eHMI-Action Scoring: Evaluate LLMs' eHMI Message-to-Action Translation Capability

Anonymous ACL submission

Abstract

The external human-machine interfaces (eHMIs) play a critical role as communication mediators between autonomous vehicles and other road users. However, current eHMI studies are typically evaluated in predefined scenarios that convey fixed messages through fixed action mappings, limiting their applicability in real-world environments where dynamic interactions are required. To address this limitation, we introduced Large Language Models (LLMs) into eHMI actions due to their impressive generativity and versatility across multiple tasks. This raises a key question: **Can the LLM-driven eHMI system consistently translate intended messages into actions that other road users can accurately interpret?** To answer this question, we created an `eHMI-Action Scoring` dataset consisting of eight interaction scenarios with intended messages, four eHMI modalities, ten actions generated by LLMs and human designers for each scenario-modality pair, rendered animations of these actions, and human scores evaluating the actions shown in the animations. Furthermore, we asked visual LLMs to evaluate these action clips, and the results demonstrate that their scores are consistent with those provided by humans, suggesting the feasibility of automated scoring. Finally, we benchmarked the capabilities of other state-of-the-art LLM models.

1 Introduction

With the advancement of autonomous vehicles (AVs), external human-machine interfaces (eHMIs) have emerged as a critical research field to address the communication gap between AVs and human road users (Oudshoorn et al., 2021; Dey et al., 2020a; Bazilinskyy et al., 2019). These interfaces utilize diverse forms, such as displays, projections, and robots, to convey vehicle intentions through text, signals, or non-verbal motions (Dey et al., 2020b; Al-Taie et al., 2024). While promising, current eHMI systems face significant limitations: they are evaluated in narrow, predefined scenarios (e.g., pedestrian crossings, blind spot notification) with fixed messages (e.g., "Please stop", "Watch out!") and fixed eHMI action mappings (Chang et al., 2022; Chauhan et al., 2024; Gui et al., 2024a, 2022). This approach restricts their scalability in real-world environments, where dynamic interactions require adaptive communication.

To address this limitation, we propose leveraging Large Language Models (LLMs) as automated action designers for eHMI systems (Radford et al., 2019). Pre-trained LLMs offer unique advan-

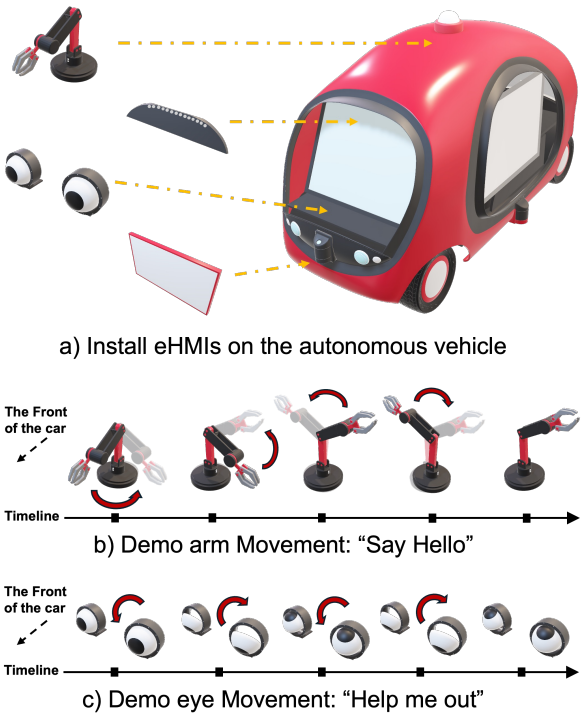


Figure 1: The eHMI setup illustration and action demos. a)Four types of eHMIs are installed on the vehicle separately; b) The demo actions of arm convey the message: "Say Hello". The shaded action indicates the subsequent status.; c)The demo actions of eye convey the message "Help me out".

tages, including contextual reasoning (Kojima et al., 2022; Huang et al., 2022b) and generative capabilities (Mirchandani et al., 2023), which may enable scenario-specific, human-understandable communication. However, the application of LLMs in eHMI action design remains under-explored, raising a critical research question: **Can LLM-driven eHMI systems consistently translate intended messages into actions that other road users can interpret accurately?**

Answering this question involves two key challenges. First, existing methodologies lack a systematic pipeline for translating intended messages into understandable eHMI actions. To bridge this gap, we adapt prompt engineering strategies from task & motion planning with LLMs (Ding et al., 2023; Chen et al., 2024) and customize them for message-to-action translation. Second, there is no empirical evidence validating whether humans can correctly interpret LLM-designed eHMI actions. That is, a benchmark is needed.

To evaluate the consistency between intended messages and perceived meanings, we introduced a user-rated `eHMI-Action Scoring` dataset as a novel benchmark. We designed eight interaction scenarios, each featuring an intended message for the eHMI to convey, and selected four representative eHMI modalities. For each scenario–modality pair, we generated ten actions: eight produced by state-of-the-art LLMs and two designed by human designers. These actions were rendered using Blender, resulting in 320 video clips of eHMI actions. Subsequently, we conducted a video-based user study with 40 participants, in which ten participants per clip rated the consistency between the LLM-designed action and its intended message. The dataset provides averaged human scores for each action, enabling a comparative benchmark for existing LLMs.

Furthermore, we asked visual LLMs (VLLMs) to perform the same task as human raters to evaluate these action clips, and the results show that their scores follow the same relative order as those provided by human raters. This consistency enables us to further benchmark the capabilities of other LLMs. In our benchmark, we found that larger LLMs typically achieve better average scores, and reasoning LLMs exhibit superior performance, even for small distilled LLMs.

The results yield three key findings:

- LLMs demonstrate the capability to generate contextually appropriate eHMI actions.
- Reasoning-enabled LLMs consistently outperform other approaches, even in distilled, smaller versions.
- Visual LLMs (VLLMs) can serve as human-level raters for scoring new clips, providing an automated scoring pipeline.

The future applications of our dataset are twofold. First, eHMI researchers can use our pipeline to render their self-developed or customized eHMIs and scenarios into action clips. Second, language model researchers can adopt this dataset as a benchmark to evaluate their models. Our `eHMI-Action Scoring` dataset and clip rendering pipeline will be publicly released.

2 Related Works

2.1 eHMI design

Current eHMI action planning follows a fixed design approach. Human designers establish behavioral rules based on the specific features of different eHMI modalities. For example, in text- and icon-based eHMIs, designers created contents by referencing traffic regulation icons or messages (Eisele and Petzoldt, 2022; Eisma et al., 2021). In color- and light-band-based eHMIs, they designed the content relying on human intuitive empathy with colors and blinking frequencies (Bazilinskyy et al., 2019; Dey et al., 2020b). For human-like eHMIs, such as eyes or arms, designers mimicked nonverbal communication cues based on common human-human interactions(Mahadevan et al., 2018; Ochiai and Toyoshima, 2011).

To sum up, traditionally, experts observed real-world examples and derived design rules to guide eHMI action planning. However, different eHMI modalities vary in their expressiveness. Low-expressiveness eHMIs, such as arrow icons, are relatively simple, as they convey static directional cues, making it easier to define behavioral rules (Fridman et al., 2017). On the other hand, high-expressiveness eHMIs can exhibit complex actions, allowing them to communicate richer messages (Chang et al., 2024). However, defining rules for these actions is challenging for human experts due to their intricacy and variability (Gui et al., 2023; de Winter and Dodou, 2022). In this project, we addressed this challenge by leveraging LLMs to assist in eHMI action planning, enabling more complex and dynamic communication.

2.2 Task & Motion Planning with LLMs

The existing Task and Motion Planning (TAMP) (Garrett et al., 2021) involves decomposing high-level task instructions into sequences of low-level motion planning problems. Pre-trained Large Language Models (LLMs) (Huang et al., 2022a; Xiang et al., 2024) have demonstrated impressive zero-shot capabilities in utilizing world knowledge and the emerging ability to plan for the TAMP task (Wang et al., 2024). For example, LLM-GROP (Ding et al., 2023) employs a pre-trained LLM to translate requests into symbolic goals which are fed into a low-level planner. AutoTAMP (Chen et al., 2024) uses a zero-shot LLM to translate instructions and state observations into a formal language processable by simple TAMP algorithms.

Our eHMI-planning pipeline shared a similar concept, where the intended messages are treated as high-level instructions that the pre-trained LLM translates into corresponding sets of low-level actions. When grounded into details, the low-level control relies on the predefined functions — often provided by frameworks such as ROS (Quigley et al., 2009) — to generate continuous trajectories and execute precise motion commands. However, unlike the clear task (e.g., grasping the object), our eHMI action planning involved actions that are difficult to define in advance. Therefore, we provided the LLM with detailed structural description prompts for each eHMI, enabling it to control the lowest-level actions such as angular movement and transition speed.

3 eHMI-Action Scoring Dataset

3.1 eHMI Modalities

Four representative eHMIs are selected based on different levels of expressiveness, as shown in Figure 3(a). We designed detailed prompts for each modality of eHMI to ensure that LLMs can fully understand both what they can control and how to control it. Each step of the designed action consists of a next status and transition time.

We first defined the description of status for the four modalities of eHMI. The following status design spaces are described from the perspective of the autonomous vehicle:

Eyes Robotic eyes are mounted at 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, and each of them 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).

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).

We then provided various transition speed options (e.g., "slow", "medium", "fast") when transitioning to the next status. In addition, we included a special transition speed ("super fast") designed to clearly separate each action stage and enhance the readability of actions. This approach guarantees that the output actions are well-formatted and can be directly used to actuate the eHMIs. Detailed prompts are available in appendix D.

3.2 Scenarios

We classified our scenarios into three types based on communication type (Bazilinskyy et al., 2019): first-person, third-person, and one-to-many (see Figure 2). In total, we develop eight scenarios (see Figure 3(a)). Each scenario includes:

- A description from other road users' perspective (provided below).
- A description from the AV's perspective (for details, please refer to Appendix B).
- A message needs to be conveyed by the eHMI.

First-person scenarios involve sending messages about the AV itself. We designed four scenarios:

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

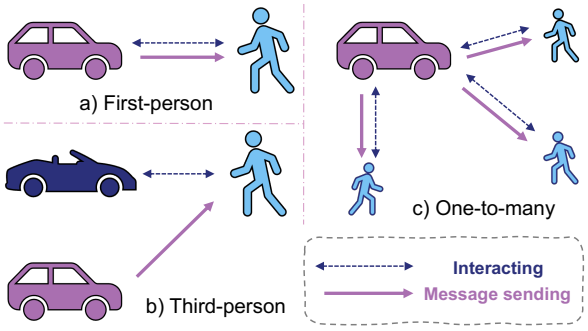


Figure 2: Three types of scenarios. The **purple** vehicle is installed with eHMIs. In the third-person scenario, the message-sending direction is not between interacting pairs. For one-to-many scenarios, the interaction and communication happen between the vehicle with multiple objects.

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 scenarios involve sending messages related to other road users. We designed two scenarios for this type:

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 scenarios involve broadcasting messages from an autonomous system to many individuals. We designed two scenarios:

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."

3.3 Action Clips & Human Scoring

Our `eHMI-Action Scoring` dataset contains two components: eHMI-action clips generation and human scores. The data collection contains two steps, shown in Figure 3(b, c).

In step 1 (Figure 3(b)), we prepared the material containing 320 eHMI-action clips. We first designed an action-rendering pipeline that converts the designed actions into video clips. We used two types of 3D vehicle models (an autonomous car and a delivery robot). The autonomous car model is proprietary, while the delivery robot model is available under an open-source license. We equipped them with four modalities of eHMI individually. They are designed by ourselves. For the eight scenarios, we designed the corresponding 3D environments in Blender with a paid addon named *The city generator 2.0* (Blendermarket, 2025). Then, we obtained a total of 32 scenario-modality pairs. The designed actions are used to change the status of components over different transition durations.

We then developed 10 motions for each scenario-modality pair. Specifically, we asked four state-of-the-art LLMs (GPT-4o (Achiam et al., 2023), Claude Sonnet 3.5 (Anthropic, 2024), Gemini 2 Flash (DeepMind, 2024), and GPT-o1 (OpenAI, 2024b)) to design two distinct actions for each pair. In addition, two human designers performed the same task, yielding a total of 320 rendered motion clips. The scenario prompt used for message-to-action translation was described from the perspective of the AV. With a GPU-equipped device, the overall clip rendering time for 320 clips took 100 hours with the average rendering time for a 10-second clip being approximately 20 minutes.

In step 2 (Figure 3(c)), We invited 40 partici-

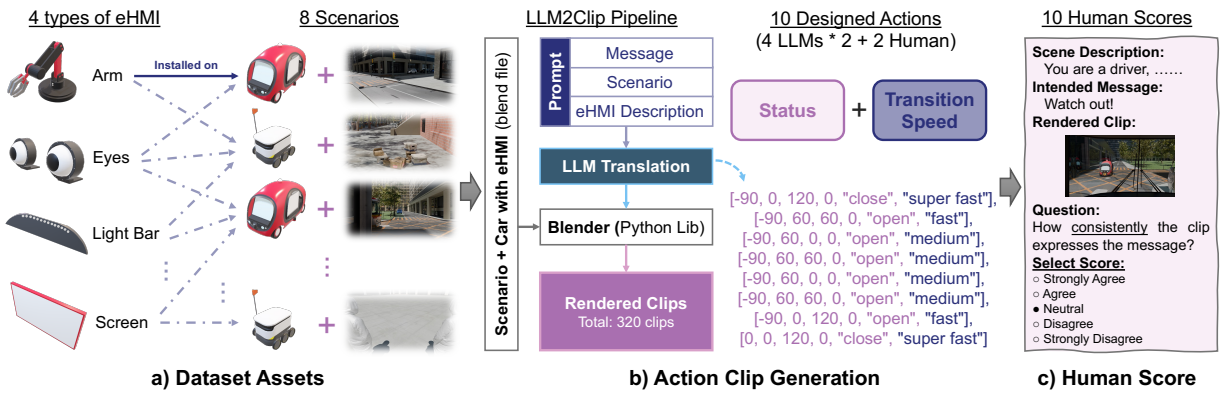


Figure 3: Dataset Asset, Generation, and Scoring. For asset creation, we selected four modalities of eHMIs. Then, we design eight scenarios covering different communication types. In the generation pipeline, we create specific Blender 3D scenarios and employed LLM-designed actions to actuate the eHMIs, resulting in rendered clips. During the scoring phase, 10 participants evaluated each action clip using a 5-point Likert scale.

pants to score the action clips. Each participant received one survey and was asked to answer the question: *"How consistently does the movement express the message?"* Participants rated each clip using a 5-point Likert scale. In total, we collected 3200 scores, with each action clip being scored by 10 different raters. We removed the incomplete responses and computed the average of these selected scores to obtain 320 averaged scores.

3.4 Automated Scoring System

In the future, one might want to evaluate the message-to-action translation capability of novel LLMs. In such cases, employing human raters to score generated actions can be tedious. To address this, we proposed two substitutes.

First, we introduced an Action Reference Score (ARS) that automatically generates a score for a new action by retrieving the most similar actions from our dataset. We used Dynamic Time Warping (DTW) to compute the similarity between action sequences. DTW is particularly effective because it calculates similarity even when identical patterns appear at different positions or when sequences vary in length.

Our approach started by converting the parameters of each action step into numerical values. For example, the angle variable (e.g., 60°) was transformed into its sine and cosine components to capture its cyclical nature accurately. Similarly, categorical variables (e.g., "close") were assigned predefined integer values, and transition times were quantified by assigning "slow" as 4, "medium" as 3, "fast" as 2, and "super fast" as 1. Experimental results indicated that adjusting these predefined values only leads to minor variations in the final output score.

Second, we adopted VLMMs to rate actions based on their common knowledge (Zhang et al., 2023; Gu et al., 2024). We leveraged the inherent multimodal understanding and reasoning capabilities of VLMMs to assess whether the designed actions are contextually appropriate and semantically consistent with the input message. For each action clip, we sampled one frame every six frames, preserving the original temporal order, and downsampled each to 512×512. We used a self-designed prompt with the same scenario description given to human participants to ensure consistent machine and human ratings. The results confirmed the effectiveness of VLLM scores.

4 Experiments and Discussion

In this section, we presented our three experiments progressively: 1) we assessed the LLM's ability to translate messages into corresponding actions; 2) we evaluated the performance of VLLMs in scoring the actions depicted in clips; 3) we benchmarked the actions generated by other LLM models. In addition, we explored the bias toward ratings concerning the lengths of actions.

4.1 LLMs' translation capability

Table 1 shows the statistics of our `eHMI-Action scoring` dataset, while Figure 4 compared the human-rated score distributions between four LLMs and human designers.

First, Table 1 revealed that state-of-the-art (SOTA) LLMs can achieve performance comparable to that of human designers. Notably, the

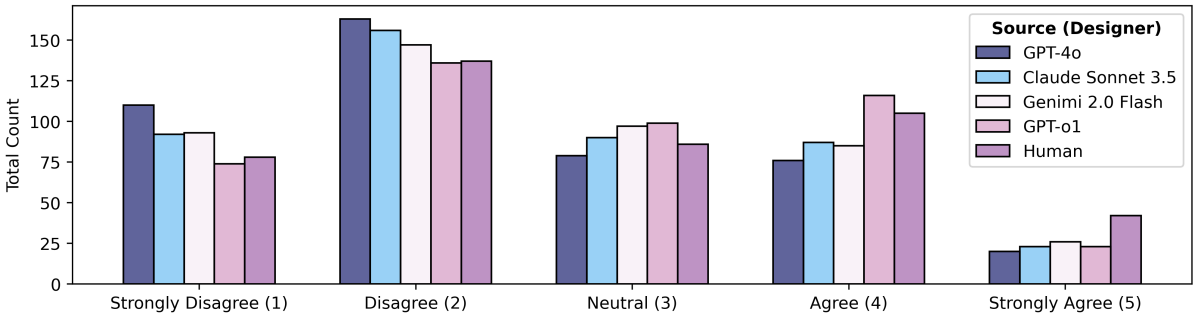


Figure 4: Comparative Distribution of eHMI-Action scores from different sources (designer). Participants rated each action clip using a 5-point Likert scale. Human designers were most frequently awarded a score of 5 (Strongly Agree), while GPT-o1 received the highest number of 4 (Agree) scores.

Source (Designer)	Average	Scenario 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
Claude 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 1: Statistics of the `eHMI-Action Scoring` Dataset: Average scores indicated that LLMs perform comparably to human designers across various scenarios and eHMI modalities. Krippendorff's alpha was also calculated to assess Inter-Rater Reliability (IRR) among human raters.

average score of GPT-o1 was very close to that of human designers. Figure 4 showed a similar trend: human designers most frequently awarded a score of 5 (Strongly Agree), followed by a score of 4 (Agree). In contrast, GPT-o1 received the second-highest number of 5 (Strongly Agree) scores and the highest number of 4 (Agree) scores. Additionally, when examining different scenario types and eHMI modalities, we observed that for the eHMI modality ("eyes"), GPT-o1 achieved an average score of 2.795, which is higher than the 2.536 achieved by human designers. In third-person scenarios, GPT-o1 (3.098) also outperformed human designers (3.045). For those interested in the score distribution from eHMI modalities, please refer to Appendix C and Figure 6.

Second, we noticed that the scores are highly related to both the scenario (message) and the eHMI modalities. For example, the average score for third-person scenarios was higher than for other scenarios. This may be because the intended message design in third-person scenarios was relatively simpler. Meanwhile, regarding different eHMI modalities, the arm modality outperformed the others, while facial expressions scored noticeably lower. This might be due to the types of messages we designed. In our eight scenarios, the majority require conveying spatial information, where the arm modality is advantageous. There was no emotional message (e.g., "I am scary"), which led to the limited performance of the facial expression.

Third, we found that the reasoning-enabled LLM, GPT-o1, outperformed other LLMs, which was supported by its superior performance across different scenario types and eHMI modalities. Additional benchmarking with other reasoning-enabled LLMs in Section 4.3 reinforced this.

Fourth, we validated the effectiveness of our collected data by computing the Inter-Rater Reliability (IRR), which reflects the level of agreement among all participants' scores for all clips generated by one source. We computed Krippendorff's alpha, as a metric of IRR, and found it to be moderate, thereby demonstrating the reliability of our dataset for a subjective task (Wong et al., 2021).

Together, these findings suggested that *LLMs can translate intended messages into eHMI actions that other road users can interpret at a human level*. The current performance is impressive, considering it was solely based on the common knowledge embedded within the LLMs. However, to further enhance performance, it is essential to develop tailored LLMs for specific eHMI modalities. It implied the need for a universal metric that can be

Scene ID	GPT-4o-mini		Qwen-QvQ-72B	
	ρ p-value	τ p-value	ρ p-value	τ p-value
First-person Scenario				
0	0.274 0.08	0.218 0.08	0.304 0.05	0.252 0.05
1	0.256 0.11	0.177 0.13	0.289 0.06	0.224 0.06
2	0.210 0.19	0.156 0.20	0.379 0.01	0.322 0.01
3	0.198 0.22	0.148 0.21	0.196 0.23	0.152 0.21
Third-person Scenario				
4	0.260 0.10	0.200 0.10	0.371 0.01	0.280 0.02
5	0.317 0.04	0.238 0.05	0.233 0.14	0.166 0.16
One-to-many Scenario				
6	0.460 0.01	0.351 0.01	0.271 0.09	0.203 0.08
7	0.210 0.19	0.157 0.20	0.210 0.19	0.165 0.19

Table 2: Association between scores from human raters and that from VLLM raters (GPT-4o-mini and Qwen-QvQ-32B) measured by two rank correlation coefficients: Spearman's ρ and Kendall's τ. ρ measures the strength of a monotonic relationship, while τ focuses solely on the order of the data. Larger ρ and τ both mean higher correlation and a smaller p-value is better.

Source (Designer)	Human	ARS	VLLM	
			4o-mini	QvQ
Human	2.768	-	0.516	0.531
Proprietary models				
GPT-4o	2.369	-	0.481	0.507
Sonnet3.5	2.492	-	0.474	0.514
Gemini2F.	2.509	-	0.479	0.486
GPT-o1	**2.658**	-	0.494	**0.538**
Open source, large models				
Deepseek-R1	-	**2.766**	0.512	0.519
Deepseek-V3	-	2.504	0.488	0.467
Llama3.3-70B	-	2.625	0.490	0.461
QwQ-32B	-	2.596	0.485	0.455
Open source, distilled small models				
Qwen-14B†	-	2.621	**0.524**	0.502
Llama-8B†	-	2.502	0.510	0.486

Table 3: Benchmark for different LLMs using action-reference score (ARS) and VLLM scores. † means these models are distilled by Deepseek-R1. Reasoning-enabled models like Deepseek-R1 and GPT-o1 outperform other LLMs, and even their smaller distilled versions, such as Qwen-14B and Llama-8B, match the performance of larger models.

applied both to the evaluation and reinforcement fine-tuning of future LLMs.

4.2 Visual LLMs validation

To assess the scoring reliability of visual LLMs (VLLMs), we ran an additional experiment. We showed the video clips to the Visual LLMs and asked them to do the same task that the participants did (i.e., scoring consistency).

We adopted one proprietary model (GPT-4o-mini (OpenAI, 2024a)) and one open-source model (Qwen-QvQ-72B (Qwen Team, 2024)) as VLLM raters, considering cost and inference speed. Each VLLM was used to score the clips three times.

We evaluated the results using two rank correlation coefficients: Spearman's ρ and Kendall's τ. Spearman's ρ measures the strength of a monotonic relationship by assessing the absolute differences between these scores, emphasizing how far apart the scores are. In contrast, Kendall's τ focuses solely on the order of the data by comparing the number of concordant and discordant pairs, thus evaluating the consistency of their ordering rather than the magnitude of differences. We presented statistics for the eight scenarios respectively. Table 2 showed the scoring association between human raters and VLLM raters.

First, for most scenarios, we observed a low Spearman's ρ, indicating that the absolute differences between the scores of VLLM and human raters do not consistently follow a monotonic trend. This suggests there is little overall agreement in how the magnitudes of the scores change together. However, we noticed that Kendall's τ values are at a moderate level, implying a fair consistency in the ordering of the score pairs. In other words, many VLLM scores were in the same relative order as human raters scored. Therefore, it is better to use score orders to evaluate new action clips.

Second, we found that both rank correlation coefficients for some scenarios (No. 3 and No. 7) were low. This may be attributed to the downsampling procedure when inputting image series into VLLMs. In those cases, some specific environments blended eHMI details into backgrounds due to downsampling. For example, scenario No. 3 occurred at night under complex lighting conditions, and in scenario No. 7, the autonomous vehicle was far from the camera (observer), causing the eHMI too small to be recognized. These issues can undermine the reliability of scoring action clips, which are important to address in future work.

Finally, These findings indicated that *VLLMs can serve as an effective tool for rating action clips*. It also laid the groundwork for further benchmarking the capabilities of other LLMs.

4.3 Benchmark

To evaluate the performance of different sizes and types of LLMs, we benchmarked them us-

ing two different metrics: the self-designed action-reference score (ARS) (Section 3.4) and VLLM raters, as shown in Table 3.

First, we used ARS to assess large open-source LLMs (e.g., Deepseek-R1 (Guo et al., 2025), Deepseek-V3 (Liu et al., 2024), Llama3.3-70B (Dubey et al., 2024), and Qwen-QwQ-32B (Team, 2024)) as well as small reasoning LLMs distilled by Deepseek-R1 (Qwen-14B and Llama-8B). We found that Deepseek-R1 demonstrated superior performance compared to the other models by achieving an average score of 2.766. Additionally, despite their smaller parameter sizes, both Qwen-14B and Llama-8B attained performance comparable to that of larger LLMs. These findings suggested that LLMs with inherent reasoning capabilities are especially well-suited for the eHMI action design task.

Second, for the VLLM scores, in addition to the LLMs benchmarked by ARS, we included actions designed by proprietary models and human designers. Each action clip was rated three times using GPT-4o-mini and Qwen-QvQ-72B. Based on the findings in Section 4.2, we computed VLLM scores according to the ranking of the scores. First, we combined the scores from all sources (designers) and assigned them ranks in ascending order. Next, we normalized these scores to a $0 \sim 1$ range. Finally, we calculated the average score for each source or designer. Similarly, we observed that reasoning-enabled LLMs, specifically Deepseek-R1 and GPT-o1, achieved human-level performance, while the distilled models (Qwen-14B and LLama-8B) also demonstrated impressive results despite their relatively small sizes.

Overall, these results showed that *reasoning-enabled LLMs perform better for the eHMI action design task*. This finding suggested the potential to develop specialized small LLMs for specific types of eHMI through fine-tuning.

4.4 Action Length and Scores

Past research discussed factors that bias existing in the evaluation of LLMs (Gu et al., 2024), leading us to briefly explore whether action length affects scoring. Figure 5 compared the rendered action clip lengths as evaluated by two scoring sources: human raters and VLLM (GPT-4o-mini).

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

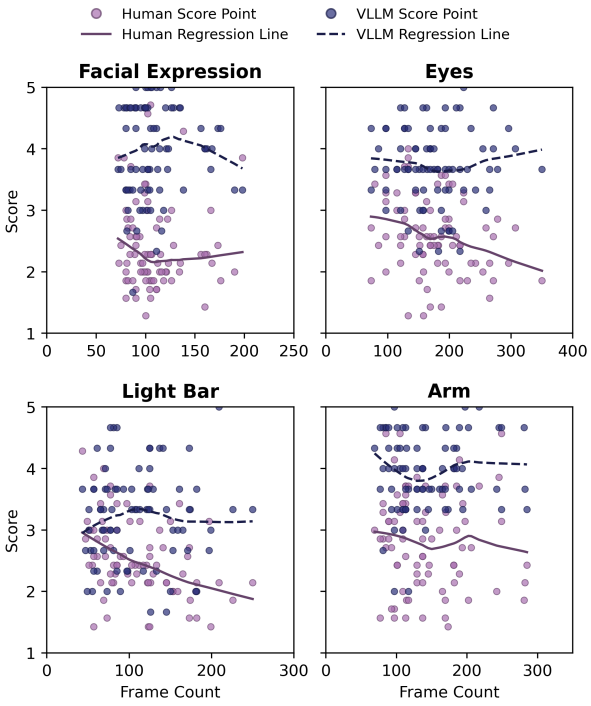


Figure 5: Relationship between action clip length and evaluation scores. The plot compares scores from human raters and the VLLM (GPT-4o-mini). Human raters tend to assign higher scores to shorter clips, whereas the VLLM scores remain relatively unaffected by clip length. Besides, the VLLM consistently gives a higher average score than human raters.

5 Conclusion

We introduced the `eHMI-Action Scoring` dataset with 320 averaged human-rated scores, built on our self-developed asset including eight scenarios, four eHMI modalities, and ten actions per scenario-modality pair. Through three experiments, we answered that **LLM-driven eHMI systems can generate eHMI actions at a human level**. The results also showed that VLLMs are effective for rating eHMI action clips, while reasoning-enabled LLMs prove to be the most suitable for our task. We believe this dataset and our findings provided valuable insights and inspired further research on LLM applications in the eHMI domain.

Limitations

Our research represented a significant step forward in incorporating large language models (LLMs) into the eHMI system; however, challenges remain.

First, one of the main contributions of our paper was the collection of the `eHMI-action scoring` dataset. However, the current study primarily focused on action design for a specific eHMI modality. It would be both interesting and valuable to explore the mixed-eHMI, as each offers distinct advantages for conveying various types of messages (e.g., spatial information or emotional content).

Second, we proposed leveraging the automated scoring system to fine-tune specific LLMs in our future work. We plan to carry out these fine-tuning tasks within a virtual world environment that demands rapid rendering speeds. Unfortunately, our current pipeline requires approximately 20 minutes to render a 10-second clip—a delay that makes it impractical for such applications. To overcome this bottleneck, we aim to accelerate the process by either adopting a more efficient renderer or by rendering only keyframes, thereby reducing the overall time required per clip.

Third, when using VLLMs to score action clips, we sampled one frame every six frames and downsample each to a resolution of 512×512. This method may lead to a loss of detail, potentially undermining the scoring reliability of the VLLMs. In future work, we aim to reduce this information loss to further enhance the reliability of the VLLM raters.

Ethics Statement

All data in `eHMI-Action Scoring` dataset were de-identified and safeguarding privacy concerns. Our data construction processes were carried out by skilled researchers. Participants included students from Chinese and Japanese universities, all of whom receive fair compensation for their contributions.

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.

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.

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.

Vishal Chauhan, Anubhav, Chia-Ming Chang, Jin Nakazato, Ehsan Javanmardi, Alex Orsholits, Takeo Igarashi, Kantaro Fujiwara, and Manabu Tsukada. 2024. Transforming pedestrian and autonomous vehicles interactions in shared spaces: A think-tank study on exploring human-centric designs. In *Adjunct Proceedings of the 16th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 136–142.

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.

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.

Yan Ding, Xiaohan Zhang, Chris Paxton, and Shiqi Zhang. 2023. Task and motion planning with large language models for object rearrangement. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2086–2092. IEEE.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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.

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.

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*.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022a. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International conference on machine learning*, pages 9118–9147. PMLR.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. 2022b. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

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*.

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.

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*.

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.

Max Oudshoorn, Joost de Winter, Pavlo Bazilinskyy, and Dimitra Dodou. 2021. Bio-inspired intent communication for automated vehicles. *Transportation research part F: traffic psychology and behaviour*, 80:127–140.

Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, Andrew Y Ng, et al. 2009. Ros: an open-source robot operating system. In *ICRA workshop on open source software*, volume 3, page 5. Kobe, Japan.

Qwen Team. 2024. Qvq: To see the world with wisdom. https://qwenlm.github.io/blog/qvq-72b-preview/. Accessed: 2025-02-15.

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.

Qwen Team. 2024. Qwq: Reflect deeply on the boundaries of the unknown. https://qwenlm.github.io/blog/qwq-32b-preview/. Accessed: 2025-02-15.

Shu Wang, Muzhi Han, Ziyuan Jiao, Zeyu Zhang, Ying Nian Wu, Song-Chun Zhu, and Hangxin Liu. 2024. Llmˆ 3: Large language model-based task and motion planning with motion failure reasoning. *arXiv preprint arXiv:2403.11552*.

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

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.

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*.

A Cost Analysis

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

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

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

LLM API Calls We utilized APIs from multiple sources:

- For proprietary models (GPT-4o, GPT-4o-mini, GPT-o1, Claude Sonnet 3.5, Gemini 2 Flash), we accessed the APIs available on their official websites, which incurred a total cost of $65.
- For open-source models (such as Deepseek-R1, Deepseek-V3, Llama3.3-70B, Qwen-QvQ-72B, QwQ-32B, etc.), we used services provided by Siliconflow [1] and Aliyun Bailian[2], resulting in an additional cost of $10.

Total The overall cost for the study $535.

B Scenario Prompt (AVs' perspective)

Four first-person scenarios:

Send intention You are an autonomous taxi that receives a ride request and arrives to pick up the passenger (will be on the right roadside of you). Upon arrival, you detect the passenger standing in an area where parking is not permitted within a 5-meter radius. To ensure a safe and legal pickup, you send the following message to the pedestrian: "I am unable to pick you up here. Please walk forward in my direction to a suitable pickup spot."

Status report You are a stopped autonomous vehicle parked near a park, positioned just before a crosswalk. At a specific moment, you plan to start moving. A student is approaching the crosswalk and is about to cross to the other side of the road. Your objective is to get the student's attention and send a message: "I am about to start moving. Please watch out."

Request help You are a delivery robot that has accidentally become trapped by a pile of boxes (or was maliciously pushed). Feeling eager to free yourself and continue delivering the items to your

[1] https://cloud.siliconflow.cn

[2] https://cn.aliyun.com/product/bailian

customer on time, you notice a passerby who sees your situation but hesitates to assist. To encourage him, you send the following message: "I am stuck. Could you please help me?"

Refuse help You are an expensive and fragile delivery robot stuck in the snow due to your low wheels. Your owner is monitoring your status remotely and has decided to rush to the scene for repairs. Meanwhile, a passerby notices you and contemplates whether to offer assistance. To politely inform them, you send the following message: "Thank you for your kindness. Please refrain from touching me."

Two third-person scenarios:

Pedestrian Blind Spot Alert You are a stopped 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 coming from his left (from your right), heading toward the intersection. To ensure his safety, you send the following urgent message to the pedestrian: "Please watch out for the vehicle coming from your left blind spot."

Driver Blind Spot Warning You are a stopped autonomous vehicle parking near an intersection with no traffic lights. A pedestrian is preparing to cross the crosswalk on the opposite side of the road (from your left). A bus traveling in the opposite direction to you is also approaching the intersection. However, the bus's view is obstructed by a building, preventing it from seeing the pedestrian approaching from its right. To ensure the pedestrian's safety, you send the following message to the bus: "Caution: Please watch out for the pedestrian coming from your right blind spot."

Two one-to-many scenarios:

Target Identification You are a delivery robot tasked with delivering a package to a customer in a crowded area. Three individuals are standing in front of you, unaware of who the package is for. Your recipient is standing directly in front of you and is taller than you. To avoid confusion, you send a message to all three: "I am sending the package only to this person."

Broadcast Communication You are a delivery robot carrying a package at a crowded intersection. To navigate through the crowd and turn right without causing disruptions, you broadcast the following message: "I am about to turn right. Kindly make a way to avoid any conflict."

C Additional Dataset Analysis

Figure 6 compared the score distribution of human-rated action scores in our dataset. The results showed that the arm modality most frequently receives scores of 5 (Strongly Agree) and 4 (Agree). In contrast, the facial expression modality most often received a score of 1 (Strongly Disagree), and the eyes modality most often received a score of 2 (Disagree). This finding supported the same hypothesis presented in the second discovery in Section 4.1. This observation might be attributed to the types of messages we designed. In our eight scenarios, the majority required conveying spatial information, where the arm modality is advantageous. Moreover, the absence of emotional messages (e.g., "I am scary") limited the performance of the facial expression modality.

D eHMI description prompts

These system prompts were designed with four sections: character profile, eHMI description, demo actions, and design guidance.

Figure 7 shows the prompt for the eye; Figure 8 displays the prompt for the arm; Figure 9 illustrates the prompt for the light bar; Figure 10 depicts the prompt for facial expressions.

E VLLM rating Prompt

Figure 11 shows the system prompt we use for VLLM raters.

F Survey Screenshots

We provided detailed guidance in our data collection process.

Figure 12 in the introduction page of our survey; Figure 13 is a demo; Figure 14 is an introduction of the next rating scenario; Figure 15 is the page participants used to rating action clips.

G Scenario-Modality pairs visualization

Figure 16 shows screenshots from eight scenario-modality pairs.

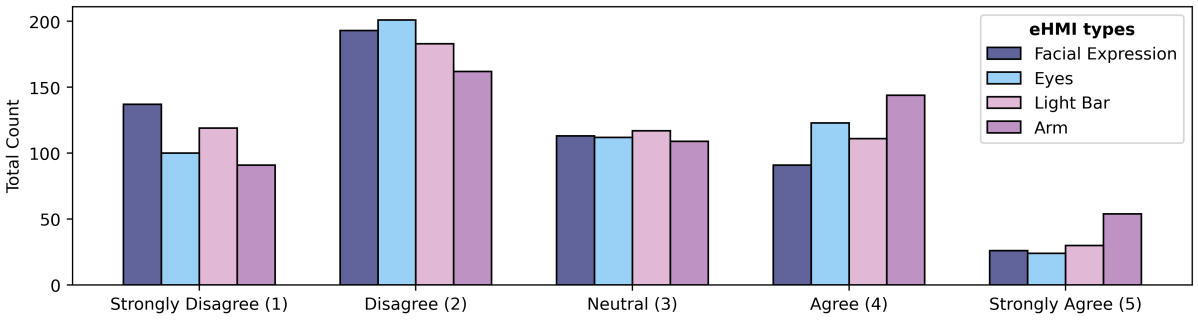


Figure 6: Comparative Distribution of eHMI-Action scores from different eHMI modalities.

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 7: 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 8: 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 9: 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 10: 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.
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 that could indicate specific emotions or messages.
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.
5. Select a Rating
- Based on your analysis, choose one of the following ratings that best reflects the degree of alignment between the eHMI's behavior and the intended message:
1) "Strongly Agree": The eHMI's behavior completely and clearly conveys the intended message.
2) "Agree": The eHMI somewhat conveys the intended message, with only minor discrepancies.
3) "Neutral": The eHMI partially conveys the intended message, resulting in an average overall impression.
4) "Disagree": The eHMI's conveyed message somewhat deviates from the intended message.
5) Strongly Disagree: The eHMI's behavior fails almost entirely to convey the intended message.
6. 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.
- The key details that influenced your selected rating.
- Specific areas where the eHMI's behavior aligned or did not align with the intended message.

Important Notes
1. Strict Evaluation: Base your evaluation solely on the eHMI's behavior as depicted in the video, without being influenced by other aspects of the autonomous system or the broader context.
2. Objectivity: Ensure your judgment remains objective and rigorous like a human. Award higher ratings only when the eHMI's behavior fully aligns with the intended message.
3.*No "Correct" Answers: There are no right or wrong answers. Your evaluation should reflect your intuitive understanding and detailed analysis of the provided materials.

Figure 11: Evaluate prompt for VLLMs.

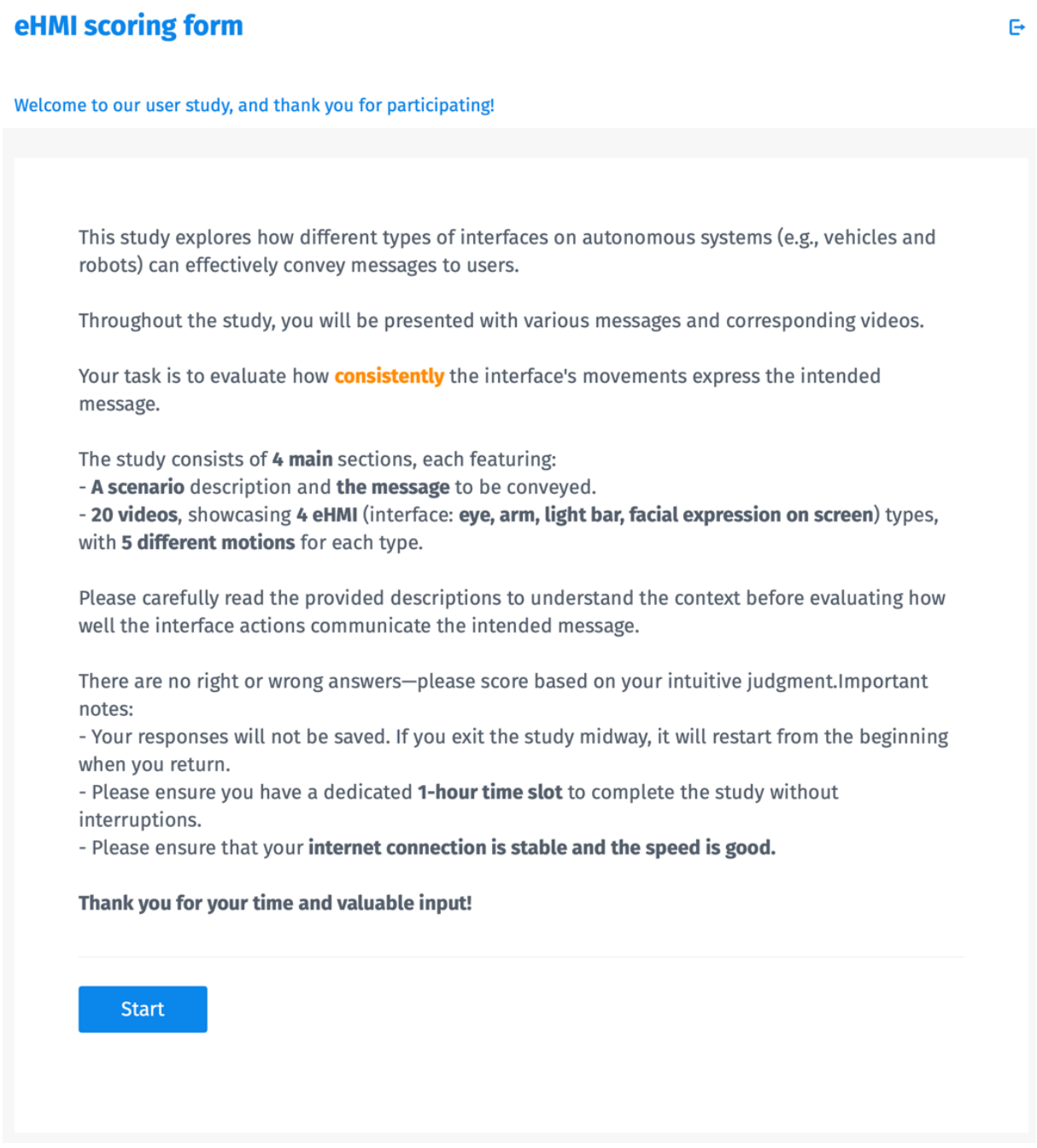


Figure 12: Introduction page of our action scoring survey

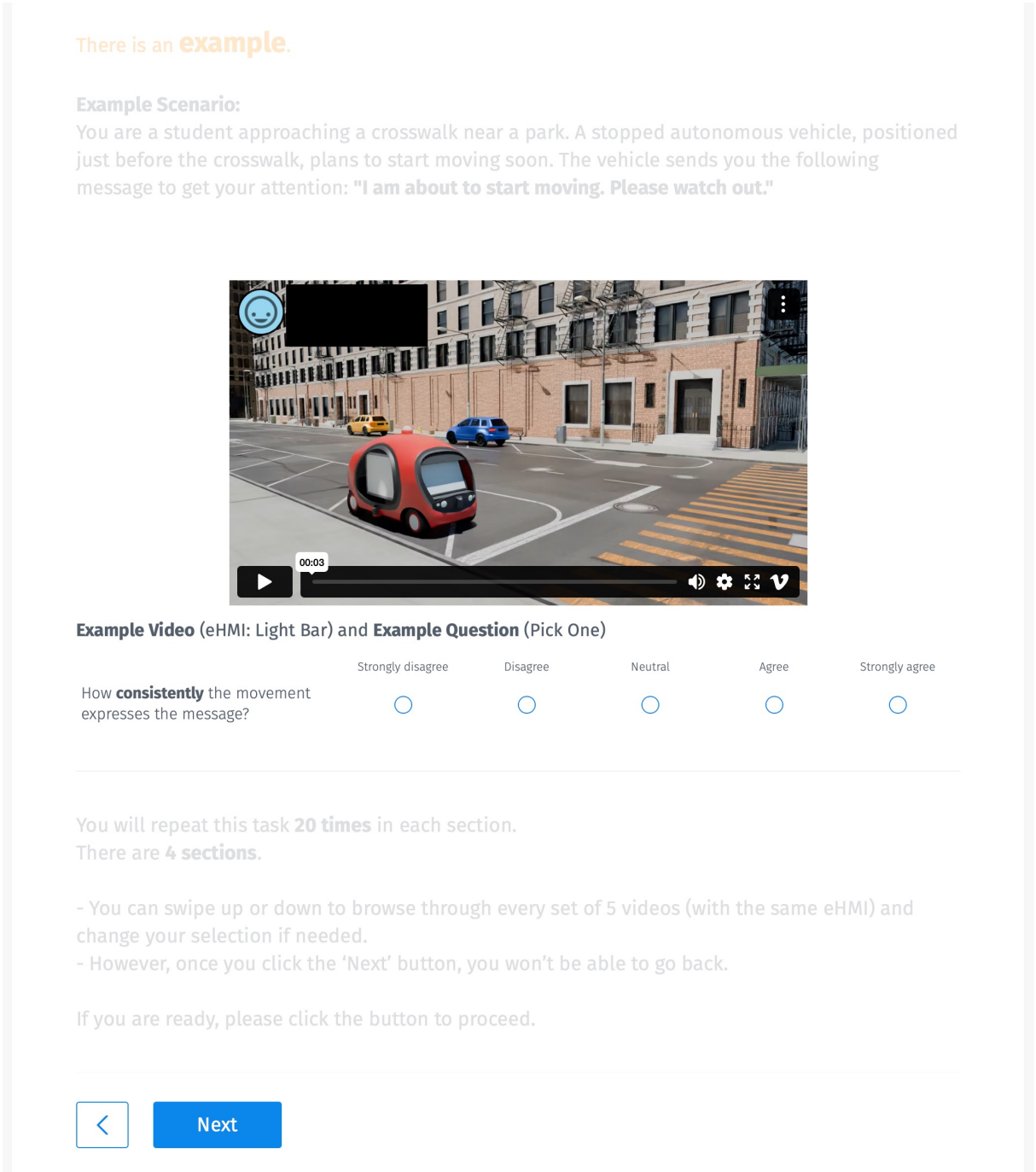


Figure 13: Demo page of our action scoring survey

Section 1:

Scenario:
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."**

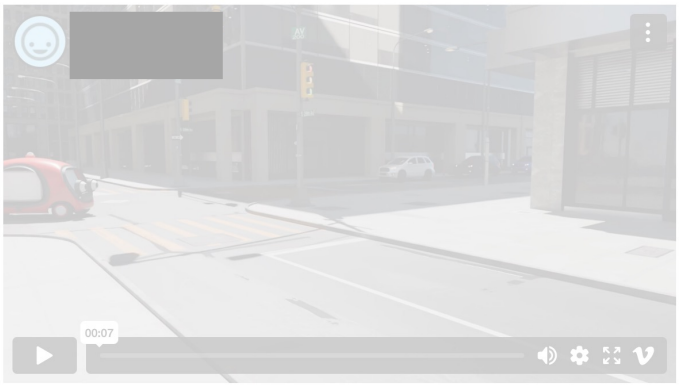
Next

Figure 14: Scenario introduction page of our action scoring survey

Scenario1: eHMI (eyes) Motion No. 3

Message: "I am unable to pick you up here. Please walk forward in my direction to a suitable pickup spot."

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
How **consistently** the movement expresses the message?	○	○	○	○	○

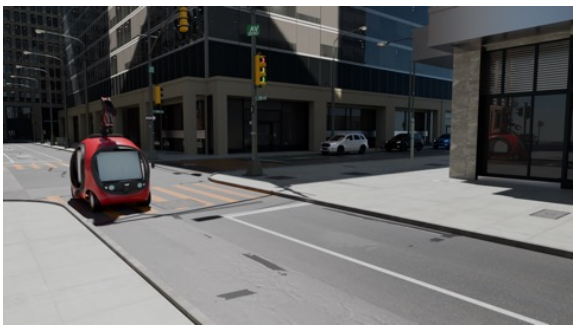


Scenario1: eHMI (eyes) Motion No. 1

Message: "I am unable to pick you up here. Please walk forward in my direction to a suitable pickup spot."

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
How **consistently** the movement expresses the message?	○	○	○	○	○

Figure 15: Participant rating page of our action scoring survey



Scenario 0: Send intention

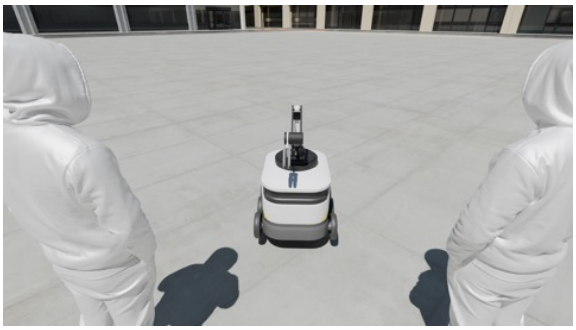
Scenario 1: Status report

Scenario 2: Request help

Scenario 3: Refuse help

Scenario 4: Pedestrian Blind Spot Alert

Scenario 5: Driver Blind Spot Warning



Scenario 6: Target Identification

Scenario 7: Broadcast Communication

Figure 16: Visualization of eight scenario-modality pairs we designed.