# SEE2REFINE: Vision-Language Feedback Improves LLM-Based eHMI Action Designers

Anonymous ACL submission


## Abstract

Automated vehicles lack natural communication channels with other road users, making external Human–Machine Interfaces (eHMIs) essential for conveying intent and maintaining trust in shared environments. However, most eHMI studies rely on developer-crafted message–action pairs, which are difficult to adapt to diverse and dynamic traffic contexts. A promising alternative is to use Large Language Models (LLMs) as action designers that generate context-conditioned eHMI actions, yet such designers lack perceptual verification and typically depend on fixed prompts or costly human-annotated feedback for improvement. We present SEE2REFINE, a human-free, closed-loop framework that uses vision–language model (VLM) perceptual evaluation as automated visual feedback to improve an LLM-based eHMI action designer. Given a driving context and a candidate eHMI action, the VLM evaluates the perceived appropriateness of the action, and this feedback is used to iteratively revise the designer's outputs, enabling systematic refinement without human supervision. We evaluate our framework across three eHMI modalities (lightbar, eyes, and arm) and multiple LLM model sizes. Across settings, our framework consistently outperforms prompt-only LLM designers and manually specified baselines in both VLM-based metrics and human-subject evaluations. Results further indicate that the improvements generalize across modalities and that VLM evaluations are well aligned with human preferences, supporting the robustness and effectiveness of SEE2REFINE for scalable action design. [1]


## 1 Introduction

Automated vehicles (AVs) are expected to reach Level 4 autonomy within a decade (Agrawal et al., 2023; Chen et al., 2023). Without human drivers, AVs usually lack effective communication channels with other road users, such as pedestrians, cyclists, and human drivers (Fagnant and Kockelman, 2015). This gap causes confusion and distrust, leading to traffic issues (Colley et al., 2025). External Human-Machine Interfaces (eHMIs) are emerging as promising solutions to address this communication challenge (Dey et al., 2020a). eHMIs can comprise devices such as mechanical eyes (Chang et al., 2022), mechanical arms (Gui et al., 2024b), and lightbars (Dey et al., 2020b) mounted on the exterior of AVs. These devices serve as communication interfaces that convey messages like "I am stopping". For instance, an eHMI can effectively express directional intent through eye gazing, finger pointing, or illuminating particular sections of a lightbar (Gui et al., 2024a).

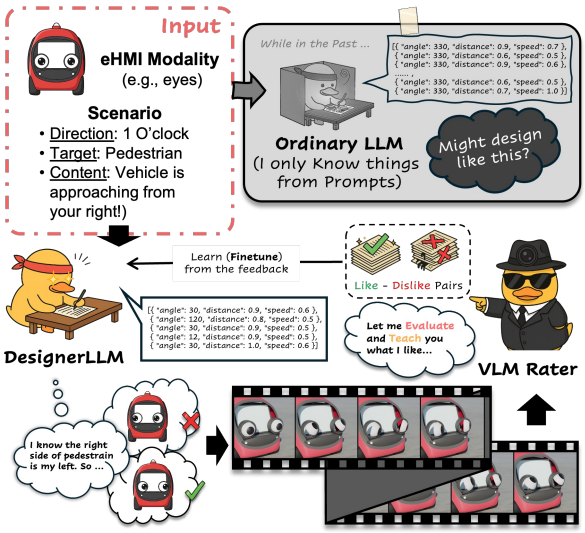



Figure 1: SEE2REFINE uses VLM-based perceptual evaluation as automated visual feedback to design, evaluate, and iteratively refine LLM-based eHMI action designers without human supervision. In contrast, standard LLM-based designers rely on static prompts and lack perceptual grounding for improvement.

[1] The source code, prompts, Blender scenarios, and rendered clips are available at https://anonymous.4open.science/r/eHMI_LLM_finetuning-3A74

With eHMIs, researchers typically design eHMI interactions via predefined actions, limiting adaptability in dynamic and unpredictable environments (Dey et al., 2020a; de Winter and Dodou, 2022). To address this rigidity, LLMs offer a data-driven alternative by generating context-dependent eHMI actions from traffic scenarios, reducing the reliance on manually defined rules (Radford et al., 2019). Moreover, prior work shows that, when equipped with carefully crafted system prompts, LLM-based action designers can perform at a level comparable to that of human designers (Xia et al., 2025). However, text-based prompts can only capture a portion of the details involved in eHMI installation and are unable to convey rich visual information, such as sizes, speeds, and observation angles, which are essential for designing actions in dynamic scenarios (Tellex et al., 2020; Cao et al., 2024; Majumder et al., 2024).

Using human feedback to guide the extraction of knowledge related to specific types of eHMI serves as a viable solution. Reinforcement Learning from Human Feedback (RLHF) (Bai et al., 2022a) has proven effective in developing LLMs at the expert-level in specific tasks. However, applying RLHF to the eHMI LLM designer involves extensive human annotations, similar to other RLHF studies (Kaufmann et al., 2023; Jin et al., 2023; Wang et al., 2025). Reinforcement Learning from AI Feedback (RLAIF) offers an alternative approach by leveraging visual perception capabilities of Vision-Language Models (VLMs) to provide judgments that are close to human-level (Lee et al., 2024a; Lu et al., 2024). To develop a framework capable of designing, evaluating, and self-improving in a cost-effective, rapid, and automated way, we pose the following research question:

> *How to use VLMs' perception feedback to improve an LLM-based eHMI action designer without human annotations?*

In this work, we propose SEE2REFINE that features LLM designers to autonomously design, evaluate, and iteratively refine eHMI actions, utilizing visual perception from VLMs as automated feedback in a cost-effective manner. In our experiments, we trained separate DesignerLLM models for each of the three eHMI modalities: lightbar, eyes, and arm. After three rounds of iterative learning, the DesignerLLM models exhibited a clear alignment in preferences with VLM raters. Additionally, our sampling method to expand the action database effectively reduces annotation costs and time without compromising performance. Subsequently, we recruited 18 participants to evaluate the eHMI actions generated by five LLMs: two 7B models (one base model, one DesignerLLM) and three state-of-the-art models. Extensive results show that utilizing the visual perception of VLMs can also enhance human evaluation scores. The main contributions can be summarized as follows:

- We propose a new SEE2REFINE framework that integrates VLM-based perceptual evaluation as automated visual feedback to improve LLM-based eHMI action designers (DesignerLLM).
- We discuss methods for effectively expanding the action database while maintaining equivalent VLM preference alignment performance.
- We conduct a user study to demonstrate that aligning with VLM raters' preferences can also enhance human evaluation scores.

## 2 Related Works

**eHMI Action Planning.** Existing eHMI action-planning mostly follows predefined, designer-authored rules per modality: text/icons use traffic-rules messages (Eisele and Petzoldt, 2022; Eisma et al., 2021); color/light use flashing patterns based on intuitive color associations (Bazilinskyy et al., 2019; Dey et al., 2020b); and anthropomorphic cues (e.g., eyes, arm gestures) draw on nonverbal communication principles (Mahadevan et al., 2018; Ochiai and Toyoshima, 2011). While effective, these manual designs are labor-intensive and hard to scale to complex real-world scenarios (Gui et al., 2023; de Winter and Dodou, 2022). Recent work uses LLMs for eHMI action planning via prompt engineering (Xia et al., 2025), but this remains prompt-dependent and largely task-level, lacking mechanism-level understanding, requiring more advanced methods for higher-quality action planning.

**VLM-based perceptual evaluation as feedback.** Reinforcement Learning from AI Feedback (RLAIF) replaces the collection of human preferences in RLHF by using stronger teacher models to generate reward signals (Bai et al., 2022b; Lee et al., 2023). In multimodal settings, VLMs can be fine-tuned as rubric-following judges for images/videos with human-level agreement competitive with GPT-4V (Lee et al., 2024a; Lu et al., 2024; Chen et al., 2024). Systems such as Prometheus-Vision, WildVision-Arena/Bench, and MLLM-as-a-Judge show that VLM judges can

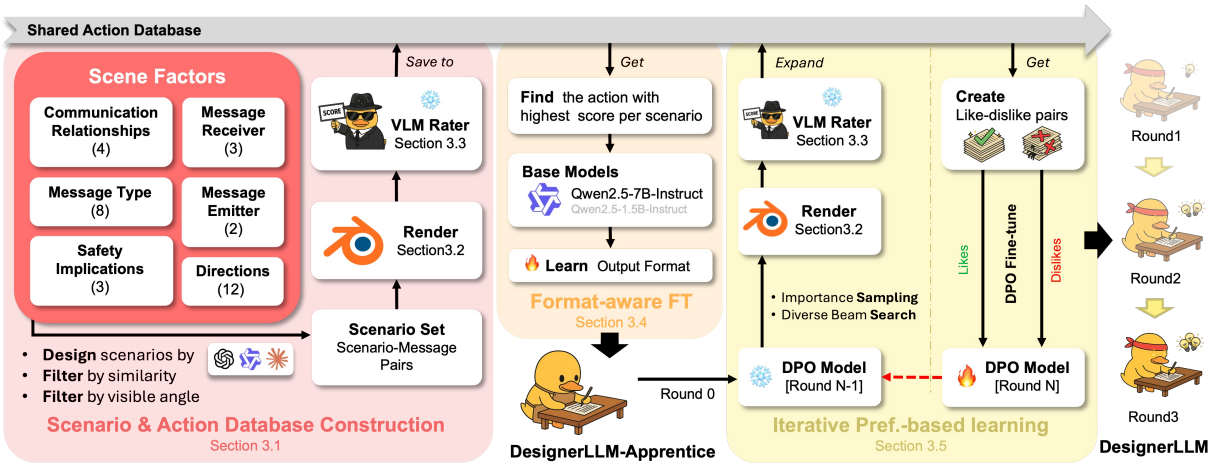



Figure 2: Our SEE2REFINE framework includes: scenario and action database construction (Section 3.1), format-aware fine-tuning (Section 3.4), and iterative preference-based learning (Section 3.5). A shared action database supports all three components, storing generated actions and expanding to enhance DesignerLLM's performance.

follow instructions and output pointwise/pairwise scores that preserve human preference orderings (Lee et al., 2024a; Lu et al., 2024; Chen et al., 2024), while VHELM and LLaVA-Critic further standardize multimodal evaluation and reduce reliance on closed models (Lee et al., 2024b; Xiong et al., 2025). Multimodal variants of RLAIF/RLHF-V use AI-generated preferences to train reward models and optimize LVLM policies with minimal human labeling (Ahn et al., 2024; Yu et al., 2025; Sun et al., 2023). For eHMI, VLM judges correlate well with human ratings on LLM-generated action clips, enabling closed-loop optimization where an LLM proposes actions, a VLM scores them, and updates (e.g., policy gradient or DPO-style) refine the space without human studies (Wang et al., 2024).

## 3 Method

As shown in Figure 2, our SEE2REFINE framework contains (1) Scenario-Message Pair Generation, (2) Model Asset & Action Rendering, (3) Multi-Metric Evaluation System, (4) Format-aware Fine-tuning, and (5) Iterative Preference-based Learning.

### 3.1 Scenario-Message Pair Generation

We first generate traffic scenarios by combining six different factors, inspired by existing eHMI scenario studies (Dey et al., 2020a; Colley and Rukzio, 2020). The factors we consider are listed below. By combining these factors, we generate a total of 6,912 unique condition combinations.

- **Communication Relationships (4):** 1st-/3rd-person perspective × one-to-one/one-to-many.
- **Emitter (2):** Self-driving car; delivery robot.
- **Receiver (4):** Vehicle driver; pedestrian; cyclist; motorcyclist.
- **Message Type (8):** Instruction, advisory, question, answer, current, historical, predictive, warn.
- **Direction (12):** 1 to 12 o'clock.
- **Safety Implications (3):** Critical (0–5 m), moderate (5–10 m), routine (>10 m).

Second, conditioned on these scenarios, we generate 20,736 intended messages containing scenario information using larger LLMs, including OpenAI GPT-4.1 (OpenAI, 2025), Claude 3.7 Sonnet (Anthropic, 2025), and Qwen3-235B-A22B (Qwen Team, Alibaba Cloud, 2025). To eliminate redundancy, we employ a sentence transformer (Reimers and Gurevych, 2019) to compare messages. We sample 70% of these messages using the farthest point sampling algorithm (Eldar et al., 1997) to ensure sufficient variance. In addition, we conduct a small user-rating study on 200 randomly selected scenario–message pairs. Two eHMI researchers independently rate the validity of each pair on a 7-point Likert scale (1=Strongly Disagree, 7=Strongly Agree). The mean rating is 5.3 (SD=1.9), between "Somewhat Agree" (5) and "Agree" (6), suggesting that the generated scenarios were generally of high quality.

Third, we build an initial action database for the later co-learning process. Using the same LLMs as in the second step, we generate two structured action sequences per model for each modality (six candidates per scenario). We then render these actions (Section 3.2) and score them with VLM raters (Section 3.3), saving the scored actions to the shared database.

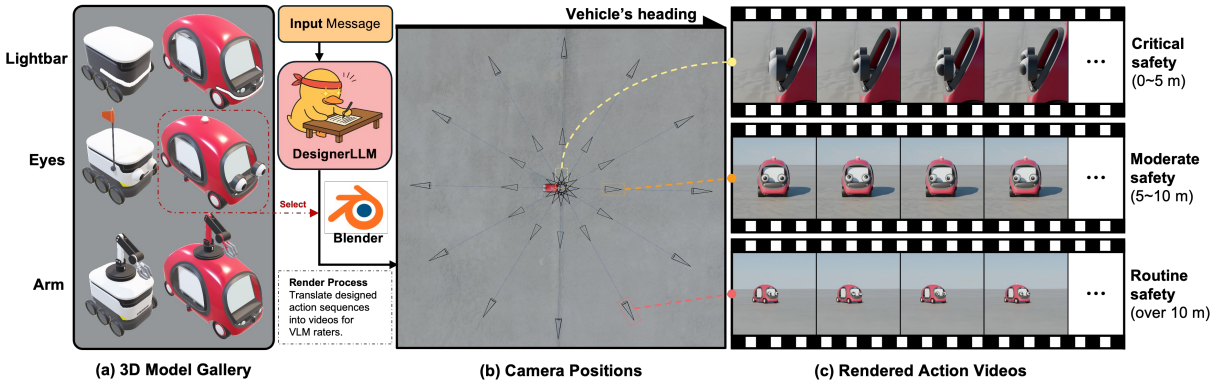



Figure 3: Overview of the six eHMI 3D models combining three modalities (lightbar, eyes, arm) and two emitter types (self-driving car, delivery robot). Rendered videos are generated in Blender from the message receivers' perspective under a defined camera direction and distance.

## 3.2 Model Asset and Action Rendering

### 3.2.1 Model Development and Definition

In this study, we focus on three eHMI modalities: lightbar, eyes, and arm. The first two modalities represent classic eHMI designs (Dey et al., 2020a; Deb et al., 2018; Rouchitsas and Alm, 2019; Benderius et al., 2017; Colley et al., 2020) with simpler control mechanisms, while the arm modality requires managing more parameters, which presents additional challenges for **DesignerLLM**. In Blender (Blender Foundation, 2025), we implemented each modality in two message emitter types (self-driving cars and delivery robots), resulting in six 3D models (Figure 3 (a)). Modality descriptions and permitted actions are as follows:

- **Eyes** are mounted on the front of the vehicle. The pupil is parameterized in polar coordinates: angle in $[0°, 360°]$ (0°=up, counterclockwise) and radius in $[0, 1]$ (0=center, 1=edge) (Chang et al., 2022; Gui et al., 2022).
- **Light Bar** is mounted on the vehicle front, slightly extending to both wings. It has 16 binary lighting regions (0=off, 1=on); e.g., "0000000011111111" denotes left-half off and right-half on (vehicle perspective). Brightness is fixed and color is cyan (Dey et al., 2020b).
- **Arm** is mounted on the vehicle top and consists of five single-axis joints (shoulder, upper arm, forearm, hand, fingers) operating within limited ranges (Gui et al., 2024b).

We also specify transition speeds between action states. The eye and arm transitions range from 0.5 to 2.0s (0.1s steps), while the lightbar transitions range from 0.1 to 1.0s (0.1s per step). This enables smooth sequences while preserving temporal detail at the video FPS.

### 3.2.2 Rendering Pipeline

For rendering, we typically consider three conditions: camera direction, camera distance (relative to the object), and background scenes, as illustrated in Figure 3 (b). Regarding camera direction, aligned with Section 3.1, we create 12 cameras arranged in clockwise order. Regarding camera distance, which reflects the safety implications (Section 3.1), we position the cameras at three different distances. These positions can be accessed via Python to generate videos that simulate the perspectives of other road users. We use a plain background, unlike the previous work (Xia et al., 2025), to avoid contextual cues and reduce rendering time. Removing scene backgrounds decreases rendering time to approximately $1/10$, while lower GPU memory usage also enables more parallel threads, further reducing wall-clock time by around 4× for batch rendering.

We render action videos using Blender version 4.5 (Blender Foundation, 2025) on a GPU-equipped device featuring four NVIDIA GTX Titan cards with 24 GB of RAM each, which allows for up to 16 concurrent threads. The average rendering time for 1,000 clips is approximately one hour, with each thread consuming around 1.1 GB of GPU RAM. Each frame is rendered at a resolution of 512×512 pixels, centered on the eHMI-equipped AV, as shown in Figure 3 (c). These clips are rendered at four FPS, with a maximum action change rate for each eHMI set to ensure that the frame rate captures all relevant details, following Nyquist-Shannon sampling theorem (Nyquist, 1928).

## 3.3 Multi-Metric Evaluation System

Recent studies show that VLMs demonstrate capabilities and preferences similar to those of hu-

mans (Lee et al., 2024a; Lu et al., 2024; Chen et al., 2024). In our task, we aim to borrow knowledge from existing Human-Computer Interaction research (Colley et al., 2020; Colley and Rukzio, 2020; Colley et al., 2025) and use the features of VLMs to develop a multi-metric evaluation system.

To simulate the human perception process, we design a two-phase assessment: Phase 1 (without revealing the intended message) and Phase 2 (with the intended message disclosed). Each phase assesses different aspects using standardized 9-point Likert rating scales.

Phase 1 assesses how effectively the eHMI communicates without prior knowledge of the message, simulating real-world reception scenarios.

- **Intention Recognition (text + certainty, 9-pt):** Infer the message from the animation only; output (i) an interpreted message sentence and (ii) a 9-point certainty score (1=unclear, 5=neutral, 9=unambiguous).
- **Targeting (9-pt):** Rate the confidence that the message is directed to the receiver (1=not for the receiver, 5=uncertain, 9= clear for the receiver).
- **Trust (9-pt):** Rate trust in AV/eHMI (1=distrust/avoid, 5=neutral, 9=full trust).
- **Similarity (9-pt, post-hoc):** Compare the VLM-interpreted message with the intended message (from the emitter scenario) using a small LLM judge (1=contradictory, 5=partially similar, 9=equivalent).

Phase 2 reveals the intended message to the VLM evaluator and assesses how well the observed eHMI behavior aligns with this communication goal. This phase evaluates the quality of the design from an informed perspective.

- **User Acceptance (9-pt):** Willingness to accept the eHMI in daily life given the scenario, intended message, and actions (1=reject, 5=uncertain, 9=fully accept).
- **Consistency (9-pt):** Alignment between the perceived meaning of the actions shown and the revealed intended message (1=contradictory, 5=mixed, 9=fully aligned).

#### 3.3.1 Kernel Score: Composite Quality Metric

We later define a kernel scoring function that aggregates the evaluation metrics as a unified quality indicator. Given the six scores obtained from the VLM evaluation: target score $t$, user acceptance $u$, consistency $c$, certainty $\kappa$, similarity $s$, and trust $\tau$. We compute the kernel score $K$ as:

$$K = (\kappa \times s) + t + \tau + u + c. \quad (1)$$

### 3.4 Format-aware Fine-tuning

This phase trains **DesignerLLM-Apprentice** to reliably output eHMI-specific action formats, as illustrated in Figure 2. We fine-tune two base model sizes, Qwen2.5-7B-Instruct and Qwen2.5-1.5B-Instruct (Qwen Team, 2024) (Qwen Team, 2024) as "apprentices" for later optimization. Training data is sampled from the shared action database by selecting the candidate action with the best kernel score per scenario.

### 3.5 Iterative Preference-based learning

We utilize an iterative Direct Preference Optimization (DPO) approach (Du et al., 2024; Rafailov et al., 2023) to progressively enhance the ability of DesignerLLM-base models to generate high-quality eHMI action designs. Our pipeline involves multiple cycles of sampling, rendering, evaluation, and fine-tuning.

#### 3.5.1 Importance-based Scenario Sampling

Expanding the action database is common in DPO-style iterative training (Du et al., 2024), but regenerating actions for all scenarios is costly, and random sampling can overlook hard cases. We therefore use importance sampling to prioritize scenarios that benefit most from additional generations.

For each scenario $i$, the shared action database stores tuples $(i, m, K_i^m)$, where $m$ indexes the source model/round and $K_i^m$ is the kernel score. From these records, we compute $K_i^{\text{best}} = \max_m K_i^m$, $K_i^{\text{worst}} = \min_m K_i^m$, $\Delta K_i = K_i^{\text{best}} - K_i^{\text{worst}}$, and the candidate count $N_i$. We define the importance score:

$$I_i = \frac{(\Delta K_{\max} - \Delta K_i)\, K_i^{\text{worst}}}{(K_i^{\text{best}})^3} \times 0.5^{n_i}, \quad (2)$$

where $\Delta K_{\max} = \max_j \Delta K_j$ and $n_i = N_i/6 - 1$ (assuming 6 generations of baseline). It prioritizes:

- Low best scores $K_i^{\text{best}}$ (room for improvement)
- Small gaps $\Delta K_i$ (limited diversity)
- High worst scores $K_i^{\text{worst}}$ (easier to improve)
- Fewer previous sampling attempts

We normalize scores such that $\max_i I_i = 1$, then sample 20% of scenarios per round by default.

#### 3.5.2 Diverse Action Generation

For each sampled scenario, we generate diverse eHMI action sequences with the current fine-tuned

model using diverse beam search (Vijayakumar et al., 2016) (six beam groups with one beam each). Each output is validated against modality-specific JSON schemas (eyes/arm/lightbar) via Pydantic, and invalid formats are discarded. We then render and evaluate the valid actions using the same pipeline as the initial database, compute their kernel scores, and append them to the shared action database. Repeating this step iteratively expands the candidate set and improves both diversity and quality over training rounds.

#### 3.5.3 Preference Pair Construction

After expanding the shared score database, we form DPO preference pairs $(y^+, y^-)$ per scenario, where $y^+$ is a higher-quality action than $y^-$. This setup is analogous to human preference ordering, as discussed in prior work (Xia et al., 2025). Pair extraction is two-stage. (1) **Max–min:** for each scenario $i$, select the highest and lowest-scoring actions,

$$(y_i^+, y_i^-) = (\arg\max_m K_i^m,\ \arg\min_m K_i^m), \quad (3)$$

only if the score gap $\Delta K_i \geq \delta_{\min}$ (default $\delta_{\min} = 4.0$), to ensure a strong preference signal. (2) **High-gap extras:** collect additional pairs within the scenario with $K_i^{m_1} - K_i^{m_2} \geq \delta_{\min}$, sort by gap, and keep the top $p\%$ (default $30\%$) to enhance training.

## 4 Experiment

### 4.1 Experimental Settings

**Devices and Software.** LLM inference and video rendering run locally on a 4×RTX TITAN (24GB GPU RAM each) machine. Fine-tuning (Format-aware Fine-tuning; Iterative Preference-based learning) is performed on cloud GPUs with up to 2×A100 (40GB GPU RAM each). We use Transformers for inference (Wolf et al., 2020) and LLaMA-Factory for fine-tuning (Zheng et al., 2024). VLM inference is performed remotely using GPT-5-mini (OpenAI, 2025).

**Training Setup.** All finetuning processes use ShareGPT-formatted data (Chiang et al., 2023) with an 80/20 train/test split. We fine-tune with LoRA (Hu et al., 2022) in bfloat16: 3 epochs for Format-aware Fine-tuning and 1 epoch for DPO (to reduce overfitting, similar to prior works (Bai et al., 2022b; Touvron et al., 2023; Liu et al., 2024)). DPO uses weight=1, sigmoid loss, and a preference coefficient of 0.01.

**Computation Cost.** VLM rating achieves approximately 90 evaluations per second with 8 threads. Single-thread inference requires about 7 seconds per scenario for the 7B model and 6 seconds per scenario for the 1.5B model. The rendering takes roughly 45 seconds per clip, operating in batch parallelization. Fine-tuning duration is approximately 3 hours for the 7B model and 1 hour 10 minutes for the 1.5B model in Format-aware Fine-tuning, and around 5 hours (7B) or 3 hours 30 minutes (1.5B) per DPO round. Excluding dataset construction and assuming three DPO rounds, the total co-learning time amounts to at least 60 hours for the 7B model and 38 hours for the 1.5B model.

### 4.2 Results

#### 4.2.1 Alignment with VLM Rater Preferences.

We evaluate how well DesignerLLM aligns with VLM rater preferences to validate the effectiveness of our framework. We compare DesignerLLM-7B with DesignerLLM-1.5B, DesignerLLM-Apprentice-7B, the Qwen2.5-7B-Instruct base and the initial action database (Section 3.1). We report all learning metrics and a diversity measure, adapted from the Action Reference Score (ARS) (Xia et al., 2025), which assesses the average similarity between pairs of action sequences. Notably, these five models use three different prompts, as described in the caption of Table 1.

**Preference alignment.** Table 1 shows that all DesignerLLM variants outperform the initial action database on all metrics, indicating a successful alignment with VLM raters. Gains are larger for Phase 2 metrics (acceptance, consistency) than Phase 1 metrics (targeting, trust, similarity), with trust and similarity improving the least (also discussed in Section 4.2.3).

**Effect of multi-round learning.** Comparing the post-SFT base models (DesignerLLM-base) with the final models (DesignerLLM) confirms that iterative preference-based learning is necessary to further distill VLM preferences into DesignerLLM.

#### 4.2.2 Ablation Studies

We conduct ablations to (1) test whether Scenario Sampling and Diverse Action Generation improve the speed–performance trade-off, and (2) track DesignerLLM performance across different rounds.

**Module effectiveness.** We remove each module while fixing the sampling ratio at 20% to keep training tractable (full regeneration would exceed 13

| Modality | Source / Configuration | VLM rater Metrics | | | | | | Div. ↑ |
|---|---|---|---|---|---|---|---|---|
| | | UA ↑ | Cons. ↑ | Targ. ↑ | Trust ↑ | Sim. ↑ | K.S. ↑ | |
| Lightbar | DesignerLLM-7B† (Full) | **5.889** | **6.061** | **6.577** | **6.874** | **0.373** | **28.017** | **63.580** |
| | DesignerLLM-1.5B† | 5.725 | 5.898 | 6.516 | 6.865 | 0.351 | 27.473 | 63.888 |
| | DesignerLLM-Apprentice-7B† | 4.847 | 4.763 | 6.273 | 6.796 | 0.377 | 25.382 | 61.974 |
| | Qwen2.5-7B-Instruct* | 4.382 | 4.125 | 5.909 | 6.705 | 0.362 | 23.759 | 53.441 |
| | Initial Action Database‡ | 4.444 | 4.203 | 6.197 | 6.784 | 0.365 | 24.244 | 58.378 |
| | *w/o Div.Beam* | 5.011 | 5.234 | 6.377 | 6.806 | 0.370 | 24.534 | 56.332 |
| | *w/o Imp.Samp.* | 5.097 | 5.386 | 6.439 | 6.839 | 0.372 | 24.884 | 54.034 |
| Eyes | DesignerLLM-7B† (Full) | **6.010** | **6.083** | **6.929** | **6.925** | **0.432** | **28.839** | **8.492** |
| | DesignerLLM-1.5B† | 5.828 | 5.919 | 6.773 | 6.901 | 0.421 | 28.246 | 8.229 |
| | DesignerLLM-Apprentice-7B† | 5.461 | 5.533 | 6.598 | 6.813 | 0.400 | 27.158 | 7.242 |
| | Qwen2.5-7B-Instruct* | 5.171 | 5.042 | 6.629 | 6.869 | 0.435 | 26.642 | 10.726 |
| | Initial Action Database‡ | 5.110 | 5.103 | 6.455 | 6.843 | 0.413 | 26.269 | 7.770 |
| | *w/o Div.Beam* | 5.633 | 5.649 | 6.706 | 6.908 | 0.428 | 27.747 | 7.812 |
| | *w/o Imp.Samp.* | 5.706 | 5.732 | 6.761 | 6.900 | 0.424 | 27.936 | 7.976 |
| Arm | DesignerLLM-7B† (Full) | **6.569** | **6.974** | **7.354** | 6.759 | **0.428** | **30.531** | 14.721 |
| | DesignerLLM-1.5B† | 6.495 | 6.820 | 7.313 | **6.768** | 0.423 | 30.218 | 14.896 |
| | DesignerLLM-Apprentice-7B† | 6.109 | 6.362 | 7.156 | 6.614 | 0.413 | 29.128 | 15.022 |
| | Qwen2.5-7B-Instruct* | 5.202 | 5.156 | 6.807 | 6.686 | 0.412 | 26.640 | 10.891 |
| | Initial Action Database‡ | 5.935 | 6.090 | 7.142 | 6.730 | 0.425 | 28.759 | **16.282** |
| | *w/o Div.Beam* | 6.359 | 6.622 | 7.237 | 6.732 | 0.417 | 29.766 | 16.920 |
| | *w/o Imp.Samp.* | 6.366 | 6.594 | 7.233 | 6.755 | 0.425 | 29.694 | 16.788 |

Table 1: Comprehensive VLM evaluation and ablation results. ‡ denotes the structured prompt with explicit format guidance; † is a simplified version of ‡ without format guidance; ∗ refers to an enhanced prompt adapted from prior work (Xia et al., 2025). For ablation, we show the removal of Scenario Sampling (Imp.Samp.) and Diverse Action Generation (Div.Beam) modules.

days). As shown in Table 1, omitting either module causes a clear performance drop, especially for the lightbar modality, where the decline is roughly twice that of the other modalities on average.

**Across-round behavior.** We further compare DesignerLLM-7B across 3 rounds (Table 2), adding *Formatting Errors* (F.Err.) to quantify output validity. Notably, lightbar shows degraded format accuracy after finetuning, likely because its 16-bit status strings (e.g., "0000000011111111") encounter issues with continuous sequences of digits in current LLMs (Spathis and Kawsar, 2023; Hugging Face, 2024).

#### 4.2.3 Metric-wise Analysis

Metric choices are inherited from prior eHMI work, but their effectiveness for VLM raters is unclear. Therefore, we analyze the VLM score distributions for three milestones: (1) the initial action database, (2) DesignerLLM-Apprentice-7B, and (3) DesignerLLM-7B. As illustrated in Figure 7 in the Appendix, metrics behave differently over training. Phase 2 metrics (User Acceptance, Consistency) improve most clearly: their means increase and distributions tighten, suggesting that they are the most sensitive and reliable signals for VLM-based evaluation. In contrast, Phase 1 metrics show weaker trends; targeting improves slightly, while trust changes little and provides poor preference separation in the initial database, making it less useful for forming training pairs. This likely reflects a mismatch between VLM- and human-perceived trust in our setting.

### 4.3 Human Preference Alignment

We conducted a user study to (1) test whether improvements transfer from VLM raters to people, and (2) examine potential gaps between VLM and human preferences. Details are in Appendix A.

**Setup.** We evaluated two common modalities (eyes, lightbar) on eight scenario-message pairs from prior work (Xia et al., 2025). We compared five models: DesignerLLM-base, DesignerLLM, and three commercial LLMs (OpenAI GPT-5, Claude Opus 4.1, Grok 4). Each model generated actions that were rendered as 1080p videos at 12 FPS with full scene context (80 videos total).

**Participants and measures.** We recruited 18 par-

| Modality | Stage | VLM rater Metrics | | | | | | Div. ↑ | F.Err. (%) ↓ |
|---|---|---|---|---|---|---|---|---|---|
| | | UA ↑ | Cons. ↑ | Targ. ↑ | Trust ↑ | Sim. ↑ | K.S. ↑ | | |
| Lightbar | Apprentice | 4.847 | 4.763 | 6.273 | 6.796 | **0.377** | 25.382 | 61.974 | **<0.01** |
| | Round1 | 5.673 | 5.888 | 6.526 | 6.826 | 0.352 | 27.425 | **66.628** | 6.06 |
| | Round2 | **5.920** | **6.152** | **6.637** | 6.848 | 0.359 | **28.110** | 62.983 | 10.06 |
| | Round3 | 5.889 | 6.061 | 6.577 | **6.874** | 0.373 | 28.017 | 63.580 | 11.31 |
| Eyes | Apprentice | 5.461 | 5.533 | 6.598 | 6.813 | 0.400 | 27.158 | 7.252 | 0 |
| | Round1 | 5.707 | 5.742 | 6.733 | 6.733 | 0.423 | 27.917 | 8.229 | 0 |
| | Round2 | 5.834 | 5.921 | 6.790 | 6.912 | 0.427 | 28.305 | **8.606** | 0 |
| | Round3 | **6.010** | **6.083** | **6.929** | **6.925** | **0.432** | **28.839** | 8.492 | 0 |
| Arm | Apprentice | 6.109 | 6.362 | 7.156 | 6.614 | 0.413 | 29.128 | 15.022 | 0 |
| | Round1 | 6.377 | 6.604 | 7.235 | 6.754 | 0.426 | 29.826 | **16.458** | 0 |
| | Round2 | 6.471 | 6.743 | 7.301 | 6.755 | **0.432** | 30.168 | 15.197 | 0 |
| | Round3 | **6.569** | **6.974** | **7.354** | **6.759** | 0.428 | **30.531** | 14.721 | <0.01 |

Table 2: Different Round Performance and Output Format Error Comparison.

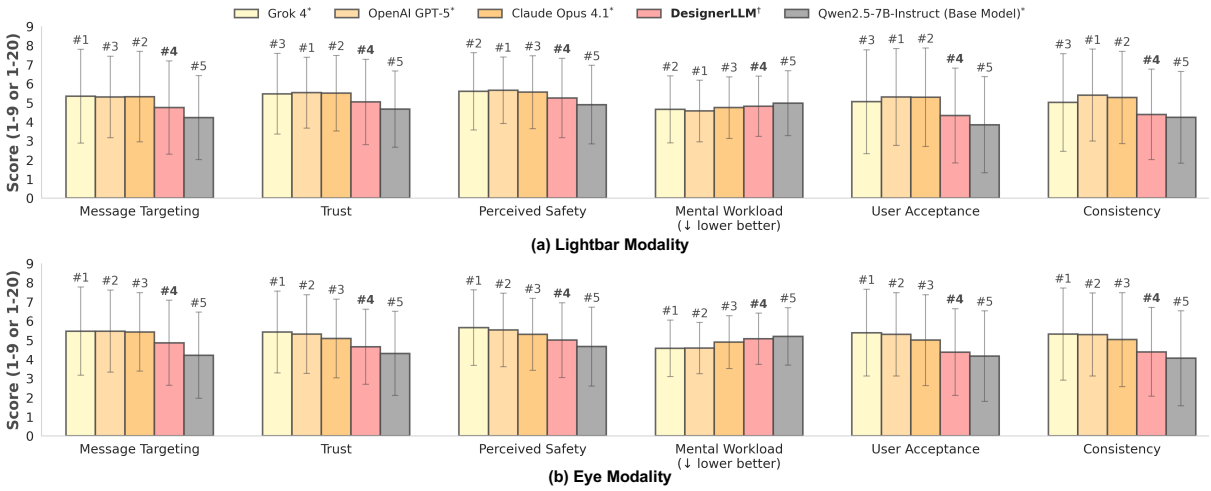



Figure 4: Human ratings of five models on six subjective metrics for (a) eyes and (b) lightbar. Bars show mean scores (1–9) with error bars; mental workload is rescaled from 1–20 to 1–9. Ranks are shown above bars (#1 best). DesignerLLM (7B) ranks 4th for both modalities, outperforming the base model and approaching commercial models. † denotes a simple eHMI description; ∗ denotes an enhanced prompt adapted from (Xia et al., 2025).

ticipants (21–30 years, $M$=26, $SD$=2.7), split evenly by modality. For each scenario, participants rated videos in two stages: without the intended message (targeting, trust, perceived safety, mental workload) and with the intended message revealed (acceptance, consistency).

**Results.** Across both modalities, DesignerLLM consistently outperformed the base model and ranked 4th overall (Figure 4). Relative to the base model, DesignerLLM improved average scores by 7.7% (eyes) and 7.9% (lightbar), with the largest gains in targeting (15.4% eyes; 12.6% lightbar) and acceptance (5.0% eyes; 12.5% lightbar). Commercial models achieved higher absolute ratings, but the improvements from our framework were statistically meaningful for most metrics, indicating that VLM-aligned training also yields measurable benefits for human-perceived eHMI quality.

## 5 Conclusion

We introduce SEE2REFINE leveraging VLM perceptual feedback as an automated approach to progressively enhance the capabilities of LLM-based eHMI action designers. We perform extensive experiments and explore methods for effectively expanding the action dataset while preserving alignment with VLM preference performance. Additionally, we conduct a user study demonstrating that insights from VLM raters can also improve human perception scores. Finally, our work serves as an important practice in improving the message delivery capabilities of LLM-actuated systems.

## Limitations

**Collaborative Learning with Human Users.** The current setup demonstrates that, with the SEE2REFINE system, LLM-based eHMI action designers can achieve higher human perception scores, despite using VLM as an efficient substitute for real human annotators. However, as noted in previous work (Xia et al., 2025), the preference order between a human and VLM is consistent only when the rating score difference between two actions is sufficiently large. Replicating the nuanced preferences of individuals from different cultural backgrounds or age groups remains challenging. Therefore, incorporating a real human annotator into the refinement process would be valuable.

There are two aspects worth exploring. First, effective fine-tuning of VLM rater models could further improve system performance. However, collecting enough data for a single round of fine-tuning is still costly, so the potential marginal gains from such fine-tuning merit careful consideration. Second, rating alone is often not an information-rich method for annotation. If we consider our system as a simulation of a real teacher, employing a teaching approach, using rich cues such as voice, gestures, posture, and other non-verbal signals, could serve as a more natural and effective way to teach LLM designers accurate scenario-action pairs. This approach could also help reduce the ambiguity associated with preference pairs, thereby enabling more effective distillation of human preferences into the LLM training process.

**Real-World Simulation and Learning.** In the current implementation (Section 3.1), we utilize generated scenarios and rendered action videos as visual cues for VLM raters. This approach has proven to be both effective and efficient. However, it primarily focuses on simulating the main elements of typical traffic scenarios, such as automated vehicles, while overlooking other factors that could influence the visual perception outputs for both VLM raters and human road users, such as lighting conditions, obstructions, and other environmental factors. Therefore, the challenge is to incorporate more visual elements that accurately simulate real-world scenarios in rendered videos, while maintaining reasonable rendering times to enhance the robustness of LLM-based action designers in practical applications. One potential trade-off solution is to combine different levels of visual detail within the same training session. For example, including videos with rich visual elements for only a small portion of the training data could improve robustness while keeping overall training costs relatively low.

## References


Shubham Agrawal, Amy M Schuster, Noah Britt, Elizabeth A Mack, Michael L Tidwell, and Shelia R Cotten. 2023. Building on the past to help prepare the workforce for the future with automated vehicles: A systematic review of automated passenger vehicle deployment timelines. *Technology in Society*, 72:102186.

Daechul Ahn, Yura Choi, Youngjae Yu, Dongyeop Kang, and Jonghyun Choi. 2024. Tuning large multimodal models for videos using reinforcement learning from ai feedback. *arXiv preprint arXiv:2402.03746*.

Anthropic. 2025. Claude 3.7 sonnet and claude code. Accessed: 2025-11-02.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

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.

Ola Benderius, Christian Berger, and Victor Malmsten Lundgren. 2017. The best rated human–machine interface design for autonomous vehicles in the 2016 grand cooperative driving challenge. *IEEE Transactions on intelligent transportation systems*, 19(4):1302–1307.

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

Bowen Cao, Deng Cai, Zhisong Zhang, Yuexian Zou, and Wai Lam. 2024. On the worst prompt performance of large language models. *Advances in Neural Information Processing Systems*, 37:69022–69042.

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.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*.

Long Chen, Yuchen Li, Chao Huang, Yang Xing, Daxin Tian, Li Li, Zhongxu Hu, Siyu Teng, Chen Lv, Jinjun Wang, and 1 others. 2023. Milestones in autonomous driving and intelligent vehicles—part i: Control, computing system design, communication, hd map, testing, and human behaviors. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 53(9):5831–5847.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

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.

Mark Colley, Marcel Walch, Jan Gugenheimer, Ali Askari, and Enrico Rukzio. 2020. Towards inclusive external communication of autonomous vehicles for pedestrians with vision impairments. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

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

Shuchisnigdha Deb, Lesley J Strawderman, and Daniel W Carruth. 2018. Investigating pedestrian suggestions for external features on fully autonomous vehicles: A virtual reality experiment. *Transportation research part F: traffic psychology and behaviour*, 59:135–149.

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.

Yuhao Du, Shunian Chen, Wenbo Zan, Peizhao Li, Mingxuan Wang, Dingjie Song, Bo Li, Yan Hu, and Benyou Wang. 2024. Blenderllm: Training large language models for computer-aided design with self-improvement. *Preprint*, arXiv:2412.14203.

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.

Yuval Eldar, Michael Lindenbaum, Moshe Porat, and Yehoshua Y Zeevi. 1997. The farthest point strategy for progressive image sampling. *IEEE transactions on image processing*, 6(9):1305–1315.

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.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Hugging Face. 2024. Number tokenization blog. https://huggingface.co/spaces/huggingface/number-tokenization-blog.

Di Jin, Shikib Mehri, Devamanyu Hazarika, Aishwarya Padmakumar, Sungjin Lee, Yang Liu, and Mahdi Namazifar. 2023. Data-efficient alignment of large language models with human feedback through natural language. *arXiv preprint arXiv:2311.14543*.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2023. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925*.

Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and 1 others. 2023. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Seongyun Lee, Seungone Kim, Sue Park, Geewook Kim, and Minjoon Seo. 2024a. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. In *Findings of the association for computational linguistics ACL 2024*, pages 11286–11315.

Tony Lee, Haoqin Tu, Chi H Wong, Wenhao Zheng, Yiyang Zhou, Yifan Mai, Josselin S Roberts, Michihiro Yasunaga, Huaxiu Yao, Cihang Xie, and 1 others. 2024b. Vhelm: A holistic evaluation of vision language models. *Advances in Neural Information Processing Systems*, 37:140632–140666.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Yujie Lu, Dongfu Jiang, Wenhu Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. 2024. Wildvision: Evaluating vision-language models in the wild with human preferences. *Advances in Neural Information Processing Systems*, 37:48224–48255.

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.

Subir Majumder, Lin Dong, Fatemeh Doudi, Yuting Cai, Chao Tian, Dileep Kalathil, Kevin Ding, Anupam A Thatte, Na Li, and Le Xie. 2024. Exploring the capabilities and limitations of large language models in the electric energy sector. *Joule*, 8(6):1544–1549.

H. Nyquist. 1928. Certain topics in telegraph transmission theory. *Transactions of the American Institute of Electrical Engineers*, 47(2):617–644.

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. 2025. Introducing gpt-4.1 in the api. Accessed: 2025-11-02.

Qwen Team. 2024. Qwen2.5 technical report. https://qwenlm.github.io/blog/qwen2.5/. Accessed: 2025-11-02.

Qwen Team, Alibaba Cloud. 2025. Qwen3-235b-a22b: Hybrid reasoning MoE model. Accessed: 2025-11-02.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Alexandros Rouchitsas and Håkan Alm. 2019. External human–machine interfaces for autonomous vehicle-to-pedestrian communication: A review of empirical work. *Frontiers in psychology*, 10:2757.

Dimitris Spathis and Fahim Kawsar. 2023. The first step is the hardest: Pitfalls of representing and tokenizing temporal data for large language models. *arXiv preprint arXiv:2309.06236*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. 2020. Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems*, 3(1):25–55.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Binghai Wang, Runji Lin, Keming Lu, Le Yu, Zhenru Zhang, Fei Huang, Chujie Zheng, Kai Dang, Yang Fan, Xingzhang Ren, and 1 others. 2025. Worldpm: Scaling human preference modeling. *arXiv preprint arXiv:2505.10527*.

Yufei Wang, Zhanyi Sun, Jesse Zhang, Zhou Xian, Erdem Biyik, David Held, and Zackory Erickson. 2024. Rl-vlm-f: Reinforcement learning from vision language foundation model feedback. In *Proceedings of the 41th International Conference on Machine Learning*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Ding Xia, Xinyue Gui, Fan Gao, Dongyuan Li, Mark Colley, and Takeo Igarashi. 2025. Automating ehmi action design with llms for automated vehicle communication. *arXiv preprint arXiv:2505.20711*.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. 2025. Llava-critic: Learning to evaluate multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13618–13628.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, and 1 others. 2025. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19985–19995.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.

# A User Study Details

In this section, we further evaluate the preference alignment of DesignerLLM models with human participants for two purposes: (1) to demonstrate that our co-learning framework can also enhance the perceptual experience of human users, and (2) to compare and analyze differences in preferences between VLMs and human raters, thereby exploring potential directions for future research.

## A.1 Study Design and Procedure

### A.1.1 Experimental Setup.

We selected two eHMI modalities (lightbar and eyes) as simpler and common modalities for the user study. Eight traffic scenario–message pairs were adopted from prior work (Xia et al., 2025) covering common interactions between AVs and other road users. Five LLMs were tested: two 7B-parameter models (DesignerLLM-base, trained after *Format-aware Fine-tuning*, and DesignerLLM, after *Iterative Preference-based Co-learning*) and three state-of-the-art models (OpenAI GPT-5, Claude Opus 4.1, and Grok 4). Each model generated eHMI actions that were rendered into 1080p (1980×1080) videos at 12 FPS to ensure smooth playback for participants in Blender. Unlike the videos provided to VLM raters (Section 3.2.2), these videos include rich contextual information. A total of 80 videos were created. Rendering all scenarios with detailed meshes and visual effects takes approximately 10 hours of computation.

### A.1.2 Interface Design.

The user study was implemented and deployed using Gradio. Separate interfaces were created for the two eHMI modalities. Each interface consisted of five components: (1) a welcome page, (2) an introduction to general eHMI concepts and the specific modality, including two demonstration videos, (3) a demographic information section, (4) main rating pages for eight scenario–message pairs, and (5) an ending page. The interface included a session-based resume function, allowing participants to continue later using their assigned ID.

### A.1.3 Data Collection and Flow.

In the demographic section, participants anonymously reported their age, gender, and familiarity with eHMIs. For the main user rating pages, eight scenarios are evaluated across different pages, each presented in two stages: the first stage, where the

intended messages are not provided, and the second stage, where the intended messages are included. This results in a total of 16 pages. On each page, five videos are shown to participants in a randomized order. As illustrated in Figure 5, each video assessment follows a horizontal layout, with the video on the left and the questionnaires on the right. To ensure no scores are missed, a verification mechanism is in place: participants must complete the current page before proceeding to the next, ensuring the completeness of the questionnaires. Each session took approximately 5 minutes to complete.

### A.2 Participants

We recruited 18 participants aged 21 to 30 (mean age: 26, SD = 2.7), consisting of nine males and nine females. They were evenly assigned to evaluate the lightbar and eye modalities. Among the participants, one was highly familiar with eHMIs, possessing extensive knowledge or experience; two had some familiarity, having heard about eHMIs before; and the remaining fifteen were unfamiliar with eHMIs. All participants received a $10 gift card as compensation.

### A.3 Measurements

To assess the quality of eHMI actions generated by different LLMs, participants rated each video using standardized perceptual and cognitive metrics commonly used in eHMI studies (Colley et al., 2020; Colley and Rukzio, 2020; Colley et al., 2025). Two measurement sets were used, depending on whether the intended message of the eHMI was provided.

**Without Intended Message.** Participants rated their initial impressions of the eHMI using four metrics:

- **Message Targeting (1–9):** The extent to which the eHMI message was directed toward the participant.
- **Trust (1–9):** The participant's confidence in and willingness to rely on the eHMI communication.
- **Perceived Safety (1–9):** The participant's sense of safety or anxiety during the interaction.
- **Mental Workload (1–20):** The cognitive effort required to interpret the eHMI actions.

**With Intended Message.** After revealing the intended messages, participants reassessed using two metrics:

- **User Acceptance (1–9):** The willingness to accept the eHMI in daily life.
- **Consistency (1–9):** The perceived alignment between the displayed eHMI motion and its intended message.

### A.4 Result Analysis

Prior to analysis, normality of these participant-level means was assessed with the Shapiro–Wilk test and showed no serious deviations ($p > 0.05$) for both modalities. Because the study used a within-subjects design with five model conditions, each metric was analyzed using repeated-measures ANOVA.

#### A.4.1 Reliability Analysis

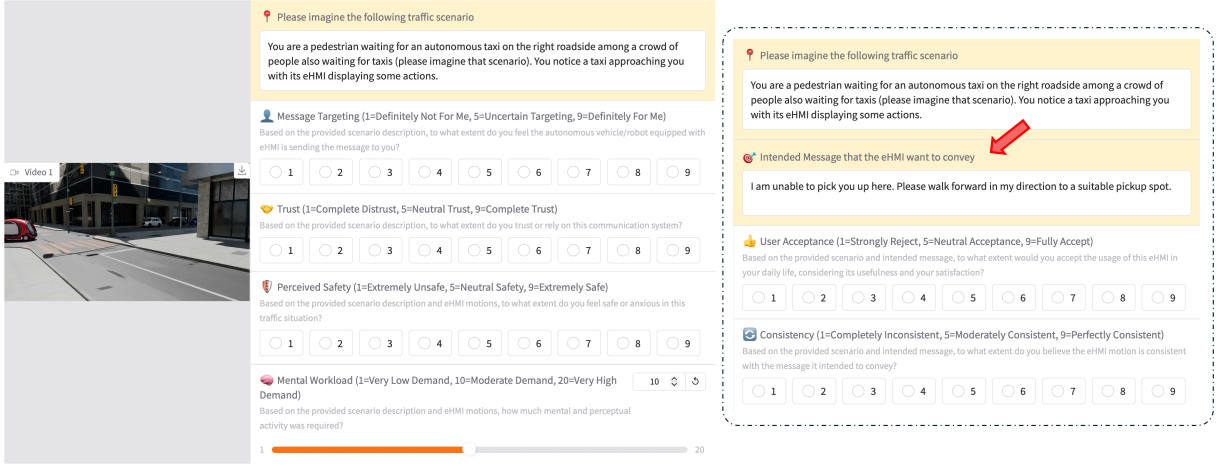
We assessed the reliability of the subjective ratings before analyzing model effects. Cronbach's alpha was used to evaluate internal consistency within each metric across the five models. Internal consistency was excellent for both modalities (eye: $\alpha = 0.925$; lightbar: $\alpha = 0.953$), with all metrics exceeding $\alpha = 0.90$. Inter-rater reliability, estimated with the intraclass correlation coefficient (ICC), was low (eye: $ICC = 0.131$; lightbar: $ICC = 0.087$), indicating limited absolute agreement across participants. Because our analyses focus on within-subject, relative comparisons in a single session, high internal consistency is the primary requirement; the low ICC limits generalization to settings requiring stable scores across raters or sessions (Koo and Li, 2016).

#### A.4.2 Eye Modality Performance

For the eye modality condition with 9 participants, repeated measures ANOVA revealed significant main effects of model type across all six metrics:

**Message Targeting.** The ANOVA showed a significant main effect of model type ($F(4, 32) = 5.41$, $p = 0.002$, $\eta^2 = 0.403$). Grok 4 ($M = 5.47$, $SD = 0.89$) and OpenAI GPT-5 ($M = 5.47$, $SD = 0.88$) achieved the highest ratings, while the base model Qwen2.5-7B-Instruct showed the lowest performance ($M = 4.21$, $SD = 1.65$). Notably, **DesignerLLM** ($M = 4.86$, $SD = 1.21$) demonstrated a 15.4% improvement over the base model, indicating that our co-learning framework successfully enhanced message targeting metric despite its modest 7B parameter size.

**Trust.** A significant main effect was observed ($F(4, 32) = 4.66$, $p = 0.005$, $\eta^2 = 0.368$). Grok 4 ($M = 5.43$, $SD = 1.04$) achieved the highest trust ratings, followed by OpenAI GPT-5 ($M = 5.32$, $SD = 1.15$). **DesignerLLM** ($M = 4.65$,




**(a) First stage** (without intended messages) **(a) Second stage** (with intended messages)

Figure 5: The layout for each video assessment is in two stages. In the first stage, participants rated their initial impressions of the eHMI using four metrics, without any reference to the intended messages. In the second stage, after the intended messages were revealed, participants re-evaluated their impressions using two different metrics.

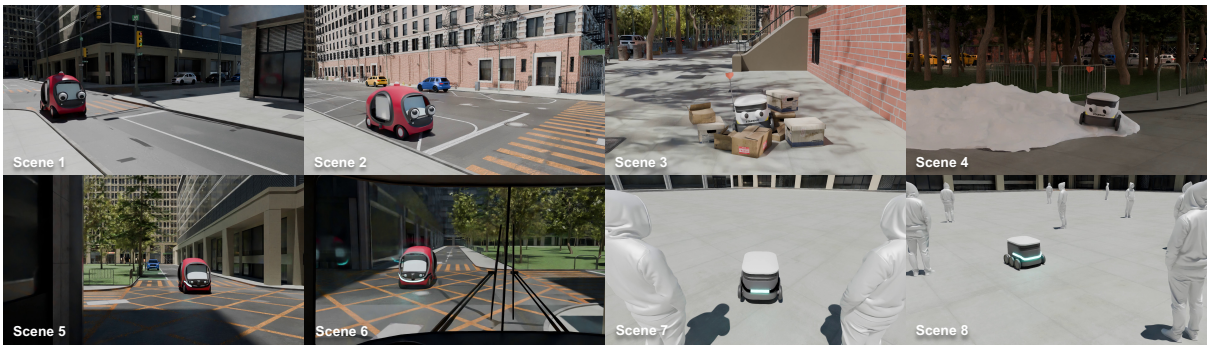



Figure 6: Demonstration of 3D scenarios used in the user study.

$SD = 1.32$) achieved a 7.9% improvement over the base model ($M = 4.31$, $SD = 1.64$), demonstrating the framework's effectiveness in generating more trustworthy eHHI action design.

**Perceived Safety.** Model type significantly affected perceived safety ratings ($F(4, 32) = 4.82$, $p = 0.004$, $\eta^2 = 0.376$). Grok 4 ($M = 5.65$, $SD = 0.90$) and OpenAI GPT-5 ($M = 5.53$, $SD = 1.04$) achieved the highest ratings. **DesignerLLM** ($M = 5.00$, $SD = 0.89$) showed a 7.1% improvement over the base model ($M = 4.67$, $SD = 1.06$), suggesting enhanced safety-oriented action generation through our framework.

**Mental Workload.** Analysis revealed significant differences in mental workload ($F(4, 32) = 4.23$, $p = 0.007$, $\eta^2 = 0.346$). Lower scores indicate better performance; Grok 4 ($M = 9.49$, $SD = 1.52$) and OpenAI GPT-5 ($M = 9.51$, $SD = 1.40$) imposed the lowest cognitive load, while the base model imposed the highest ($M = 10.97$, $SD = 1.38$). **DesignerLLM** ($M = 10.67$, $SD = 1.42$) demonstrated a 2.7% reduction in mental workload compared to the base model, indicating more intuitive eHMI outputs.

**User Acceptance.** The ANOVA indicated a significant main effect ($F(4, 32) = 4.23$, $p = 0.007$, $\eta^2 = 0.346$). Grok 4 ($M = 5.39$, $SD = 1.19$) and OpenAI GPT-5 ($M = 5.31$, $SD = 1.28$) achieved the highest acceptance ratings. **DesignerLLM** ($M = 4.38$, $SD = 1.48$) showed a 5.0% improvement in acceptance over the base model ($M = 4.17$, $SD = 1.46$), validating the practical applicability of our framework's outputs.

**Consistency.** Significant differences emerged across models ($F(4, 32) = 3.73$, $p = 0.013$, $\eta^2 = 0.318$). Grok 4 ($M = 5.32$, $SD = 1.27$) and OpenAI GPT-5 ($M = 5.29$, $SD = 1.31$) achieved the highest consistency ratings. **DesignerLLM** ($M = 4.39$, $SD = 1.36$) exhibited 8.1% better consistency than the base model ($M = 4.06$, $SD = 1.64$), demonstrating more reliable message convey capability.

Overall, **DesignerLLM** achieved an average rank of 4 across all metrics in the eye modality, substantially outperforming the base model (rank 5) with an average improvement of 7.7% across all metrics while maintaining competitive performance relative to commercial models with significantly larger parameter counts.

### A.4.3 Lightbar Modality Performance

For the lightbar modality condition with 9 participants, repeated measures ANOVA also revealed significant main effects for five of six metrics:

**Message Targeting.** The ANOVA showed a significant main effect ($F(4, 32) = 6.34$, $p < 0.001$, $\eta^2 = 0.442$). Grok 4 ($M = 5.35$, $SD = 1.03$), OpenAI GPT-5 ($M = 5.31$, $SD = 1.08$), and Claude Opus 4.1 ($M = 5.32$, $SD = 0.95$) achieved the highest ratings. **DesignerLLM** ($M = 4.75$, $SD = 1.39$) demonstrated a 12.6% improvement over the base model ($M = 4.22$, $SD = 1.33$).

**Trust.** A significant main effect was observed ($F(4, 32) = 4.40$, $p = 0.006$, $\eta^2 = 0.355$). OpenAI GPT-5 achieved the highest trust ($M = 5.53$, $SD = 1.09$), followed by Claude Opus 4.1 ($M = 5.50$, $SD = 0.97$) and Grok 4 ($M = 5.47$, $SD = 1.00$). **DesignerLLM** ($M = 5.04$, $SD = 1.17$) substantially outperformed the base model ($M = 4.67$, $SD = 1.20$) by 7.9%, approaching the performance of commercial models.

**Perceived Safety.** Model type significantly affected safety perceptions ($F(4, 32) = 3.94$, $p = 0.010$, $\eta^2 = 0.330$). OpenAI GPT-5 ($M = 5.65$, $SD = 0.90$) and Grok 4 ($M = 5.60$, $SD = 1.02$) achieved the highest ratings. **DesignerLLM** ($M = 5.25$, $SD = 1.12$) showed a 7.1% improvement over the base model ($M = 4.90$, $SD = 1.21$).

**Mental Workload.** No significant difference was observed across models ($F(4, 32) = 1.20$, $p = 0.331$, $\eta^2 = 0.130$). Nonetheless, OpenAI GPT-5 ($M = 9.47$, $SD = 1.90$) imposed the lowest cognitive load numerically. **DesignerLLM** ($M = 10.06$, $SD = 1.74$) showed a 3.8% reduction in workload compared to the base model ($M = 10.46$, $SD = 1.83$).

**User Acceptance.** The ANOVA indicated highly significant effects ($F(4, 32) = 13.55$, $p < 0.001$, $\eta^2 = 0.629$), representing the largest effect size across all metrics and modalities. OpenAI GPT-5 ($M = 5.31$, $SD = 1.47$), Claude Opus 4.1 ($M = 5.29$, $SD = 1.29$), and Grok 4 ($M = 5.06$, $SD = 1.58$) achieved the highest acceptance ratings. **DesignerLLM** ($M = 4.33$, $SD = 1.66$) achieved a substantial 12.5% improvement over the base model ($M = 3.85$, $SD = 1.75$), representing the largest single-metric gain.

**Consistency.** Significant model differences were found ($F(4, 32) = 7.54$, $p < 0.001$, $\eta^2 = 0.485$). OpenAI GPT-5 ($M = 5.40$, $SD = 1.39$) and Claude Opus 4.1 ($M = 5.28$, $SD = 1.31$) achieved the highest ratings. **DesignerLLM** ($M = 4.39$, $SD = 1.23$) demonstrated 3.5% improved consistency over the base model ($M = 4.24$, $SD = 1.15$).

In the lightbar modality, **DesignerLLM** maintained an average rank of 4.0, consistently outperforming the base model (rank 5.0) across all metrics with an average improvement of 7.9%, demonstrating cross-modality robustness of our co-learning framework.

### A.4.4 Summary

The human evaluation results demonstrate that our co-learning framework successfully enhances a 7B parameter model (**DesignerLLM**) to generate significantly more acceptable eHMI actions compared to its base model counterpart (Qwen2.5-7B-Instruct). Across both modalities, **DesignerLLM** achieved an average improvement of 7.8% across all metrics, with the most substantial gains observed in message targeting (15.4% for eye, 12.6% for lightbar) and user acceptance (5.0% for eye, 12.5% for lightbar). While commercial models with substantially larger parameter counts (GPT-5, Grok 4, Claude Opus 4.1) achieved higher absolute ratings, **DesignerLLM**'s consistent rank-4 performance across both modalities and all six metrics validates the effectiveness of our framework. Notably, the large effect sizes observed across most metrics ($\eta^2$ ranging from 0.318 to 0.629) indicate that model choice has a substantial impact on perceived eHMI quality. These results confirm that our framework not only improves automated VLM-based evaluation scores but also translates to meaningful enhancements in real human user experience, achieving competitive performance with a fraction of the computational resources required by commercial alternatives.

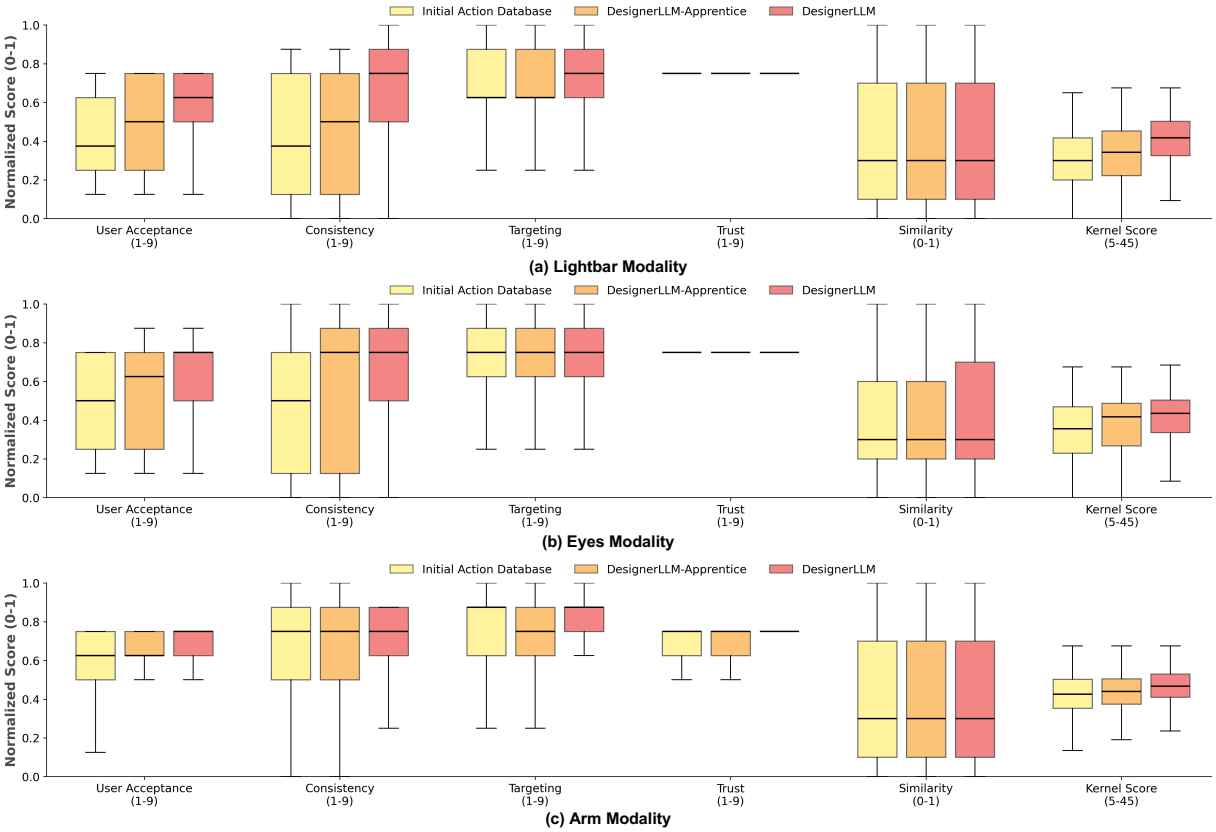



Figure 7: Boxplot of metric distributions at different training milestones. User Acceptance and Consistency both show increasing averages, with their distributions becoming more concentrated. In contrast, Trust exhibits minimal change.