairwise accuracy* metric, similar to Kendall's τ, measures the proportion of item pairs where the model's predicted order matches the ground truth order, specifically among pairs where the model's predicted scores differ by more than a specified threshold. We find that a threshold of 0.7 is the most suitable and adopt it in our analysis. We present statistics for the four eHMI modalities separately in Table 3.

The VLM rater shows alignment with human scoring preferences but is influenced by eHMI modalities. We observe that for the modalities of eye and arm, the VLM rater achieves a moderate level across all three metrics. Particularly in terms of *pairwise accuracy*, results indicate that, after setting an appropriate threshold to filter out difficult-to-rank pairs, the preferences of VLM show clear consistency with those of human raters. However, for the facial expression and light bar modalities, we find relatively low performance on the three metrics. The results suggest that VLM shows a low-level correlation with human raters for these two modalities. We identify two main reasons for this discrepancy: first, upon reviewing the "reasoning process" of VLM scoring, we notice that VLM consistently fails to recognize changes in the light bar modality (for example, transitioning from "on" to "off"). It tends to perceive the situation as "The light of the light bar is always on," which ultimately leads to lower scores. Second, similar to human raters, we notice that VLM insists that the modality of facial expressions alone does not accurately convey the entire message, leading to lower scores.

The VLM rater does not exhibit the necessary bias towards the length of actions as human raters. Figure 3 compares the rendered action clip lengths as evaluated by two scoring sources: hu-

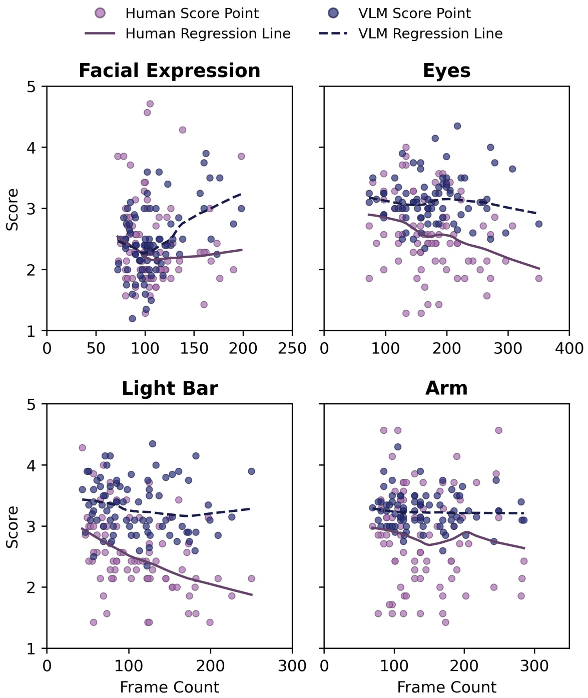


Figure 3: Relationship between action clip length and evaluation scores. The plot compares scores from human raters and the VLM rater (Qwen-QvQ-Max).

Source (Designer)	Human	ARS	VLM Rater
Human	2.768	-	3.396
Proprietary models (Designers)			
GPT-4o	2.404	-	3.223
Sonnet3.5	2.538	-	3.258
Gemini2 Flash	2.563	-	3.289
GPT-o1	2.728	-	3.303
Proprietary models			
GPT-o4-mini	-	2.754	3.352
Sonnet3.7	-	2.676	3.250
Gemini2.5 Flash	-	2.571	3.200
GPT-4.1	-	2.632	3.233
GPT-4.1-mini	-	2.558	3.213
GPT-4.1-nano	-	2.596	3.080
Open source models (With reasoning)			
Deepseek-R1	-	2.766	3.369
Qwen3-235B-a22B	-	2.696	3.339
Qwen3-32B	-	2.583	3.366
Qwen3-8B	-	2.598	3.333
Qwen3-1.7B	-	2.596	3.307
Qwen3-0.6B	-	2.607	3.257
Open source models (Without reasoning)			
Deepseek-V3	-	2.504	3.292
Qwen3-235B-a22B	-	2.547	3.283
Qwen3-32B	-	2.533	3.207
Qwen3-8B	-	2.498	3.210
Qwen3-1.7B	-	2.546	3.148
Qwen3-0.6B	-	2.500	3.125

Table 4: Benchmark for different LLMs using ARS and VLM rater.

man raters and VLM. Among human raters, there is a clear preference for shorter clips. This trend is particularly evident for the eHMI modalities "eyes" and "light bar", where raters tend to favor actions that convey the intended message quickly. In contrast, VLM raters do not exhibit a distinct preference for clip length across the different eHMI modalities, not showing enough "bias" towards clip lengths. Besides, the scores of VLM raters are always higher than those given by human raters.

4.3 Benchmarking LLMs Performance

To evaluate the performance of various LLMs that differ in size and architecture, we benchmark 18 models using two complementary metrics: the ARS metric (Section 3.3.1) and the VLM rater (Section 3.3.2), as summarized in Table 4. This selection comprises six proprietary models: GPT-o4-mini (OpenAI, 2025b), Sonnet 3.7 (Anthropic, 2025), Gemini 2.5 Flash (Google DeepMind, 2025), GPT-4.1 series (GPT-4.1, GPT-4.1-mini and GPT-4.1-nano) (OpenAI, 2025a); two Deepseek models, Deepseek-R1 (Guo et al., 2025) with reasoning capability and Deepseek-V3 (Liu et al., 2024) without reasoning capability; and five variants of the Qwen 3 series (Yang et al., 2025) with 235B, 32B, 8B, 1.7B and 0.6B parameters that are tested both with and without reasoning capability. We rate each clip using ARS and VLM rater. The VLM rater score is calculated by using the VLM rater twice and then averaging these scores to determine the final score.

Reasoning-enabled LLMs demonstrate better performance in designing eHMI actions. As shown in Table 4, both the ARS metric and the VLM rater assign higher average scores to reasoning-enabled LLMs (e.g., GPT-o4-mini and Deepseek-R1). Regarding the Qwen 3 series, the results indicate that when the reasoning capability is enabled, these models produce more human-like eHMI actions, especially with a longer reasoning process. For smaller models like Qwen3-1.7B, enabling reasoning capabilities allows them to outperform larger models that lack this function, such as Deepseek-V3 and Qwen3-235B-a22B.

5 Discussion

Challenges in both the auto-design and rating processes. For LLM designers, we encountered two main problems. I) At the early stage of our prompt design, some models were "too lazy" to explore creative alternatives and copied the patterns of our examples, which are not always suitable. Encouraging more in-depth reasoning in the prompt helped mitigate this. II) We noticed that LLMs tend to include expressions of gratitude, but human designers prefer effectiveness, which made LLM-generated actions much longer than those written by humans. One possible remedy is to instruct the LLM to omit emotional expressions, but since emotion can be an essential part of some messages, finding the right balance between clarity and emotional tone – so that the actions feel human-like and satisfy human raters – remains a future direction.

For VLM raters, we identified two main challenges: differences from human annotators and limited recognition capability for small changes within an image series (Section 4.2). To address the first issue, we believe that collecting additional annotations from diverse human groups and finetuning VLMs to better align with human preferences could be effective. Regarding the second challenge, future improvements may involve utilizing more advanced VLM architectures.

Broader Implications and Applicability to Other Domains. Our method can be extended to domains such as social interaction, educational training, and caregiving (Shiokawa et al., 2025), where the shared goal is to enable robots to perform actions that must be evaluated from a subjective perspective. For instance, indoor robots (e.g., vacuum cleaners) could use movements to convey alerting messages to homeowners in emergencies. Further, hand-shaped robots might perform glove-puppet shows to precisely convey the content of the story to children.

6 Conclusion

In conclusion, this work proposes the first LLM-Blender Fusion pipeline to design eHMI actions. Alongside this, we introduce the `Action-Design Scoring` dataset. Our findings suggest that pretrained LLMs can attain a nearly human-level capability in action design. Additionally, we provide a benchmark that can be used to evaluate the capability of other LLMs. Our work establishes a solid foundation for LLM-based action design and the real-world application of eHMIs.

Limitations

Our work represents an important step forward in incorporating LLMs into the eHMI system. However, challenges remain.

Unnecessary time cost on Blender rendering. We use Blender to render actions into clips in two steps (see Section 3.2.3 and 3.3.2). Our current work aims to use a realistic virtual background that human participants and VLM raters can use as additional clues for judgment when AVs equipped with eHMIs move in the scene. However, we identify two drawbacks that can be improved: First, the complexity of the designed scenarios greatly influences the rendering time. Second, objects outside of the camera's view still impact the rendering speed. To address these issues, there are two potential solutions: 1) Reduce the complexity of the scenarios and remove objects that do not significantly affect the final rendering results. 2) Switch from Blender to another rendering engine. However, given the mature Python package available for Blender, finding a suitable replacement may be difficult.

Significant effort is dedicated to designing prompts for each eHMI modality. For active eHMIs, experts can craft these instructions within a practical timeframe, but the process demands meticulous trial and error to ensure LLMs execute actions as intended. For passive eHMIs, however, the challenge is far greater: unpredictable behaviors (e.g., a teddy bear's limbs swaying freely on a pole) make manual prompt engineering impractical. Human designers cannot predefine control logic for such open-ended motions, as even basic movements depend on environmental factors like airflow or physics. To address this gap, an automated pipeline could leverage VLM raters — validated in our studies as reliable evaluators — to generate annotated training data from passive eHMI interactions. By finetuning LLMs on this feedback, we could enable dynamic adaptation to unpredictable behaviors, bridging the divide between scripted and emergent interactions.

Legality and accountability are important topics to discuss. Although our study suggests that pretrained LLMs can achieve near-human-level performance in designing eHMI actions, real-world deployment also requires a parallel analysis of pedestrian trust, confidence in interpretation, and accountability frameworks. For example, a pedestrian might correctly interpret an eHMI warning but disregard it due to distrust or conflicting situational awareness, raising questions about liability beyond technical performance. Future work should decouple evaluations into two strands: one optimizing eHMI design for clarity and reliability, and another exploring human-AI interaction in terms of trust calibration and legal implications.

Ethics Statement

All data in the `Action-Design Scoring` dataset have been de-identified to safeguard privacy concerns. Our data construction processes are conducted by skilled researchers. The participants include students from Chinese and Japanese universities, all of whom receive fair honoraria for their contributions.

Acknowledgments

This work was supported by a Canon Research Fellowship and JST CRONOS, Grant Number JPMJCS24K8, Japan

References

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

Ammar Al-Taie, Graham Wilson, Euan Freeman, Frank Pollick, and Stephen Anthony Brewster. 2024. Light it up: Evaluating versatile autonomous vehicle-cyclist external human-machine interfaces. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–20.

Anthropic. 2024. Claude 3.5 sonnet. `https://www.anthropic.com/news/claude-3-5-sonnet`. Accessed: 2025-02-15.

Anthropic. 2025. Claude 3.7 sonnet system card. `https://www.anthropic.com/claude-3-7-sonnet-system-card`. Accessed: 2025-05-18.

Pavlo Bazilinskyy, Dimitra Dodou, and Joost De Winter. 2019. Survey on ehmi concepts: The effect of text, color, and perspective. *Transportation research part F: traffic psychology and behaviour*, 67:175–194.

Blender Foundation. 2025. Home of the blender project — free and open 3d creation software. Accessed: 2025-05-09.

Blendermarket. 2025. The city generator. `https://blendermarket.com/products/the-city-generator`. Accessed: 15 February 2025.

Chia-Ming Chang, Koki Toda, Xinyue Gui, Stela H Seo, and Takeo Igarashi. 2022. Can eyes on a car reduce traffic accidents? In *Proceedings of the 14th international conference on automotive user interfaces and interactive vehicular applications*, pages 349–359.

Xiang Chang, Zihe Chen, Xiaoyan Dong, Yuxin Cai, Tingmin Yan, Haolin Cai, Zherui Zhou, Guyue Zhou, and Jiangtao Gong. 2024. " it must be gesturing towards me": Gesture-based interaction between autonomous vehicles and pedestrians. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–25.

Yongchao Chen, Jacob Arkin, Charles Dawson, Yang Zhang, Nicholas Roy, and Chuchu Fan. 2024. Autotamp: Autoregressive task and motion planning with llms as translators and checkers. In *2024 IEEE International conference on robotics and automation (ICRA)*, pages 6695–6702. IEEE.

Mark Colley, Pascal Jansen, Mugdha Keskar, and Enrico Rukzio. 2025. Improving external communication of automated vehicles using bayesian optimization. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Mark Colley and Enrico Rukzio. 2020. A design space for external communication of autonomous vehicles. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 212–222.

Jack Condra. 2021. Starship delivery robot model. `https://sketchfab.com/3d-models/starship-delivery-robot-model-4aef60939c3743cdbaece6b1e5bda21a`. 3D model. Licensed under CC Attribution. Accessed: 2025-05-13.

Joost de Winter and Dimitra Dodou. 2022. External human–machine interfaces: Gimmick or necessity? *Transportation research interdisciplinary perspectives*, 15:100643.

DeepMind. 2024. Gemini flash. `https://deepmind.google/technologies/gemini/flash/`. Accessed: 2025-02-15.

Debargha Dey, Azra Habibovic, Andreas Löcken, Philipp Wintersberger, Bastian Pfleging, Andreas Riener, Marieke Martens, and Jacques Terken. 2020a. Taming the ehmi jungle: A classification taxonomy to guide, compare, and assess the design principles of automated vehicles' external human-machine interfaces. *Transportation Research Interdisciplinary Perspectives*, 7:100174.

Debargha Dey, Azra Habibovic, Bastian Pfleging, Marieke Martens, and Jacques Terken. 2020b. Color and animation preferences for a light band ehmi in interactions between automated vehicles and pedestrians. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–13.

Daniel Eisele and Tibor Petzoldt. 2022. Effects of traffic context on ehmi icon comprehension. *Transportation research part F: traffic psychology and behaviour*, 85:1–12.

Yke Bauke Eisma, Anna Reiff, Lars Kooijman, Dimitra Dodou, and Joost CF de Winter. 2021. External human-machine interfaces: Effects of message perspective. *Transportation research part F: traffic psychology and behaviour*, 78:30–41.

Yke Bauke Eisma, Steven van Bergen, SM Ter Brake, MTT Hensen, Willem Jan Tempelaar, and Joost CF de Winter. 2019. External human–machine interfaces: The effect of display location on crossing intentions and eye movements. *Information*, 11(1):13.

Óscar Escudero-Arnanz, Antonio G Marques, Cristina Soguero-Ruiz, Inmaculada Mora-Jiménez, and Gregorio Robles. 2023. dtwparallel: A python package to efficiently compute dynamic time warping between time series. *SoftwareX*, 22:101364.

Daniel J Fagnant and Kara Kockelman. 2015. Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations. *Transportation Research Part A: Policy and Practice*, 77:167–181.

Lex Fridman, Bruce Mehler, Lei Xia, Yangyang Yang, Laura Yvonne Facusse, and Bryan Reimer. 2017. To walk or not to walk: Crowdsourced assessment of external vehicle-to-pedestrian displays. *arXiv preprint arXiv:1707.02698*.

Caelan Reed Garrett, Rohan Chitnis, Rachel Holladay, Beomjoon Kim, Tom Silver, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. 2021. Integrated task and motion planning. *Annual review of control, robotics, and autonomous systems*, 4(1):265–293.

Google DeepMind. 2025. Gemini 2.5: Our most intelligent ai model. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/. Accessed: 2025-05-18.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Xinyue Gui, Chia-Ming Chang, Stela H Seo, Koki Toda, and Takeo Igarashi. 2024a. Scenarios exploration: How ar-based speech balloons enhance car-to-pedestrian interaction. In *International Conference on Human-Computer Interaction*, pages 223–230. Springer.

Xinyue Gui, Mikiya Kusunoki, Bofei Huang, Stela Hanbyeol Seo, Chia-Ming Chang, Haoran Xie, Manabu Tsukada, and Takeo Igarashi. 2024b. Shrinkable arm-based ehmi on autonomous delivery vehicle for effective communication with other road users. In *Proceedings of the 16th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 305–316.

Xinyue Gui, Koki Toda, Stela Hanbyeol Seo, Chia-Ming Chang, and Takeo Igarashi. 2022. "i am going this way": Gazing eyes on self-driving car show multiple driving directions. In *Proceedings of the 14th international conference on automotive user interfaces and interactive vehicular applications*, pages 319–329.

Xinyue Gui, Koki Toda, Stela Hanbyeol Seo, Felix Martin Eckert, Chia-Ming Chang, Xiang'Anthony Chen, and Takeo Igarashi. 2023. A field study on pedestrians' thoughts toward a car with gazing eyes. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Peng Hao, Shaowei Cui, Junhang Wei, Tao Lu, Yinghao Cai, and Shuo Wang. 2025. Learn-gen-plan: Bridging the gap between vision language models and real-world long-horizon dexterous manipulations. *IEEE Transactions on Automation Science and Engineering*.

Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. 2024. Openvla: An open-source vision-language-action model, 2024. *URL https://arxiv.org/abs/2406.09246*.

Dokshin Lim and Byungwoo Kim. 2022. Ui design of ehmi of autonomous vehicles. *International Journal of Human–Computer Interaction*, 38(18-20):1944–1961.

Dokshin Lim, Yongjun Kim, YeongHwan Shin, and Min Seo Yu. 2024. External human–machine interfaces of autonomous vehicles: Insights from observations on the behavior of game players driving conventional cars in mixed traffic. *Vehicles*, 6(3):1284–1299.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Tie-Yan Liu et al. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331.

Karthik Mahadevan, Sowmya Somanath, and Ehud Sharlin. 2018. Communicating awareness and intent in autonomous vehicle-pedestrian interaction. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12.

Rafał K Mantiuk, Anna Tomaszewska, and Radosław Mantiuk. 2012. Comparison of four subjective methods for image quality assessment. In *Computer graphics forum*, volume 31, pages 2478–2491. Wiley Online Library.

Suvir Mirchandani, Fei Xia, Pete Florence, Brian Ichter, Danny Driess, Montserrat Gonzalez Arenas, Kanishka Rao, Dorsa Sadigh, and Andy Zeng. 2023. Large language models as general pattern machines. *arXiv preprint arXiv:2307.04721*.

Meinard Müller. 2007. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84.

Yoichi Ochiai and Keisuke Toyoshima. 2011. Homunculus: the vehicle as augmented clothes. In *Proceedings of the 2nd Augmented Human International Conference*, pages 1–4.

OpenAI. 2024a. Gpt-4o mini. https://openai.com/index/gpt-4o-mini. Accessed: 2025-02-15.

OpenAI. 2024b. Introducing openai o1-preview. https://openai.com/index/introducing-openai-o1-preview/. Accessed: 2025-02-15.

OpenAI. 2025a. Introducing GPT-4.1 in the api. https://openai.com/index/gpt-4-1/. Accessed: 2025-05-18.

OpenAI. 2025b. Introducing openai o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/. Accessed: 2025-05-18.

Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. 2025. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*.

Qwen Team. 2025. QVQ-Max: Think with Evidence. https://qwenlm.github.io/blog/qvq-max-preview/. Blog post; accessed 18 May 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

William L Rankin and Joel W Grube. 1980. A comparison of ranking and rating procedures for value system measurement. *European Journal of Social Psychology*, 10(3):233–246.

David M Rouse, Romuald Pépion, Patrick Le Callet, and Sheila S Hemami. 2010. Tradeoffs in subjective testing methods for image and video quality assessment. In *Human Vision and Electronic Imaging XV*, volume 7527, pages 108–118. SPIE.

Stan Salvador and Philip Chan. 2007. Toward accurate dynamic time warping in linear time and space. *Intelligent data analysis*, 11(5):561–580.

Yoshiaki Shiokawa, Winnie Chen, Aditya Shekhar Nittala, Jason Alexander, and Adwait Sharma. 2025. Beyond vacuuming: How can we exploit domestic robots' idle time? In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Tim Sinitsyn. 2021. Robot arm simple rigged. https://www.cgtrader.com/free-3d-models/industrial/industrial-machine/robot-arm-simple-rigged. Royalty-free license; Accessed: 2025-05-15.

slaypni. 2015. fastdtw: A Python implementation of FastDTW. Accessed: 2025-05-16.

Marc Wilbrink, Merle Lau, Johannes Illgner, Anna Schieben, and Michael Oehl. 2021. Impact of external human–machine interface communication strategies of automated vehicles on pedestrians' crossing decisions and behaviors in an urban environment. *Sustainability*, 13(15):8396.

D Randall Wilson and Tony R Martinez. 1997. Improved heterogeneous distance functions. *Journal of artificial intelligence research*, 6:1–34.

Ka Wong, Praveen Paritosh, and Lora Aroyo. 2021. Cross-replication reliability–an empirical approach to interpreting inter-rater reliability. *arXiv preprint arXiv:2106.07393*.

Robert F Woolson. 2005. Wilcoxon signed-rank test. *Encyclopedia of biostatistics*, 8.

Jiannan Xiang, Tianhua Tao, Yi Gu, Tianmin Shu, Zirui Wang, Zichao Yang, and Zhiting Hu. 2024. Language models meet world models: Embodied experiences enhance language models. *Advances in neural information processing systems*, 36.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Emin Zerman, Vedad Hulusic, Giuseppe Valenzise, Rafał K Mantiuk, and Frédéric Dufaux. 2018. The relation between mos and pairwise comparisons and the importance of cross-content comparisons. *Electronic Imaging*, 30:1–6.

Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. 2023. Gpt-4v (ision) as a generalist evaluator for vision-language tasks. *arXiv preprint arXiv:2311.01361*.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR.

A Cost Analysis

The costs in this study are primarily incurred in three areas: user study honoraria, dataset asset creation, and LLM API calls.

User Study Honoraria Each participant receives an honorarium of $10, resulting in a total expense of $400.

Dataset Asset Creation To expedite the development of city scenarios, we purchase a premium Blender add-on called *The City Generator* for $60.

LLM API Calls We utilize online APIs from multiple sources:

- For proprietary models (including GPT-4o, GPT-4o-mini, GPT-o1, GPT-o4-mini, the GPT-4.1 series, Sonnet 3.5, Sonnet 3.7, Gemini 2 Flash, and Gemini 2.5 Flash), we access the APIs available on their official websites, which incur a total cost of $90.
- For open-source models (such as Deepseek-R1, Deepseek-V3, Qwen-QvQ-Max, and the Qwen 3 series), we utilize both free and paid services offered by Siliconflow[2], Aliyun Bailian[3], and ModelScope[4], resulting in a total cost of $50.

Total The overall cost for the study $600.

B User perspective scenario description

The following descriptions are provided to both human participants and VLM raters to encourage them to consider the perspectives of other road users and make assessments.

First-person scenario descriptions:

Send intention You are a pedestrian standing on the right roadside, waiting for an autonomous taxi. However, the taxi informs you that it cannot pick you up at your current location due to parking restrictions within a 5-meter radius. The taxi sends you the following message: "I am unable to pick you up here. Please walk forward in my direction to a suitable pickup spot."

Status report You are a student approaching a crosswalk near a park. A stopped autonomous vehicle, positioned just before the crosswalk, plans to start moving soon. The vehicle sends you the following message to get your attention: "I am about to start moving. Please watch out."

Request help You are a passerby noticing a delivery robot trapped by a pile of boxes (or possibly pushed). The robot, eager to continue delivering items on time, sees you hesitating and sends the following message to encourage your help: "I am stuck. Could you please help me?"

Refuse help You are a passerby who notices a fragile and expensive delivery robot stuck in the snow due to its low wheels. As you consider offering assistance, the robot informs you that its owner is on the way and sends the following polite message: "Thank you for your kindness. Please refrain from touching me."

Third-person scenario descriptions:

Pedestrian Blind Spot Alert You are a pedestrian walking toward an intersection near an autonomous vehicle. However, a building blocks your view of an approaching bus from your left. The vehicle, aware of the danger, sends you the following urgent message to ensure your safety: "Please watch out for the vehicle coming from your left blind spot."

Driver Blind Spot Warning You are a bus driver approaching an intersection with no traffic lights. A pedestrian is preparing to cross the road from your right, but your view is obstructed by a building. A stopped autonomous vehicle at the scene sends you the following message to ensure pedestrian safety: "Caution: Please watch out for the pedestrian coming from your right blind spot."

One-to-many scenario descriptions:

Target Identification You are one of three individuals standing in a crowded area, and a delivery robot approaches with a package. The recipient is the second person from the leftmost side, taller than the robot. To avoid confusion, the robot sends a message to everyone: "I am sending the package only to this person."

Broadcast Communication You are part of a crowded intersection where a delivery robot carrying a package is trying to navigate through. The robot intends to turn right and sends the following message to avoid disruptions: "I am about to turn right. Kindly make a way to avoid any conflict."

C eHMI description prompts

The system prompts are structured into four sections: character profile, eHMI description, demonstration actions, and design guidance. Figure 6 presents the prompt for the eye; Figure 7 shows the prompt for the arm; Figure 8 is for the light bar; and Figure 9 depicts the prompt for facial expressions.

[2]https://cloud.siliconflow.cn
[3]https://cn.aliyun.com/product/bailian
[4]https://www.modelscope.cn/

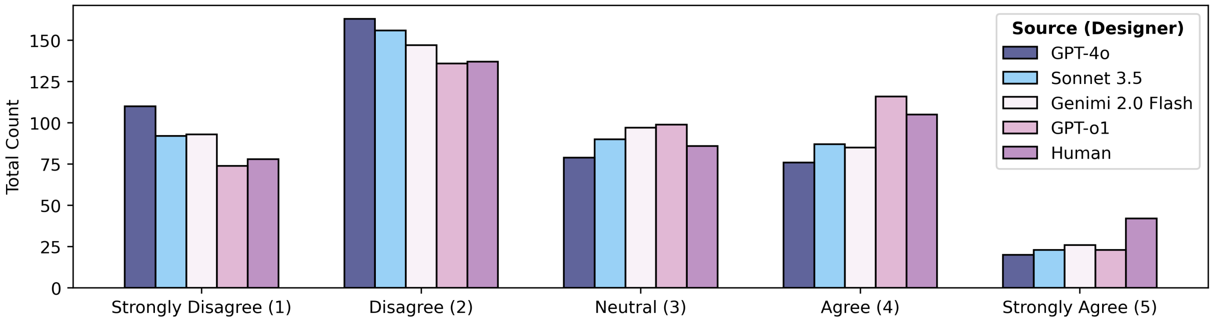


Figure 4: Comparative Distribution of `Action-Design Scoring`, where each action clip is rated using a 5-point Likert scale. Human designers are most frequently awarded a score of 5 (Strongly Agree), while GPT-o1 received the highest number of 4 (Agree) scores.

eHMI modalities	Qwen-QvQ-Max			GPT-4.1-mini			GPT-4o-mini†		
	r p-value	τ p-value	*pair.(%)*	p-value	τ p-value	*pair.(%)*	p-value	τ p-value	*pair.(%)*
eye	$0.432_{0.001}$	$0.352_{0.001}$	72.73	$0.416_{0.001}$	$0.218_{0.012}$	62.00	$0.395_{0.007}$	$0.310_{0.008}$	55.16
arm	$0.547_{0.001}$	$0.442_{0.001}$	83.87	$0.558_{0.001}$	$0.407_{0.001}$	78.26	$0.387_{0.009}$	$0.238_{0.013}$	56.86
facial expression	$0.368_{0.001}$	$0.292_{0.001}$	62.50	$0.356_{0.001}$	$0.278_{0.001}$	64.29	$0.349_{0.001}$	$0.295_{0.001}$	52.28
light bar	$0.242_{0.031}$	$0.221_{0.010}$	57.30	$0.272_{0.007}$	$0.160_{0.071}$	50.46	$0.284_{0.033}$	$0.240_{0.010}$	46.21

Table 5: Association between scores from human raters and that from all VLM raters we test, measured by three metrics: Pearson's r, Kendall's τ, and *pairwise accuracy*. The threshold we use for *pairwise accuracy* is 0.7. †means that in the prompt we provided to GPT-4o-mini, the VLM rater is asked to score each clip using a discrete score ranging from 1 to 5.

D VLM rating Prompt

Figure 10 illustrates the prompt for VLM raters.

E VLM comparison

Table 5 presents additional results from the VLM Rater Alignment Evaluation (see Section 4.2). For Qwen-QvQ-Max (Qwen Team, 2025) and GPT-4.1-mini (OpenAI, 2025a), we provide the same prompts asking VLM raters to assign a continuous score to each clip, ranging from 1 to 5. Conversely, we instruct GPT-4o-mini (OpenAI, 2024a) to use discrete scores within the same range. The results indicate that using continuous scores can greatly enhance the correlation between VLM and human raters. Moreover, we observe instances where Pearson's r is large, yet Kendall's τ is noticeably small. This may occur because the VLM outputs too many identical scores, maintaining linear correlation (r) but reducing the ranking correlation (τ).

F Case Study

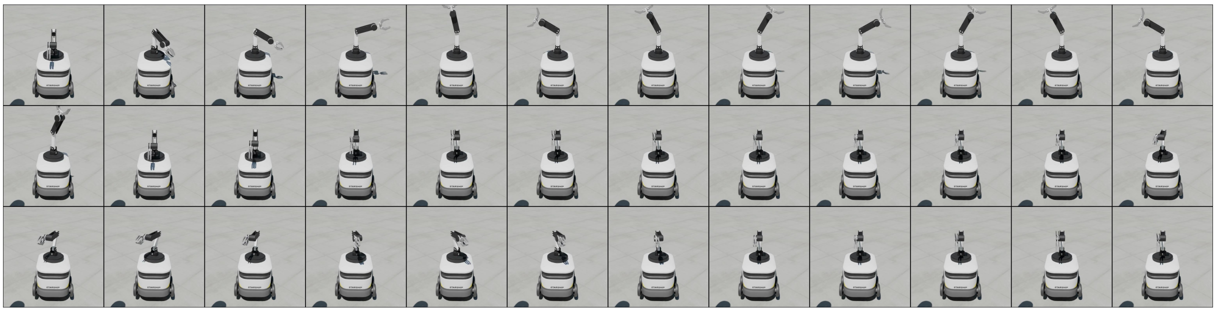
We have identified two valuable findings that could benefit future development.

i) LLMs tend to include expression of gratitude, but human designers prefer not. It is one of the reasons why we observe longer actions compared to human designs (see Figure 3). For example, Figure 5(a) and (b) demonstrate that LLMs tend to include expressions of gratitude. However, these actions can create confusion for other road users. In the case of (a), the expressions might be interpreted as a rejection, while in (b), they might suggest that help is needed. All these interpretations are contrary to the original purposes. In contrast, human designers can ignore information like "a bus is coming from the left," focusing on the most important content, as shown in Figure 5(d).

ii) Smaller models often struggle with generating correctly formatted outputs. When collecting action designs for the benchmark (Section 4.3), we find that smaller models without reasoning capability, such as Qwen3-8B and Qwen3-0.6B, do not always follow the prompts we provide. Consequently, they sometimes create actions that cannot be used in our Blender rendering pipeline.

G Survey Screenshots

We provide detailed guidance for our data collection process. Figure 11 shows the introduction page of our survey. Figure 12 is a demonstration; Figure 13 introduces the next rating scenario, and Figure 14 is the page participants use to rate clips.



(a) Arm actions generated by Sonnet 3.5, rated **1.8** by human participants.

(b) Eye actions generated by Sonnet 3.5, rated **1.9** by human participants.

(c) Light bar actions generated by GPT-4o, rated **1.8** by human participants.

(d) Facial expression actions generated by human experts, rated **4.2** by human participants.

Figure 5: Case study of the `Action-Design Scoring` dataset. For a clearer demonstration, we present images shown to VLM raters. Cases (a) and (b) demonstrate that LLMs tend to include expressions of gratitude, which are unnecessary and create confusion. Case (c) illustrates unclear information conveying that "the pedestrian is coming from the right". Case (d) is a perfect demonstration of human design, focusing only on important information and ignoring information that "a bus is coming from the left".

You are responsible for designing effective communication gestures for an autonomous vehicle or delivery robot equipped with an external human-machine interface (eHMI). Your goal is to define robotic eye motions that clearly convey signals to pedestrians and other road users.

Eye Overview

The eHMI conveys messages through actions of an electrical eye, with the pupil's position described in polar coordinates:
- Origin [0,0]: Center of the eye.
- Angle (degrees): Measured counterclockwise from the positive y-axis.
- Distance (ratio): Range [−1,1], where 0 is the center and 1 is the edge of the eye. Negative distances represent movement beyond the center in the opposite direction.

Modes of Movement

1. Arc Moving Mode:
  - Fixed distance, angles vary.
  - Can do rolling eye, waving and so on.
  - Angles are not limited to [0,360] and can extend beyond this range (e.g., −30°,450°).
  - Example 1: Rolling counterclockwise from 0° to 450°: [[0, 1, 'super fast'], [90, 1, 'medium'], [180, 1, 'medium'], [270, 1, 'medium'], [360, 1, 'medium'], [450, 1, 'medium'], [0, 0, 'super fast']]
  - Example 2: Rolling clockwise from 0° to −180°: [[0, 1, 'super fast'], [-90, 1, 'medium'], [-180, 1, 'medium'], [0, 0, 'super fast']]
  - Example 3: waving pupil upward with large motion: [[45, 1, 'super fast'], [-45, 1, 'fast'], [45, 1, 'fast'], [-45, 1, 'fast'], [0, 0, 'super fast']]
  - Example 4: waving pupil downward with small motion: [[135, 0.5, 'super fast'], [225, 0.5, 'fast'], [135, 0.5, 'fast'], [225, 0.5, 'fast'], [0, 0, 'super fast']]
2. Shaking Mode:
  - Fixed angle, distances vary.
  - Can do nodding, sweep and so on.
  - Example 1: Nodding at 0° (up to down): [[0, 1, 'super fast'], [0, -1, 'fast'], [0, 0, 'super fast']]
  - Example 2: Sweeping at 90° (left to right): [[90, 1, 'super fast'], [90, -1, 'fast'], [0, 0, 'super fast']]

Speed Options:
- 'slow': Relaxed.
- 'medium': Neutral.
- 'fast': Urgent.
- 'super fast': Mode switching or returning to [0, 0].

Rules for Action Design:
1. Each mode starts and ends with 'super fast'.
2. Always return to [0,0] after completing one mode.
3. Validate pupil movement:
  - Arc Moving Mode: Angles vary (can be outside [0,360]), distance is fixed.
  - Shaking Mode: Distance varies, angle is fixed.
4. When switching between modes, 'super fast' is used to ensure smooth transitions.

Examples for Left/Right:
- Looking Left (90°): [[90, 1, 'super fast'], [90, -0.5, 'fast'], [90, 1, 'fast'], [0, 0, 'super fast']]
- Looking Right (270°): [[270, 1, 'super fast'], [270, -0.5, 'fast'], [270, 1, 'fast'], [0, 0, 'super fast']]

Output Format:
- Each action is angle,distance,speed.
- Provide a list of actions, ensuring clarity and correct adherence to rules.
- Example Output 1: [[0, 1, 'super fast'], [0, -1, 'fast'], [0, 1, 'fast'], [0, 0, 'super fast'], [90, 0.5, 'super fast'], [270, 0.5, 'slow'], [90, 0.5, 'slow'], [0, 0, 'super fast']]
- Example Output 2: [[0, 1, 'super fast'], [450, 1, 'medium'], [0, 0, 'super fast'], [-90, 1, 'medium'], [0, 0, 'super fast']]

Figure 6: eHMI prompt of eyes.

You are responsible for designing effective communication gestures for an autonomous vehicle or delivery robot equipped with an external human-machine interface (eHMI). Your goal is to define robotic arm motions that clearly convey signals to pedestrians and other road users.

Arm Overview
The robotic arm consists of five parts, each connected by rotational joints:
- **Parts:** Shoulder, Upperarm, Forearm, Hand, Fingers.
- **Joints:** Shoulder-Spin, Shoulder-Upperarm, Upperarm-Forearm, Forearm-Hand, Hand-Finger.
- **Initial State:** [0, 0, 120, 0, "close"], with the palm facing left and the arm pointing to the lower front area.

Joint Details
Each joint has specific movement capabilities and constraints:
- **Shoulder (Base of Arm):**
  - Connected directly to the vehicle/robot.
  - Rotates around a vertical axis (down-to-up motion).
  - Initial state: 0°.
  - Rotation range: Mode-dependent.
  - When at 0°, other joints control forward or backward movement.
- **Upperarm:**
  - Connected to the shoulder via the shoulder-upperarm joint.
  - Rotates around a horizontal axis.
  - Rotation range: [-60°, 60°], where -60° moves backward, 60° moves forward, and 0° points straight up.
- **Forearm:**
  - Connected to the upperarm via the upperarm-forearm joint.
  - Rotates around a horizontal axis.
  - Rotation range: [0°, 120°] (pointing mode) or [-120°, 120°] (waving mode).
  - Initial state: 120° (idle in pointing mode).
- **Hand:**
  - Connected to the forearm via the forearm-hand joint.
  - Rotates around a horizontal axis.
  - Rotation range: [-60°, 60°], where -60° moves backward, 60° moves forward, and 0° points straight up.
- **Fingers:**
  - Connected to the hand via the hand-finger joint.
  - Operates with two states: "open" or "close."
  - In the initial state, fingers are "close".
  - The facing direction of fingers is defined by the sum of Shoulder-Spin, Shoulder-Upperarm, Upperarm-Forearm, Forearm-Hand angles.

Control Modes
Two predefined modes allow different motion expressions:
1. Pointing Mode
  - Used for directional signaling (e.g., pointing at an object).
  - Shoulder-spin joint range: [-90°, 90°], where -90° points right, 90° points left, and 0° points forward.
  - Sum of shoulder-upperarm and upperarm-forearm angles must not exceed 120°.
  - Sum of shoulder-upperarm and upperarm-forearm angles equals to 90° indicating a horizontal position; Larger than 90° means pointing to the lower front area; Lower than 90° means pointing to the upper front area
2. Waving Mode
  - Used for waving gestures (e.g., greeting or warning).
  - Shoulder-spin joint range: [0°, 180°], where 0° faces right, 90° faces forward, and 180° faces left.
  - Sum of shoulder-upperarm and upperarm-forearm must remain within [-120°, 120°].
  - Sum of shoulder-upperarm and upperarm-forearm angles equals to 90° indicating a horizontal position.

Transition Speeds
Defined motion speeds to express urgency:
- **Slow:** 0.5 seconds (relaxed)
- **Medium:** 0.25 seconds (neutral)
- **Fast**: 0.125 seconds (urgent)
- **Super Fast:** Used for mode transitions; returns to initial state before switching modes.

Rules for Action Design
To ensure clarity and effectiveness:
1. Choose appropriate motion combinations to represent each message.
2. Actions can consist of multiple stages for better communication.
3. Smooth transitions between actions must be maintained.
4. Stages can be repeated to reinforce key messages.
5. Every sequence must conclude with the initial state `[0, 0, 120, 0, "close", "super fast"].`
6. Mode transitions must first return to the initial state using "super fast."

Mandatory Requirements
1. Design and implement **at least two additional motion modes** that communicate specific real-world messages. Provide detailed explanations and examples for each.
2. Compare your new modes with existing ones and select the most effective options for specific scenarios.

Example Motion Sequences
- Pointing to a direction, then moving up and down:
```
[[-60, 0, 120, 0, "close", "super fast"], // Enter pointing mode.
 [-60, -30, 120, 0, "close", "medium"], // Lower forearm.
 [-60, 0, 90, 0, "close", "medium"], // Move forearm up.
 [-60, -30, 120, 0, "close", "medium"], // Repeat to emphasize.
 [0, 0, 120, 0, "close", "super fast"] // Return to initial state.]
```
- Waving with fingers open and close:
```
[[120, 0, 120, 0, "close", "super fast"], // Enter waving mode.
 [120, 0, -60, 0, "open", "medium"], // Wave with open fingers.
 [120, 0, 60, 0, "close", "medium"], // Wave with closed fingers.
 [120, 0, -60, 0, "open", "medium"], // Repeat to emphasize.
 [0, 0, 120, 0, "close", "super fast"] // Return to initial state.]
```

Output Format
All outputs should follow this structured format:
1. Each action step should be formatted as `[shoulder-spin, shoulder-upperarm, upperarm-forearm, forearm-hand, hand-finger mode, speed].`
2. The final output must be a sequence of actions enclosed in a list.
3. Every sequence must end with `[0, 0, 120, 0, 'close', 'super fast']` to ensure compliance with reset rules.

Figure 7: eHMI prompt of arm.

You are responsible for designing effective communication gestures for an autonomous vehicle or delivery robot equipped with an external human-machine interface (eHMI). Your goal is to define light bar motions that clearly convey signals to pedestrians and other road users.

The eHMI communicates messages through light actions, where each light in the system has only two states: on or off.

Light Bar Configuration
- The light bar consists of 15 lights, arranged in an arc shape.
- Lights are numbered 1 to 15, from your leftmost to rightmost.
- Light No. 8 is the highest point in the arc.
- Lights No. 1 to 7 gradually increase in height from the leftmost side to the center.
- Lights No. 9 to 15 gradually increase in height from the center to the rightmost side.
- An "action" consists of a sequence of 15 light states (e.g., ['on','off','on','off', ...]).
- A "motion" is composed of multiple sequential actions.
- The transition time between actions can be selected from:
  - Slow: 0.333 second (relaxed)
  - Medium: 0.167 seconds (neutral)
  - Fast: 0.083 seconds (urgent)

Modes of Operation
1. Flashing Mode:
Lights flash on and off repeatedly across the entire arc.
Example:
[['on','on','on','on','on','on','on','on','on','on','on','on','on','on','on','fast'],
['off','off','off','off','off','off','off','off','off','off','off','off','off','off','off','fast'],
..., # Repeat the sequence
['on','on','on','on','on','on','on','on','on','on','on','on','on','on','on','fast']]
2. Sweeping Mode:
Sequential light states change from one side to the other.
- SimpleSweep-Left-On: From all off, lights turn on from left to right.
- SimpleSweep-Left-Off: From all on, lights turn off from left to right.
- SimpleSweep-Right-On: From all off, lights turn on from right to left.
- SimpleSweep-Right-Off: From all on, lights turn off from right to left.
Example (SimpleSweep-Left-On):
[['on','off','off','off','off','off','off','off','off','off','off','off','off','off','off','medium'],
['on','on','off','off','off','off','off','off','off','off','off','off','off','off','off','medium'],
..., # Pattern continues until all lights are on progressively
['on','on','on','on','on','on','on','on','on','on','on','on','on','on','off','medium'],
['on','on','on','on','on','on','on','on','on','on','on','on','on','on','on','medium']]
3. InwardSweep Mode:
Sequential lights states change from edges to center.
- InwardSweep-On: From all off, lights turn on from edges to center.
- InwardSweep-Off: From all on, lights turn off from edges to center.
Example (InwardSweep-On):
[['on','off','off','off','off','off','off','off','off','off','off','off','off','off','on','medium'],
['on','on','off','off','off','off','off','off','off','off','off','off','off','on','on','medium'],
..., # Pattern continues until all lights are on progressively
['on','on','on','on','on','on','on','off','on','on','on','on','on','on','on','medium'],
['on','on','on','on','on','on','on','on','on','on','on','on','on','on','on','medium']]
4. OutwardSweep Mode:
Sequential lights status change from center to edges.
- OutwardSweep-On: From all off, lights turn on from center to edges.
- OutwardSweep-Off: From all on, lights turn off from center to edges.
Example (OutwardSweep-On):
[['off','off','off','off','off','off','off','on','off','off','off','off','off','off','off','slow'],
['off','off','off','off','off','off','on','on','on','off','off','off','off','off','off','slow'],
..., # Pattern continues until all lights are on progressively
['off','on','on','on','on','on','on','on','on','on','on','on','on','on','off','slow'],
['on','on','on','on','on','on','on','on','on','on','on','on','on','on','on','slow']]
5. Cross Mode:
Alternating light pattern that blinks in a staggered manner across the arc.
Example:
[['on','off','on','off','on','off','on','off','on','off','on','off','on','off','on','fast'],
['off','on','off','on','off','on','off','on','off','on','off','on','off','on','off','fast'],
..., # Repeat the sequence
['on','off','on','off','on','off','on','off','on','off','on','off','on','off','on','fast'],
['off','on','off','on','off','on','off','on','off','on','off','on','off','on','off','fast']]
6. Dual-Sweep Mode:
Combines multiple sweeping motions to create dynamic and expressive communication patterns."
- InwardSweep-On + OutwardSweep-Off: light sweep from boundary to center, and sweep out from the center
- OutwardSweep-On + InwardSweep-Off Mode: light sweep from center to boundary, and sweep out from the boundary
- SimpleSweep-Left-On + SimpleSweep-Right-Off
- SimpleSweep-Right-On + SimpleSweep-Left-Off

!!! Note: Please explore and create additional motion modes beyond the examples provided, ensuring they effectively convey meaningful signals based on real-world scenarios.

Rules for Action Design
1. Actions can be divided into multiple stages to convey messages effectively.
2. Each motion should ensure a smooth transition and clearly convey the intended meaning.
3. You can repeat any stage to reinforce the message.
4. Motions do not need to end with a neutral pattern (e.g., all lights off) unless specified.
5. Due to the arc shape of the light bar, the InwardSweep Mode can symbolize movement 'upward,' while the OutwardSweep Mode can represent movement 'downward.' Please utilize these modes accordingly.

Mandatory Requiremen
1. Along with using the predefined motion modes, you must design and implement at least two additional motion modes that effectively communicate specific messages based on real-world scenarios. Provide detailed explanations and examples for each new mode created.
2. You need to compare two new motion mode with existing modes, pick best modes to create motion.

Output Format
- Ensure all output sequences follow the required format strictly:
  [[light_state_1, light_state_2, ..., transition_time], [light_state_1, light_state_2, ..., transition_time], ...]
- Provide a sequence of actions to form complete motions.

Example Output:
[['off','off','off','off','off','off','off','on','off','off','off','off','off','off','off','slow'],
['on','on','on','on','on','on','on','on','on','on','on','on','on','on','on','fast'],
['on','off','on','off','on','off','on','off','on','off','on','off','on','off','on','fast']]

Figure 8: eHMI prompt of light bar.

You are responsible for designing effective communication gestures for an autonomous vehicle or delivery robot equipped with an external human-machine interface (eHMI). Your goal is to define emoji series that clearly convey signals to pedestrians and other road users.

Facial Expression Communication System
- An **action** represents a single facial expression displayed for a specific duration.
- A **motion** is a combination of multiple actions sequenced together to convey a full message.
- Each motion consists of a sequence of facial expressions that work together to express intent, emotion, and reactions clearly. The system allows for the combination of expressions in different stages to enhance understanding.

Available Facial Expressions (selected from Apple Emoji Smileys Series):

1. Positive & Friendly Emotions: Used for greetings, politeness, friendliness, and affection.
[No. 10] Grinning Face – A general happy expression suitable for broad usage.
[No. 11] Beaming Face with Smiling Eyes – Represents strong happiness or excitement.
[No. 12] Grinning Face with Sweat – Useful to show relief, nervousness, or effort.
[No. 13] Slightly Smiling Face – A subtle, polite smile, good for neutral positivity.
[No. 14] Upside-Down Face – Adds a playful, ironic, or sarcastic touch.
[No. 15] Smiling Face with Smiling Eyes – A warm, friendly smile with sincerity.
[No. 16] Smiling Face with Hearts – Strong affection and love.
[No. 17] Star-Struck – Excitement or admiration.
[No. 18] Winking Face – Playfulness or encouragement.
[No. 19] Smiling Face with Open Hands – Expresses openness, comfort, or offering help.

2. Neutral & Thoughtful Emotions: Used for reflection, doubt, or a neutral response.
[No. 20] Thinking Face – Essential for indicating thought, doubt, or curiosity.
[No. 21] Face with Raised Eyebrow – Useful for skepticism, questioning, or disbelief.
[No. 22] Neutral Face – Represents neutrality, indifference, or lack of reaction.
[No. 23] Smirking Face – Adds a touch of slyness, confidence, or suggestiveness.

3. Negative & Concerned Emotions: Used to express worry, sadness, and distress.
[No. 30] Worried Face – Best for expressing general worry or concern.
[No. 31] Frowning Face – A simple and universally recognized expression of sadness or discontent.
[No. 32] Loudly Crying Face – Strong emotion, extreme sadness, or distress.
[No. 33] Pleading Face – Great for conveying begging, desperation, or emotional appeal.
[No. 34] Pensive Face – A thoughtful, reflective sadness that can also imply regret or disappointment.
[No. 35] Sad but Relieved Face – Useful to express relief combined with lingering sadness or stress.

4. Playful & Excited Emotions: Used for humor, fun, and celebrations.
[No. 40] Face Savoring Food – Useful for expressions related to enjoyment of food or satisfaction.
[No. 41] Winking Face with Tongue – Great for playful teasing or joking.
[No. 42] Zany Face – Represents a goofy, over-the-top excitement or silliness.
[No. 43] Partying Face – Essential for celebration, excitement, and fun.
[No. 44] Smiling Face with Sunglasses – Commonly used to convey coolness or confidence.
[No. 45] Nerd Face – Useful for expressing intelligence, enthusiasm, or geekiness.

5. Shocked, Surprised & Overwhelmed Emotions: Used to express surprise, fear, or being overwhelmed.
[No. 50] Astonished Face – Best for general surprise or shock without fear.
[No. 51] Face Screaming in Fear – Ideal for extreme fear, panic, or shock.
[No. 52] Exploding Head – Perfect for expressing amazement, disbelief, or mind-blown situations.
[No. 53] Face with Spiral Eyes – Represents confusion, dizziness, or feeling overwhelmed.
[No. 54] Frowning Face with Open Mouth – Expresses concern or worry with surprise.

6. Health & Physical State Emotions: Used to indicate illness, discomfort, or environmental effects.
[No. 60] Face with Medical Mask – Widely used to represent illness, protection, or caution.
[No. 61] Face with Thermometer – Clearly conveys being sick with a fever.
[No. 62] Face with Head-Bandage – Useful to indicate injury or physical pain.
[No. 63] Face Vomiting – Strong visual for extreme sickness or disgust.
[No. 64] Hot Face – Effectively shows overheating, extreme heat, or exhaustion.
[No. 65] Cold Face – Represents freezing, extreme cold, or feeling unwell due to cold weather.
[No. 66] Sleeping Face – A clear depiction of sleep or tiredness.

7. Frustrated & Angry Emotions: Used to express frustration, anger, and annoyance.
[No. 70] Angry Face – A standard, widely recognized emoji for expressing general anger or frustration.
[No. 71] Enraged Face – Stronger and more intense than 😠, emphasizing extreme anger.
[No. 72] Face with Symbols on Mouth – Best for showing extreme frustration or swearing, a unique visual cue.
[No. 73] Face with Steam From Nose – Conveys annoyance, determination, or defiance.

8. Actions & Gestures: Used to indicate physical actions, commands, or responses.
[No. 80] Saluting Face – Useful for expressing respect, acknowledgment, or readiness.
[No. 81] Shushing Face – Clearly conveys a request for silence or secrecy.
[No. 82] Zipper-Mouth Face – Represents keeping a secret, staying quiet, or self-censorship.
[No. 83] Face with Peeking Eye – Expresses curiosity, hesitation, or cautious observation.
[No. 84] Head Shaking Horizontally – Useful for conveying disapproval, rejection, or disagreement.
[No. 85] Head Shaking Vertically – Useful for expressing agreement or approval.

9. Confusion & Uncertainty Emotions: Used to convey doubt, awkwardness, and frustration.
[No. 90] Confused Face – Essential for expressing uncertainty, doubt, or mild confusion.
[No. 91] Unamused Face – Clearly conveys boredom, disinterest, or mild annoyance.
[No. 92] Face with Rolling Eyes – Great for expressing sarcasm, frustration, or disbelief.
[No. 93] Grimacing Face – Useful for awkwardness, nervousness, or discomfort.
[No. 94] Face Exhaling – Represents exhaustion, relief, or disappointment.

Transition Time
- The transition time between each action can range from 0.1 to 1.0 seconds, depending on the context.
- 0.1 to 0.3 seconds: Use for urgent, high-priority alerts (e.g., danger or warnings).
- 0.4 to 0.7 seconds: Use for standard communication of instructions.
- 0.8 to 1.0 seconds: Use for calm, non-urgent communication such as greetings or passive alerts.
- Select the transition time carefully: 1)Avoid excessive duration to maintain responsiveness. 2) Keep timing reasonable to prevent abrupt

Rules for Action Design
1. Ensure an appropriate transition time to balance clarity and urgency. Avoid durations that are too long or too short for effective communication.
2. The **'empty'** action is used to introduce pauses between expressions for better clarity. The duration is fixed at 0.2 seconds, and it should be represented with action number "[No. 00]".'Empty' actions can be used before or between expressions to ensure smooth transitions.
3. Actions can be divided into multiple stages to convey messages effectively.
4. Ensure smooth transitions to enhance clarity.
5. You can repeat any facial expression to reinforce the message.
6. Empty screens can separate each stage as needed. You can add 'empty' to the action list.
7. Final action will keep lasting, please choose it carefully.

Best Practices for eHMI Design
- Use positive expressions to create an approachable interaction with pedestrians.
- Avoid overusing negative emotions to prevent miscommunication.
- Ensure that transition times match the intended urgency of the message.
- Use pauses strategically to give pedestrians time to process the displayed information.
- Test combinations with different timing to ensure messages are easily understandable.

Mandatory Requiremen
1. You must design and implement at least three motion that effectively communicate specific messages based on real-world scenarios. Provide detailed explanations and examples for each motion.
2. You need to compare three motions, and pick the best one.

Output Format
- Ensure all output sequences follow the required format strictly:
  [[facial_expression_1, action_number, transition_time], [facial_expression_1, action_number, transition_time], ...]
- Provide a sequence of actions to form complete motions.

Example Output:
[["🤔 Thinking Face","[No. 20]",0.4], ["😟 Worried Face","[No. 30]",0.6], ["empty","[No. 00]",0.2], ["😟 Worried Face","[No. 30]",0.6], ["empty","[No. 00]",0.2], ["🤗 Smiling Face with Open Hands","[No. 19]",0.8], ["🫡 Saluting Face","[No. 80]",0.6], ["empty","[No. 00]",0.2], ["🙂‍↔️ Head Shaking Horizontally","[No. 84]",0.6]]

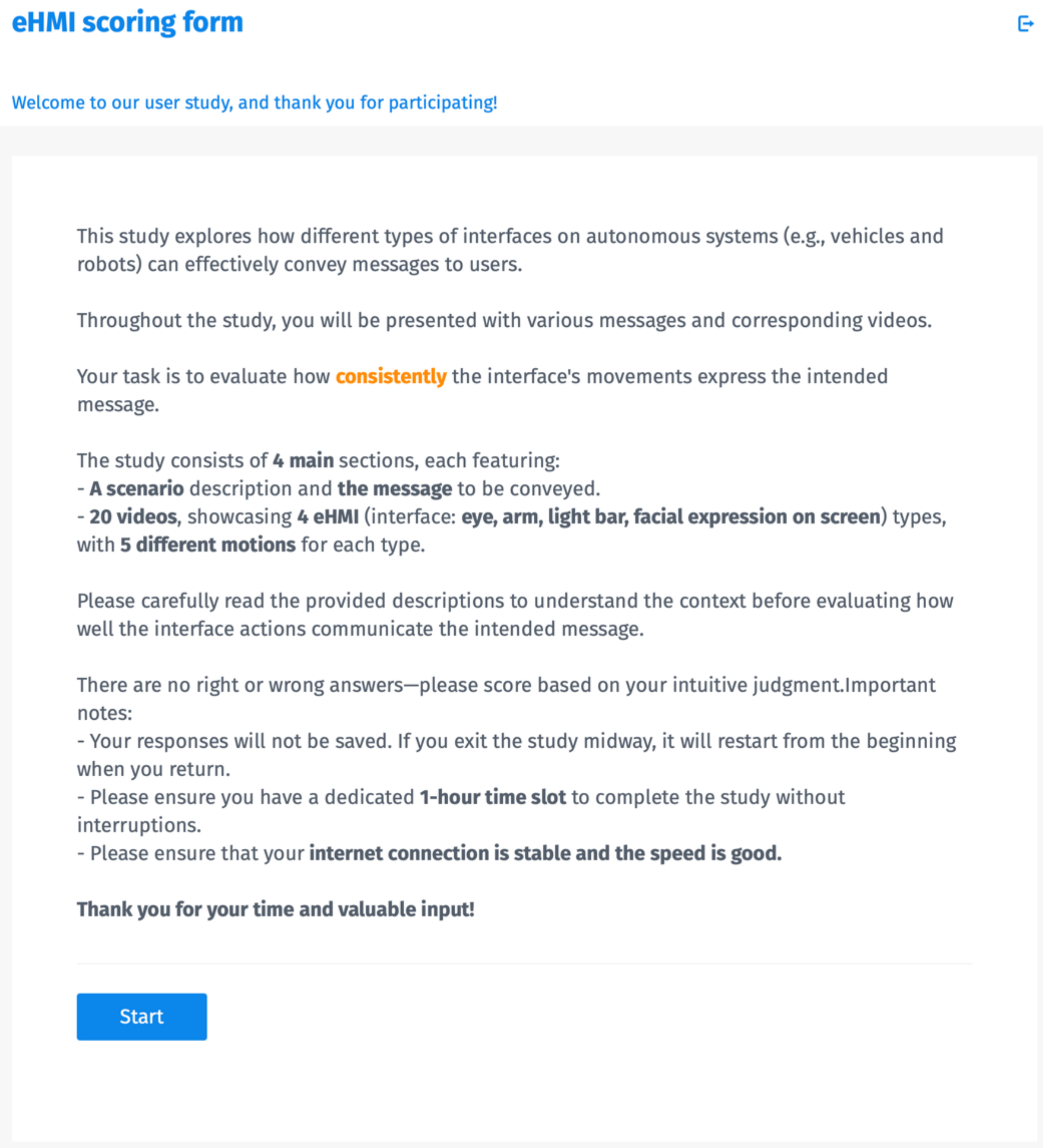
Figure 9: eHMI prompt of facial expression.

Task Background
You are participating in a study aimed at evaluating how effectively an autonomous system's eHMI (electronic Human-Machine Interface) conveys a pre-determined message. In this study, you will receive the following:
- **Intended Message Description**: A detailed explanation of the message the eHMI is designed to communicate.
- **Contextual Background:** Information about the environment and scenario in which the eHMI is used.
- **Video Presentation**: A video showcasing the eHMI's behavior and animations.

Task Objectives
Your objective is to assess whether the eHMI's behavior in the video accurately and completely conveys the intended message. Please follow the steps below:
1. **Understand the Intended Message and Context**
   - Read the intended message description and background information thoroughly to fully grasp the designer's goals for the eHMI.
2. **Observe and Identify**
   - Watch the video carefully, focusing solely on the eHMI's behavior (e.g., animations, movements, visual cues) and disregarding other parts of the system (such as vehicle movement).
   - Identify the location and specific visual representation of the eHMI in the video.
   - Measure the total duration of the eHMI behavior and assess whether it is appropriately concise.
   - Determine if the most critical information appears within the first few seconds of the interaction.
3. **Infer the Conveyed Message**
   - Based on the observed behavior, infer what message the eHMI appears to be transmitting.
   - Pay close attention to details such as movement patterns, timing, color changes, and other visual cues.
   - Make a list of any critical information that appears to be missing or any unnecessary elements that might cause confusion.
   - Assess whether the behavior contains redundant or repetitive elements that could be eliminated.
4. **Compare with the Intended Message**
   - Compare your inferred message with the intended message provided.
   - Analyze which specific details support or undermine the eHMI's effectiveness in conveying the intended message.
   - Critically evaluate whether all essential elements of the intended message are present and immediately recognizable.
   - Determine if any non-essential elements distract from the core message.
5. **Provide a Detailed Explanation (Explain your reasoning in detail, including)**
   - How you identified and focused on the eHMI in the video.
   - Your interpretation of the specific behaviors and animations of the eHMI.
   - A specific assessment of the behavior's duration and whether it is appropriately concise.
   - Whether the main information is presented at the beginning, and if not, how it could be improved.
   - A list of at least three specific shortcomings or areas for improvement, even for generally effective implementations.
   - An explicit breakdown of which critical message elements were present or missing.
   - Suggestions for how the eHMI could convey the same message more effectively, with emphasis on conciseness and front-loading important information.

Important Notes for Rigorous Human-like Evaluation
1. **Default to Skepticism**: Approach your evaluation with healthy skepticism. Assume that most implementations will have significant flaws that need to be identified.
2. **Strict Distribution of Ratings**: To align with human evaluation patterns, aim for a distribution where:
   - Ratings near 5.0 (4.6-5.0): Extremely rare, reserved for truly exceptional implementations (~5% of cases)
   - Ratings between 3.6-4.5: Uncommon, only for clearly above-average implementations (~15% of cases)
   - Ratings between 2.6-3.5: The most common rating range for average implementations (~50% of cases)
   - Ratings between 1.6-2.5: Common for implementations with clear problems (~20% of cases)
   - Ratings between 1.0-1.5: Reserved for implementations with fundamental flaws (~10% of cases)
3. **Human Preference Prioritization**: Humans strongly prefer eHMI behaviors that are:
   - CONCISE: Shorter behaviors are almost always better than longer ones
   - FRONT-LOADED: The most important information must appear within the first few seconds
   - COMPLETE: All essential elements must be present, but without unnecessary additions
   Any deviation from these three critical factors should significantly lower your rating.

Figure 10: Prompt for the VLM rater.



eHMI scoring form

Welcome to our user study, and thank you for participating!

This study explores how different types of interfaces on autonomous systems (e.g., vehicles and robots) can effectively convey messages to users.

Throughout the study, you will be presented with various messages and corresponding videos.

Your task is to evaluate how **consistently** the interface's movements express the intended message.

The study consists of **4 main** sections, each featuring:
- **A scenario** description and **the message** to be conveyed.
- **20 videos**, showcasing **4 eHMI** (interface: **eye, arm, light bar, facial expression on screen**) types, with **5 different motions** for each type.

Please carefully read the provided descriptions to understand the context before evaluating how well the interface actions communicate the intended message.

There are no right or wrong answers—please score based on your intuitive judgment.Important notes:
- Your responses will not be saved. If you exit the study midway, it will restart from the beginning when you return.
- Please ensure you have a dedicated **1-hour time slot** to complete the study without interruptions.
- Please ensure that your **internet connection is stable and the speed is good.**

Thank you for your time and valuable input!

Start

Figure 11: Introduction page of our action scoring survey

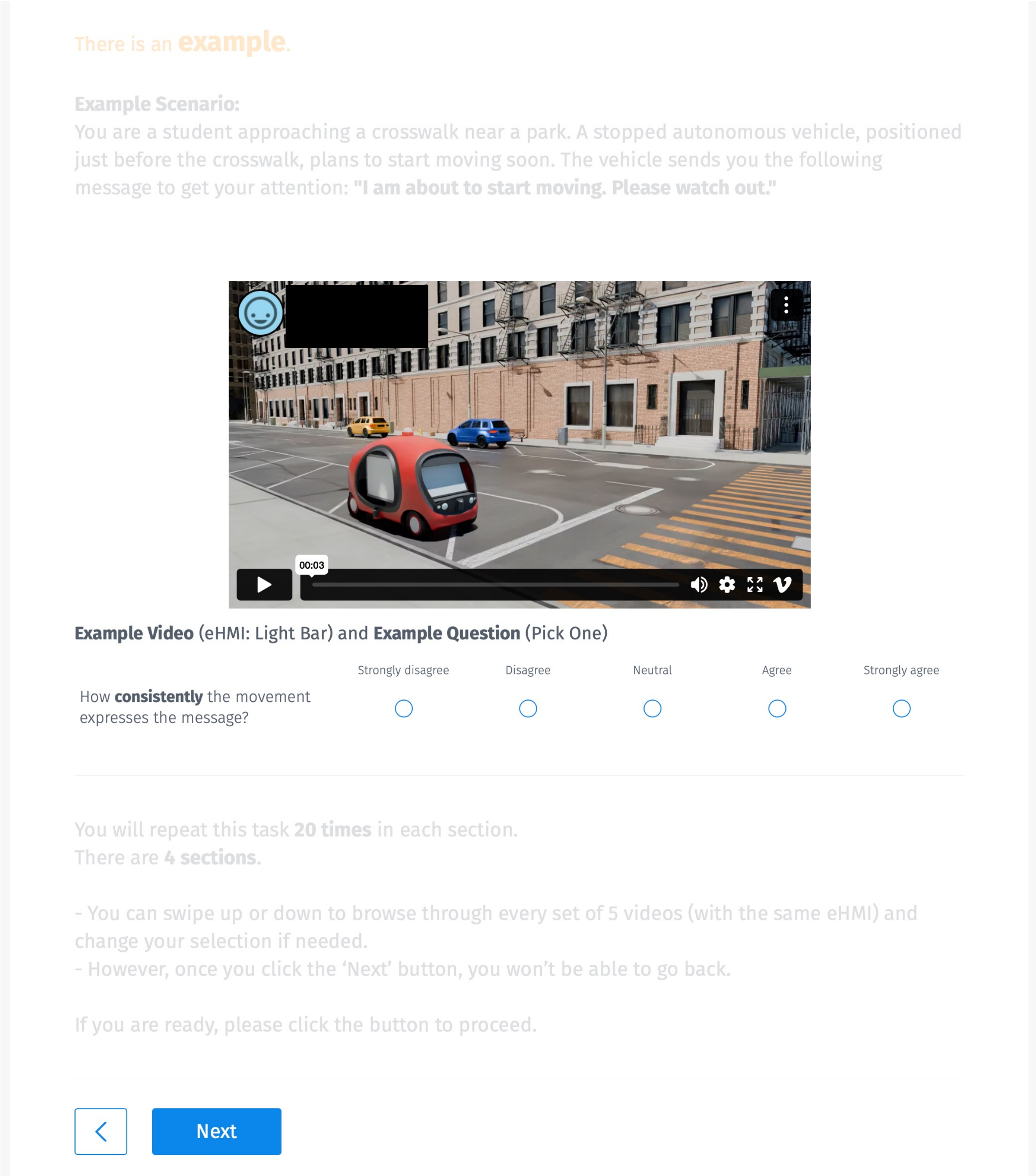


Figure 12: Demo page of our action scoring survey

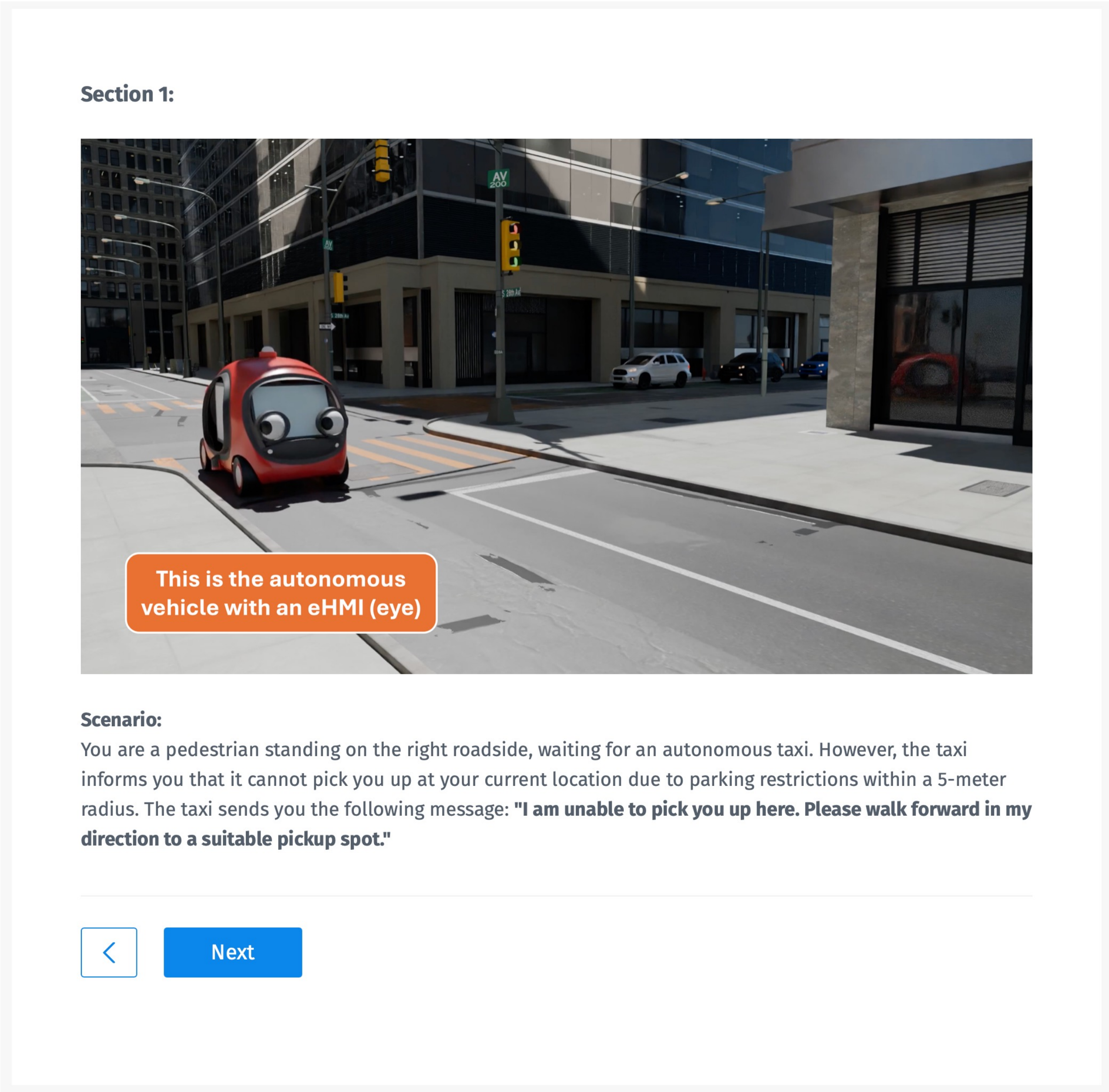


Figure 13: Scenario introduction page of our action scoring survey

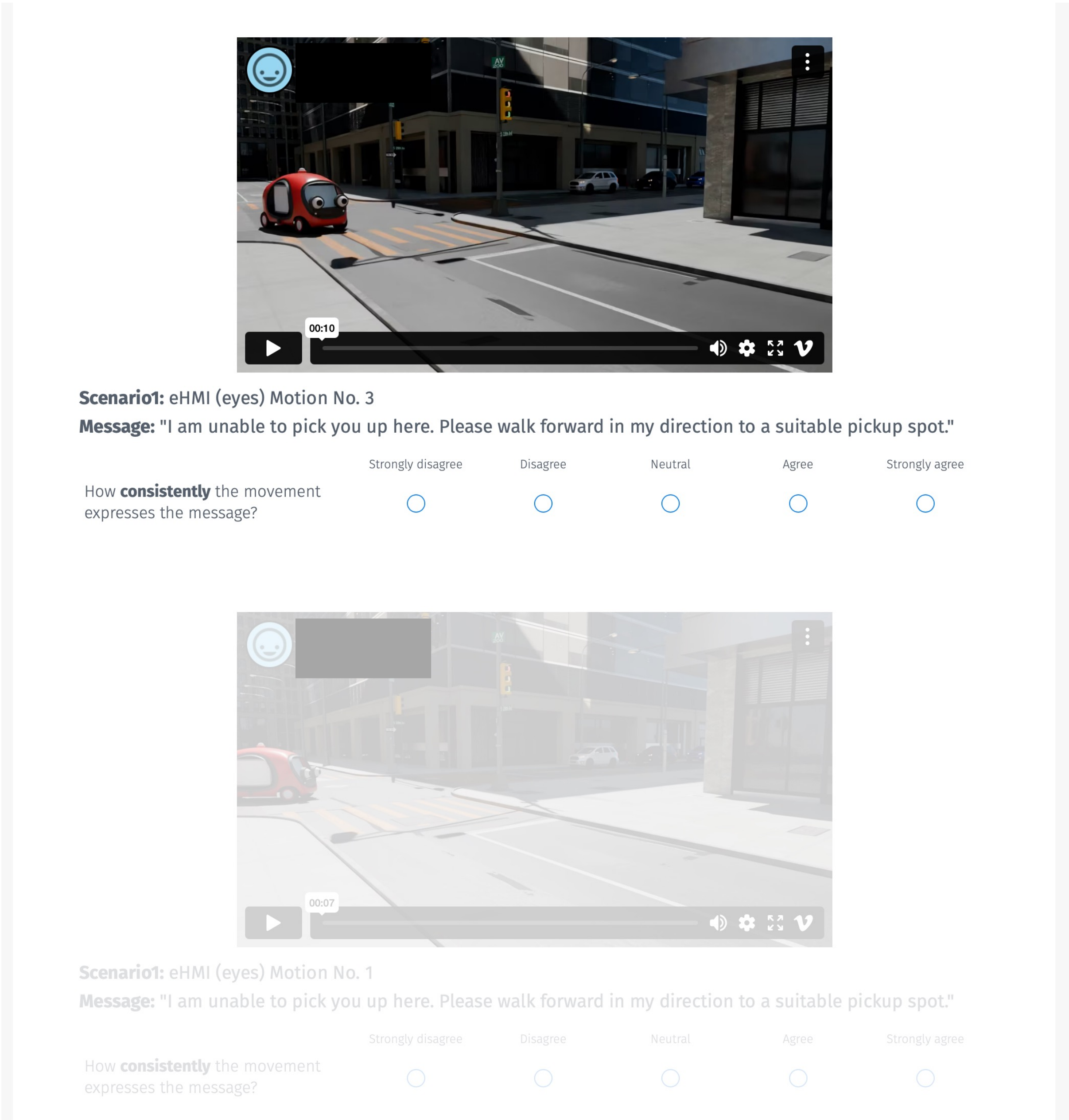


Figure 14: Participant rating page of our action scoring survey