Persona-aware and Explainable Bikeability Assessment: A Vision-Language Model Approach

Anonymous ACL submission

Abstract

Bikeability assessment is essential for advancing sustainable urban transportation and creating cyclist-friendly cities, and it requires incorporating users' perceptions of safety and comfort. Yet existing perception-based bikeability assessment approaches face key limitations in capturing the complexity of road environments and adequately accounting for heterogeneity in subjective user perceptions. This paper proposes a persona-aware Vision-Language Model framework for bikeability assessment with three novel contributions: (i) theory-grounded persona conditioning based on established cyclist typology that generates persona-specific explanations via chain-of-thought reasoning; (ii) multi-granularity supervised fine-tuning that combines scarce expert-annotated reasoning with abundant user ratings for joint prediction and explainable assessment; and (iii) AI-enabled data augmentation that creates controlled paired data to isolate infrastructure variable impacts. To test and validate this framework, we developed a panoramic image-based crowdsourcing system and collected 12,400 persona-conditioned assessments from 427 cyclists. Experiment results show that the proposed framework offers competitive bikeability rating prediction while uniquely enabling explainable factor attribution.

1 Introduction

Bikeability assessment, which refers to the systematic evaluation of roadway infrastructure to support cycling, is crucial for advancing sustainable urban transportation and creating cyclist-friendly cities. While traditional approaches focus on objective infrastructure measures such as traffic capacity and geometric design (Dowling and Reinke, 2008; Manual, 2000), growing evidence shows that people's willingness to cycle at a location depends more by how they perceive the environment than by these objectively measured features (Guo and He, 2021; Ma et al., 2014). Meanwhile, people's perceived safety at a given location can predict underlying safety risks, often identifying hazardous conditions before incidents occur (Prajapati and Tiwari, 2013; von Stülpnagel et al., 2022). Hence, there has been a growing call to incorporate users' perceptions of safety and comfort into bikeability assessment (Kellstedt et al., 2021; Gössling and McRae, 2022; Rodriguez-Valencia et al., 2022; Griswold et al., 2018). However, existing perception-based assessment approaches are limited in two key aspects: (i) comprehensively measuring complex road environments with multi-dimensional attributes and interacting factors that affect bikeability; (ii) properly accounting for cyclist heterogeneity as cyclists with varying experiences, preferences, and attitudes may perceive identical environment differently.

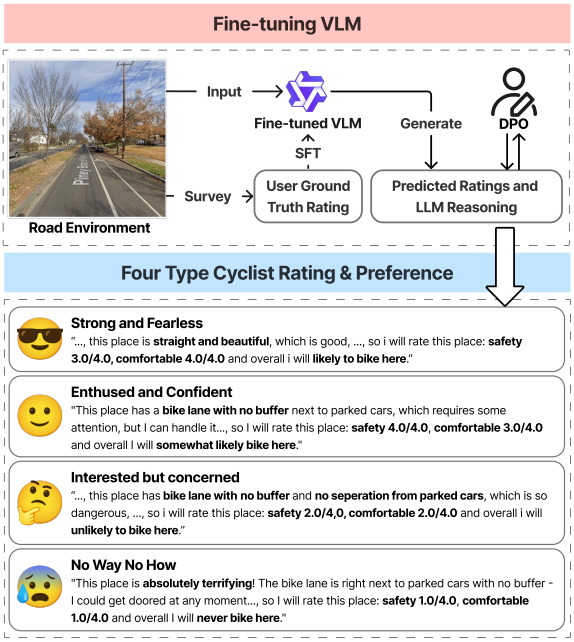

Figure 1: Overview of our persona-aware explainable bikeability assessment framework.

Recent advances in Vision-Language Models

(VLMs) offer new opportunities for perception-based bikeability assessments by facilitating the generation of high-fidelity user safety and comfort perception data at scale. VLMs have demonstrated strong capabilities in capturing and mimicking human perceptions and cognitive processes, enabling various applications such as environmental assessment (Ito and Biljecki, 2021), street design visualization (Wang et al., 2025a), and human perception modeling (Danish et al., 2025). These models can process multi-modal inputs combining visual and structured data, and generate natural language outputs that facilitate interpretability. Meanwhile, persona-based methods have shown promise in capturing diverse user behaviors through LLM conditioning (Wang et al., 2024; Chen et al., 2024).

However, applying VLMs for perception-based bikeability assessment face several key technical challenges. First, VLMs can exhibit issues such as hallucinations and unstable predictions, which limit their reliability and generalizability. Second, Effective training of VLMs that balances high inference accuracy with strong interpretability necessitates large-scale data on bikeability ratings from diverse users, alongside high-quality "reasoning chain" data capturing the rationale for each rating. Yet acquiring both types of data at scale is costly and often infeasible. Third, real-world environmental factors are highly correlated (e.g., roads with bike lanes often have better greenery), which can limit a VLM's reasoning capabilities if calibration relies solely on observational data, making it difficult to isolate the perceptual impact of individual infrastructure variables.

This paper proposes a persona-aware bikeability assessment framework (Figure 1) that addresses these three challenges with the following novel approaches: (i) a theory-grounded persona conditioning approach that first classifies cyclists based on established typology (Kim and Mokhtarian, 2023; Dill and McNeil, 2013) and generates persona-specific explanations through chain-of-thought reasoning; (ii) a multi-granularity supervised fine-tuning strategy that combines data of varying annotation depth to jointly achieve rating prediction, factor identification, and interpretable reasoning generation; and (iii) an AI-enabled data augmentation method that elicits distinct user responses under various bicycle facility design scenarios, hence enabling the model to more effectively isolate the perceptual impact of individual infrastructure variables on bikeability ratings. To test and validate this framework, we developed a panoramic image-based crowdsourcing survey system to collect hundreds of cyclists' bikeability ratings for multiple road segments across diverse road environments in Washington DC.

2 Related Work

2.1 Bikeability Assessment

Bikeability assessment requires evaluation of multidimensional attributes of the cycling environment while accounting for subjective user perceptions (Kellstedt et al., 2021; Gössling and McRae, 2022). Early approaches include the Level of Traffic Stress (LTS) framework (Dill and McNeil, 2016; Mekuria et al., 2012), though it oversimplifies user heterogeneity (Damant-Sirois et al., 2014). Traditional statistical models prioritize causal relationships over prediction accuracy (Breiman, 2001; Shmueli, 2010; Zhao et al., 2020). Recent data-driven methods leverage diverse sources: Paranga and Oda (Paranga and Oda, 2025) applied PCA/CFA on survey indicators, Zhang et al. (Zhang et al., 2025) combined COPRAS with machine learning, Ito and Biljecki (Ito and Biljecki, 2021) trained LightGBM on CV features, Zeng et al. (Zeng et al., 2024) used Random Forest, and AutoLTS (Lin et al., 2024) employed contrastive learning. However, these approaches either treat cyclists as homogeneous or rely on expert-predetermined weights that may not reflect actual user priorities, and vision-based methods typically require multi-stage pipelines limiting generalizability.

2.2 Vision-Language Models for Urban Analysis

Multi-modal approaches integrating street-view imagery, remote sensing, and geospatial data have enabled diverse urban analyses (Gebru et al., 2017; Dai et al., 2025; Zhao et al., 2023; Danish et al., 2025; Suel et al., 2021; Albert et al., 2017; Cao et al., 2020). Open-source VLMs such as Qwen-VL (Yang et al., 2025a) and LLaVA (Liu et al., 2024) now provide strong visual understanding with parameter-efficient fine-tuning capabilities (Hu et al., 2021), enabling domain adaptation for specialized tasks. These models advance urban analysis through CoT reasoning (Wei et al., 2022; Kong et al., 2024) and multi-granularity instruction tuning (Liu et al., 2024; Deng et al., 2025; Wang et al., 2022; Sanh et al., 2021; Chung et al., 2024), showing promise in navigation (Wang et al., 2025b;

Wu et al., 2025), infrastructure assessment (Ito and Biljecki, 2021), street design (Wang et al., 2025a), and decision-making (Chen et al., 2025). Meanwhile, recent advances in image generation models (Yang et al., 2025b) have enabled high-fidelity synthetic data creation, offering new possibilities for controlled data augmentation in domains where paired observations are scarce. However, domain applications require interpretability for stakeholder validation (Arrieta et al., 2019; Javed et al., 2023; Dong et al., 2025), and existing approaches focus on synthetic or web-scale data rather than domain-specific settings where expert annotations are scarce. For such subjective perception tasks, aligning model outputs with human preferences is critical, as ground truth reflects individual judgments rather than objective labels. Preference optimization methods such as RLHF (Ouyang et al., 2022) and DPO (Rafailov et al., 2024; Bai et al., 2022; Meng et al., 2024; Xiao et al., 2025) provide effective mechanisms for this alignment.

2.3 Persona-aware Modeling

User heterogeneity is fundamental to perception-based assessment, as expert-derived standards often deviate from actual user experiences (Gössling and McRae, 2022). The "Four Types of Cyclists" typology (Dill and McNeil, 2013, 2016) confirms significant differences in infrastructure needs across cyclist types. Persona-conditioning enhances LLM behavioral simulation (Wang et al., 2024; Li et al., 2024; Chen et al., 2024; Sun et al., 2025), yet existing methods rely on predefined categories or aggregated data (Chuang et al., 2024; Park et al., 2024), lacking mechanisms to model differentiated visual interpretations of streetscapes (Ito and Biljecki, 2021).

3 Method

We propose a persona-aware bikeability assessment framework that addresses both the complexity of objective road environment assessment and the heterogeneity of subjective user perceptions. Figure 2 illustrates our pipeline, which consists of three stages: (1) Multi-source Data Collection: We gather persona-specific bikeability ratings and influencing factors through crowdsourcing surveys covering diverse Washington DC road environments, with AI-based image augmentation for controlled infrastructure variations. (2) Multi-Granularity Supervised Fine-tuning: We create three types of training data (expert-annotated full reasoning chains, user-provided structured factor-rating pairs, and direct ratings) and jointly train on all three with fixed sampling ratios, enabling the model to learn interpretable reasoning while maintaining robust prediction. (3) Preference-based Reasoning Refinement: We refine reasoning outputs through Direct Preference Optimization (DPO) based on human feedback.

3.1 Data Collection and Preprocessing

3.1.1 Study Design and Sampling

We randomly sampled 200 road segments from Washington DC's road network, prioritizing spatial dispersion to ensure diverse cycling environments. The selected segments encompass varying motor vehicle lane configurations, cycling infrastructure types (from absent to fully protected), and urban contexts. We integrated GSV panoramic API into our survey system, enabling participants to conduct detailed 360-degree observations rather than being limited to static 2D images.

3.1.2 Image Augmentation Pipeline

During our road segment sampling process, we found that natural road networks rarely provide ideal paired comparisons where segments differ in only a minimal number of infrastructure attributes, even adjacent roads typically vary across multiple dimensions simultaneously. This makes it challenging for models to learn fine-grained attribute importance from limited survey data. To address this, we developed an AI-based image editing pipeline using GPT-image-1. This pipeline systematically modifies objectively defined infrastructure variables without introducing subjective developer judgments, including: Bike Lane Presence, Lane Width (Narrow/Standard/Wide), Lane Color (Green/No Paint), Buffer Type (No Buffer/Standard/Bollards/Armadillo), and Buffer Location (Adjacent to Moving Cars/Adjacent to Parked Cars). By altering these variables while preserving all other environmental factors, the pipeline enables valid paired comparisons for isolating infrastructure impacts. The original image sample and image editing result is shown in Figure 3.

3.1.3 Crowdsourcing survey

Our survey collects the following information:

Demographics: Gender, age group, race/ethnicity, and education level to ensure diverse representation.

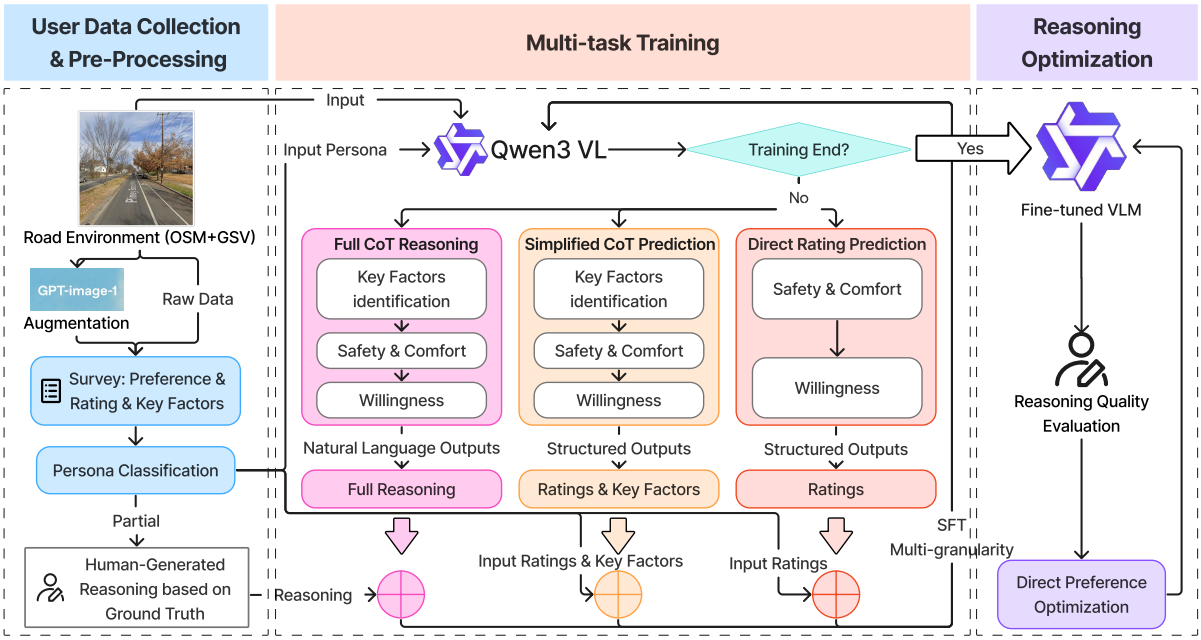

Figure 2: Three-stage model architecture: (1) Multi-source data collection with crowdsourced survey and AI-based image augmentation, (2) Multi-granularity supervised fine-tuning with three data types (Type 1: full reasoning, Type 2: factor-rating pairs, Type 3: rating-only), and (3) Preference-based reasoning refinement with DPO.

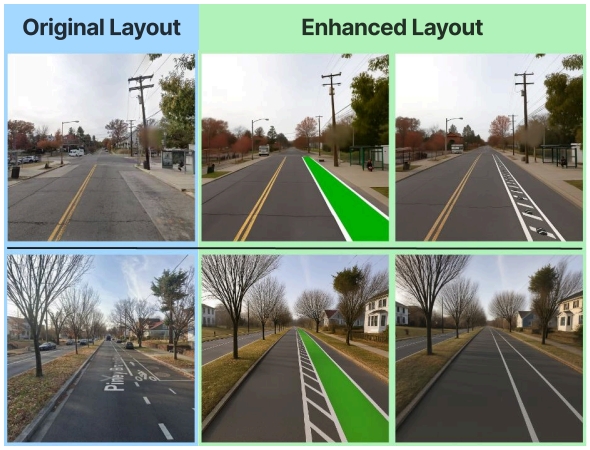

Figure 3: AI-based image augmentation: original street-view images (left) and systematically modified versions with controlled infrastructure changes (right).

Cycling preferences: Comfort ratings (5-point scale: 1=Very Uncomfortable to 5=Very Comfortable) for eight infrastructure types using reference images: (1) no bike lanes, (2) roadway shoulders, (3) off-street multi-use paths, (4) shared lanes with sharrows, (5) sidewalks, (6) striped bike lanes, (7) buffered bike lanes, and (8) protected bike lanes. These ratings enable cyclist persona identification, capturing individual risk tolerance and comfort thresholds.

Segment evaluations: For 15 locations (20 after augmentation) randomly assigned to each respondent, we collected: (1) *Safety rating* and (2) *Comfort rating* on 4-point Likert scales (1=Strongly Disagree to 4=Strongly Agree); (3) *Cycling willingness* on a 4-point scale (1=Never to 4=Absolutely); and (4) *Influencing factors* via a multi-select tag interface with predefined options (bike lane conditions, traffic, separation, environment) and open-ended fields.

3.1.4 Persona Classification

Our survey collected responses from 427 participants. Each participant first provided comfort ratings (1-5 scale) for eight infrastructure types, then rated a minimum of 20 street-view images for bikeability assessment, resulting in 12,400 total persona-conditioned assessments (mean: 29.0 images per participant).

To validate the necessity of persona-aware modeling, we analyzed within-participant rating variance. Figure 4 reveals substantial inter-participant heterogeneity: variance in street segment ratings ranges from 0 to 2.17 (median = 0.872), and is orthogonal to mean rating levels ($r = -0.10$), confirming genuine individual differences rather than scale-use artifacts.

We classified participants into four cyclist personas following the "Four Types of Cyclists" typology (Dill and McNeil, 2016). We computed mean comfort ratings under three protection lev-

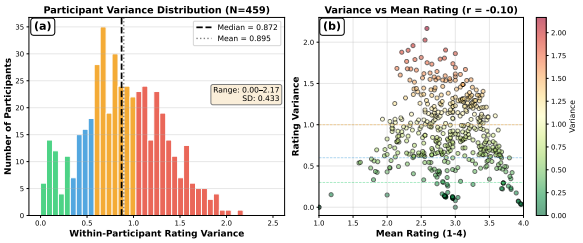

Figure 4: Within-participant rating variance distribution across 427 participants. (a) Histogram shows substantial heterogeneity (median = 0.872, range: 0–2.17). (b) Scatter plot confirms variance is orthogonal to mean rating (r = −0.10).

els: low (no bike lanes, shared lanes), medium (striped lanes, shoulders, off-street paths), and high (buffered/protected lanes), plus the gradient from low to high protection. Clustering on these indicators reveals four groups: participants with low overall comfort and minimal gradient emerge as *No Way No How* (NWNH); those with high comfort and flat gradients cluster as *Strong & Fearless* (S&F); individuals with steep gradients form the *Interested but Concerned* (IbC) group; while those with moderate comfort and gradients constitute *Enthused & Confident* (E&C). Table 1 presents the distribution.

Table 1: Persona distribution and infrastructure preference characteristics (N=427). Mean and Grad (gradient) are computed from comfort ratings (1-5 scale) for eight infrastructure types.

Persona	Pct.	Mean	Grad.
IbC	59.3%	2.97	2.73
E&C	27.6%	3.45	1.57
S&F	8.2%	3.80	0.13
NWNH	4.9%	2.02	0.50

3.2 Task Definition

We formulate persona-aware bikeability assessment as a conditional generation task where VLM generates subjective evaluations based on cyclist personas and multi-modal inputs. By combining visual street-level observations with structured infrastructure attributes as input, we ensure comprehensive coverage while avoiding human selection bias in feature engineering.

Formally, given a cyclist persona P, the model generates a subjective evaluation text $X = \{x_1, x_2, ..., x_S\}$ consisting of S words that describe the bikeability assessment from the perspective of that persona type. The model takes as input:

- A cyclist persona P from four types: S&F, E&C, IbC, or NWNH
- O: A street-view image capturing the first-person cyclist perspective, providing rich visual context
- $A = \{a_1, a_2, ..., a_I\}$: A set of I infrastructure attributes from OpenStreetMap (e.g., road type, lane configuration), ensuring comprehensive coverage of non-visual factors

Given the ground truth user evaluation X^* from our survey data, we optimize the model parameters θ to maximize the conditional probability of generating the correct evaluation:

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \log p(X^* | P, O, A; \theta) \quad (1)$$

Following established transportation theory (Dill and McNeil, 2016; Winters et al., 2011) and our empirical validation of systematic inter-participant heterogeneity (Section 3.1.4, Figure 4), we treat persona-specific assessment as a fundamental design requirement rather than an optional feature.

3.2.1 Multi-format Training Data

Since manually annotated complete Chain-of-Thought (CoT) reasoning is scarce and labor-intensive, we design three training data formats with varying supervision granularity. Our multi-granularity strategy leverages complementary strengths: Type 1 enables interpretable reasoning generation, Type 2 develops factor identification skills, and Type 3 ensures robust rating prediction. By jointly training on all three types with fixed sampling ratios (Section 4.1), the model learns a unified representation supporting multiple output formats while maintaining consistency across explanation depths. All three formats share identical multi-modal inputs (P, O, A) but differ in their prompt instructions and supervision signals.

Type 1: Full CoT Reasoning. This format trains the model to generate interpretable natural language reasoning. However, obtaining complete reasoning chains directly from crowdsourced surveys is challenging due to: (i) *cognitive burden* leading to lower completion rates (Krosnick, 1991; Deutskens et al., 2004); (ii) *expression heterogeneity* creating noise that complicates training; and

(iii) *implicit reasoning* where perceptual judgments occur subconsciously (Nisbett and Wilson, 1977; Wilson, 2002).

Given these constraints, we construct full reasoning chains through expert annotation by transportation researchers. Critically, this differs from traditional expert-driven approaches: rather than predefining which factors matter, experts receive user-provided factors and ratings from survey responses and synthesize them into coherent reasoning chains. This data-driven approach ensures reasoning patterns emerge from actual user preferences. The model receives:

$$[P, O, A; \text{prompt}_{\text{reason}}] \tag{2}$$

where $\text{prompt}_{\text{reason}}$ instructs the model to: (i) analyze environmental conditions from the image, (ii) identify key factors influencing the cyclist's perception, (iii) predict safety and comfort ratings, and (iv) synthesize an overall willingness assessment. The supervision signal is:

$$X^*_{\text{reason}} = \text{Annotate}(T^*, R^*) \tag{3}$$

where transportation researchers compose reasoning chains connecting the ground truth factors T^* and ratings R^* using natural language. While large teacher models (e.g., GPT-4) could automate annotation, such approaches introduce hallucination risks and inconsistent patterns (Bang et al., 2023; Xu et al., 2023). Our expert annotation ensures high-quality supervision that faithfully represents participant perspectives.

Type 2: Simplified CoT Prediction. Unlike Type 1, this format lacks manually completed reasoning narratives and is derived directly from survey responses. The model receives:

$$[P, O, A; \text{prompt}_{\text{struct}}] \tag{4}$$

where $\text{prompt}_{\text{struct}}$ instructs the model to perform a simplified CoT process that omits the objective condition analysis step: (i) identifying key factors that most influence the cyclist's perception, (ii) predicting safety and comfort ratings, and (iii) providing an overall willingness-to-use assessment. The supervision signals consist of paired annotations from participants:

$$(T^*, R^*) \tag{5}$$

where $T^* = \{t_1, t_2, ..., t_n\}$ represents the set of influencing factors identified by participants , and R^* contains their corresponding three-dimensional ratings. This format bridges perception and evaluation through explicit factor identification without requiring full explanatory narratives.

Type 3: Direct Rating Prediction. This format further simplifies the task by removing the factor identification step, utilizing all available survey ratings for maximum data efficiency. The model receives:

$$[P, O, A; \text{prompt}_{\text{rating}}] \tag{6}$$

where $\text{prompt}_{\text{rating}}$ requests direct rating prediction: the model predicts safety and comfort ratings, followed by an overall willingness-to-use assessment, without requiring intermediate factor identification or reasoning. The supervision signal consists of the three-dimensional rating vector:

$$R^* = \{r_{\text{safety}}, r_{\text{comfort}}, r_{\text{willingness}}\} \tag{7}$$

This format represents basic stimulus-response mapping, providing the most direct supervision signal and ensuring robust end-to-end prediction capabilities.

3.2.2 Model Training

To efficiently adapt the model given limited survey data and prevent overfitting, we employ parameter-efficient fine-tuning using LoRA (Hu et al., 2021) on the pretrained Qwen3-VL-8B-Instruct base model.

Multi-Granularity Instruction Tuning. Since manually annotated complete CoT reasoning (Type 1) is scarce and labor-intensive, we adopt a multi-granularity instruction tuning approach following LLaVA-1.5 (Liu et al., 2024). Our training data exhibits natural granularity variation: while all 12,400 survey responses provide ratings (Type 3) and user-identified factors (Type 2), only approximately 2,000 samples contain expert-annotated reasoning chains (Type 1). This data constraint necessitates multi-granularity training: relying solely on Type 1 would provide insufficient samples for learning robust visual-linguistic representations; conversely, using only Type 3 would sacrifice all interpretability. Our multi-granularity strategy leverages complementary strengths: Type 1 enables interpretable reasoning generation, Type 2 develops factor identification skills, and Type 3 ensures robust rating prediction. By jointly training on all three types with fixed sampling ratios (Section 4.1), the model learns a unified representation supporting multiple output formats while maintaining consistency across explanation depths.

Direct Preference Optimization. Following supervised fine-tuning, we apply Direct Preference Optimization (DPO) (Rafailov et al., 2024) to Type 1 outputs to refine reasoning quality. To collect preference data, we sample instances from the training set and use the SFT model to generate two Type 1 reasoning explanations per instance with different sampling temperatures. Three transportation domain experts independently evaluate both explanations, selecting the preferred one based on: (i) *Factual Accuracy*, (ii) *Logical Coherence*, and (iii) *Persona Consistency*. Majority voting determines the final preference pairs. Following the DPO framework, we directly optimize the policy to increase the likelihood of preferred responses without training a separate reward model. The DPO objective uses reference model regularization to prevent deviation from the SFT initialization, preserving performance on Type 2 and Type 3 formats.

4 Experiments

4.1 Experimental Setup

Implementation Details. We conduct all experiments on one NVIDIA A100 80GB GPU. For supervised fine-tuning, we employ LoRA with rank $r = 32$ and scaling factor $\alpha = 64$, targeting the query, key, value, and output projection matrices (`q_proj`, `k_proj`, `v_proj`, `o_proj`). We use AdamW optimizer with an initial learning rate of 2×10^{-4} and train for 5 epochs with cosine annealing learning rate schedule. The learning rate follows a decay schedule: gentle decay ($0.8\times$ per epoch) for the first 3 epochs, then rapid decay ($0.5\times$ per epoch) for the remaining epochs. We train with batch size 4 and gradient accumulation over 4 steps for an effective batch size of 16. We apply gradient checkpointing to manage memory usage and use 10% warmup steps. Mixed precision training (FP16) is employed for efficiency.

Multi-Granularity Sampling Ratios. We sample training data with fixed ratios across all epochs: 15% Type 1 (full reasoning), 40% Type 2 (structured factor-rating pairs), and 45% Type 3 (rating-only). This distribution is consistent with LLaVA-1.5's finding that detailed reasoning samples comprising 15% yields optimal performance (Liu et al., 2024). The fixed-ratio design ensures: (i) efficient utilization of scarce expert annotations without over-sampling (15% yields ∼1,860 Type 1 samples per epoch, approaching our full 2,000-sample budget), (ii) balanced multi-task learning across all supervision levels, and (iii) training stability by avoiding catastrophic forgetting (Kirkpatrick et al., 2017). We enforce strict rating constraints (1-4 scale) through explicit prompt instructions and post-processing corrections.

DPO Training Details. For preference data collection, we randomly sample 500 instances and generate two explanations per instance with sampling temperatures 0.7 and 1.0, resulting in 500 preference pairs after expert majority voting. For DPO training, we use regularization coefficient $\beta = 0.1$, learning rate 5×10^{-6}, batch size 8, and train for 3 epochs.

Baselines. We compare against three baselines: (1) *GPT-4o Zero-shot*: prompting GPT-4o with persona descriptions and multi-modal inputs without task-specific training; (2) *Kmeans-SMOTE RF* (Zeng et al., 2024): To ensure fair comparison, we replicate Zeng et al.'s preprocessing-based approach by employing YOLOv8 for comprehensive detection of observable road infrastructure elements (bike lanes, buffers, traffic signals, street furniture, greenery, etc.), combining these computer vision features with OpenStreetMap attributes (motor vehicle lane count, speed limits, bike lane types) and image-derived latent representations (encoded via a pretrained ResNet-50). These multi-source features are then fed into a Random Forest classifier following their Kmeans-SMOTE balancing strategy, extended with our LLM-normalized factor tag pool for factor identification; (3) *Kmeans-SMOTE RF (Rating-only)*: same preprocessing and feature extraction pipeline but trained only for rating prediction, isolating the contribution of joint tag-rating learning.

Evaluation Metrics. For rating prediction, we report metrics averaged across all three dimensions (safety, comfort, willingness-to-use): Mean Absolute Error (MAE), Exact Match Rate (EM), Within-One Accuracy (W1, percentage within ± 1), and Pearson correlation. For factor identification, we compute Semantic Precision, Recall, and F1-score using sentence embeddings (all-MiniLM-L6-v2) with greedy matching at threshold 0.7. For reasoning quality, we use GPT-4o as an automatic judge (Zheng et al., 2023) to assess: (1) *Factual Accuracy*, correct description of infrastructure features; (2) *Logical Coherence*, clarity and logical flow; (3) *Persona Consistency*, alignment with persona-specific concerns.

Table 2: Main experimental results comparing rating prediction and factor identification performance. MAE: Mean Absolute Error, EM: Exact Match, W1: Within-One Accuracy, Corr: Pearson Correlation, Prec: Precision, Rec: Recall, F1: F1-score (at threshold=0.7).

Method	Rating Prediction				Factor Identification		
	MAE↓	EM↑	W1↑	Corr↑	Prec↑	Rec↑	F1↑
GPT-4o Zero-shot	1.00	0.30	0.70	0.25	0.12	0.08	0.10
KS-RF (Rating-only)	**0.70**	**0.45**	0.85	**0.50**	-	-	-
KS-RF	0.80	0.38	0.82	0.45	0.33	0.30	0.31
Ours	0.71	0.41	**0.87**	0.48	**0.52**	**0.46**	**0.49**

4.2 Main Results

Table 2 presents our main results. For rating prediction, our method achieves competitive performance (MAE: 0.71, EM: 0.41, Corr: 0.48) closely matching the specialized regression approach KS-RF (Rating-only), while achieving the highest within-one accuracy (W1: 0.87).

For factor identification, our approach achieves 0.52 precision, 0.46 recall, and 0.49 F1-score, substantially outperforming KS-RF (F1: 0.31) and GPT-4o zero-shot (F1: 0.10). This demonstrates that domain-specific training enables accurate identification of bikeability-relevant factors despite wording variations between generated and ground-truth factors.

4.3 Human Preference Alignment

Table 3: Human preference alignment results. Acc.: Factual Accuracy, Coh.: Logical Coherence, Cons.: Persona Consistency.

Method	Acc.↑	Coh.↑	Cons.↑
GPT-4o	0.25	0.694	0.995
Ours (SFT)	0.58	0.580	0.920
Ours+DPO	**0.59**	**0.610**	**0.950**
Increase	+1.7%	+5.2%	+3.3%

Table 3 shows that DPO improves explanation quality across all dimensions. Our SFT model achieves substantially higher factual accuracy (0.58) than GPT-4o (0.25), reflecting domain-specific training, while GPT-4o exhibits stronger coherence and consistency due to its larger scale. DPO narrows this gap, with the largest improvement in coherence (+5.2%), validating that preference optimization refines explanation style and logical flow.

4.4 Ablation Studies

Multi-Granularity Data Ablation. We evaluate the contribution of different supervision granularities. Training with Type 3 only (rating-only) achieves MAE 0.75 and W1 0.85, but provides no factor identification or explanations. Adding Type 2 (factor-rating pairs) improves prediction to MAE 0.73 and W1 0.86, though reasoning remains absent.

Our full approach with Type 1 (CoT reasoning) at 15/40/45 ratios achieves the best accuracy (MAE: 0.71, W1: 0.87) while enabling factor identification (F1: 0.49) and interpretable reasoning. This demonstrates that CoT provides the richest supervisory signal for joint optimization. The fixed-ratio strategy ensures training stability, as aggressive epoch-wise ratio changes led to unstable convergence and catastrophic forgetting.

5 Conclusion

We presented a persona-aware VLM approach for bikeability assessment that addresses infrastructure evaluation complexity and user perception heterogeneity. Our results demonstrate that interpretability need not compromise accuracy: our framework achieves competitive rating prediction while uniquely enabling factor identification and interpretable reasoning, capabilities absent in existing approaches.

For researchers, our framework eliminates repetitive preprocessing pipelines when adapting to new urban contexts. For planners, persona-specific assessments reveal diverse community needs beyond the "average" cyclist, informing more inclusive infrastructure decisions. We hope this work inspires broader VLM adoption in transportation research, supporting human-centered infrastructure assessment that accounts for user diversity.

6 Limitations

Our study has several limitations. First, while the framework is transferable, the model trained on Washington DC data may require fine-tuning for cities with different infrastructure styles. Second,

our four-category cyclist typology is a discrete approximation of continuous preference spectra. Third, static street-view imagery does not capture dynamic factors such as traffic volume, weather, or temporal variations. Finally, a gap remains in general language capabilities compared to larger models like GPT-4o.

References

Adrian Albert, Jasleen Kaur, and Marta C Gonzalez. 2017. Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1357–1366.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2019. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Preprint*, arXiv:1910.10045.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022. Constitutional ai: Harmlessness from ai feedback. *Preprint*, arXiv:2212.08073.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *Preprint*, arXiv:2302.04023.

Leo Breiman. 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231.

Rui Cao, Wei Tu, Cuixin Yang, Qing Li, Jun Liu, Jiasong Zhu, Qian Zhang, Qingquan Li, and Guoping Qiu. 2020. Deep learning-based remote and social sensing data fusion for urban region function recognition. *ISPRS Journal of Photogrammetry and Remote Sensing*, 163:82–97.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024. From persona to personalization: A survey on role-playing language agents. *Preprint*, arXiv:2404.18231.

Ruxiao Chen, Chenguang Wang, Yuran Sun, Xilei Zhao, and Susu Xu. 2025. From perceptions to decisions: Wildfire evacuation decision prediction with behavioral theory-informed llms. *Preprint*, arXiv:2502.17701.

Yun-Shiuan Chuang, Krirk Nirunwiroj, Zach Studdiford, Agam Goyal, Vincent V. Frigo, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T. Rogers. 2024. Beyond demographics: Aligning role-playing llm-based agents using human belief networks. *Preprint*, arXiv:2406.17232.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Yilong Dai, Luyu Liu, Kaiyue Wang, Meiqing Li, and Xiang Yan. 2025. Using computer vision and street view images to assess bus stop amenities. *Computers, Environment and Urban Systems*, 117:102254.

Gabriel Damant-Sirois, Michael Grimsrud, and Ahmed M El-Geneidy. 2014. What's your type: A multidimensional cyclist typology. *Transportation*, 41(6):1153–1169.

Matthew Danish, SM Labib, Britta Ricker, and Marco Helbich. 2025. A citizen science toolkit to collect human perceptions of urban environments using open street view images. *Computers, Environment and Urban Systems*, 116:102207.

Huilin Deng, Ding Zou, Rui Ma, Hongchen Luo, Yang Cao, and Yu Kang. 2025. Boosting the generalization and reasoning of vision language models with curriculum reinforcement learning. *Preprint*, arXiv:2503.07065.

Elisabeth Deutskens, ko de ruyter, Martin Wetzels, and Paul Oosterveld. 2004. Response rate and response quality of internet-based surveys: An experimental study. *Marketing Letters*, 15:21–36.

Jennifer Dill and Nathan McNeil. 2013. Four types of cyclists? examination of typology for better understanding of bicycling behavior and potential. *Transportation Research Record*, 2387(1):129–138.

Jennifer Dill and Nathan McNeil. 2016. Revisiting the four types of cyclists: Findings from a national survey. *Transportation research record*, 2587(1):90–99.

Yifei Dong, Fengyi Wu, Kunlin Zhang, Yilong Dai, Sanjian Zhang, Wanghao Ye, Sihan Chen, and Zhi-Qi Cheng. 2025. Large language model agents in finance: A survey bridging research, practice, and real-world deployment. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17889–17907, Suzhou, China. Association for Computational Linguistics.

Richard Dowling and David Reinke. 2008. *Multimodal level of service analysis for urban streets: users guide*. Transportation Research Board Washington, DC, USA.

Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. 2017. Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. *Proceedings of the National Academy of Sciences*, 114(50):13108–13113.

Stefan Gössling and Sophia McRae. 2022. Subjectively safe cycling infrastructure: New insights for urban designs. *Journal of Transport Geography*, 101:103340.

Julia B Griswold, Mengqiao Yu, Victoria Filingeri, Offer Grembek, and Joan L Walker. 2018. A behavioral modeling approach to bicycle level of service. *Transportation research part A: policy and practice*, 116:166–177.

Yuanyuan Guo and Sylvia Y He. 2021. The role of objective and perceived built environments in affecting dockless bike-sharing as a feeder mode choice of metro commuting. *Transportation Research Part A: Policy and Practice*, 149:377–396.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Koichi Ito and Filip Biljecki. 2021. Assessing bikeability with street view imagery and computer vision. *Transportation research part C: emerging technologies*, 132:103371.

Abdul Rehman Javed, Waqas Ahmed, Sharnil Pandya, Praveen Kumar Reddy Maddikunta, Mamoun Alazab, and Thippa Reddy Gadekallu. 2023. A survey of explainable artificial intelligence for smart cities. *Electronics*, 12(4).

Debra K Kellstedt, John O Spengler, Margaret Foster, Chanam Lee, and Jay E Maddock. 2021. A scoping review of bikeability assessment methods. *Journal of community health*, 46(1):211–224.

Sung Hoo Kim and Patricia L Mokhtarian. 2023. Finite mixture (or latent class) modeling in transportation: Trends, usage, potential, and future directions. *Transportation Research Part B: Methodological*, 172:134–173.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Xianghao Kong, Jinyu Chen, Wenguan Wang, Hang Su, Xiaolin Hu, Yi Yang, and Si Liu. 2024. *Controllable Navigation Instruction Generation with Chain of Thought Prompting*, page 37–54. Springer Nature Switzerland.

Jon A. Krosnick. 1991. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5(3):213–236.

Xuchuan Li, Fei Huang, Jianrong Lv, Zhixiong Xiao, Guolong Li, and Yang Yue. 2024. Be more real: Travel diary generation using llm agents and individual profiles. *Preprint*, arXiv:2407.18932.

Bo Lin, Shoshanna Saxe, and Timothy CY Chan. 2024. Autolts: automating cycling stress assessment via contrastive learning and spatial post-processing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22222–22230.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Liang Ma, Jennifer Dill, and Cynthia Mohr. 2014. The objective versus the perceived environment: what matters for bicycling? *Transportation*, 41(6):1135–1152.

Highway Capacity Manual. 2000. Highway capacity manual. *Washington, DC*, 2(1):1.

Maaza C Mekuria, Peter G Furth, and Hilary Nixon. 2012. Low-stress bicycling and network connectivity. Technical Report CA-MTI-12-1005, Mineta Transportation Institute.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *Preprint*, arXiv:2405.14734.

Richard Nisbett and Timothy Wilson. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84:231–259.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Norbert Anthony Gerome Paranga and Takanori Oda. 2025. Correlating objective (road inventory) and subjective (user-perception) features in walkability and bikeability assessment. *Transportation Research Procedia*, 82:1452–1471.

Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. 2024. Generative agent simulations of 1,000 people. *Preprint*, arXiv:2411.10109.

Pankaj Prajapati and Geetam Tiwari. 2013. Study of relation between actual and perceived crash risk. *Procedia-social and behavioral sciences*, 104:1095–1104.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Alvaro Rodriguez-Valencia, Jose Agustin Vallejo-Borda, German A Barrero, and Hernan Alberto Ortiz-Ramirez. 2022. Towards an enriched framework of service evaluation for pedestrian and bicyclist infrastructure: acknowledging the power of users' perceptions. *Transportation*, 49(3):791–814.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, and 1 others. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Galit Shmueli. 2010. To explain or to predict? *Statistical science*, pages 289–310.

Esra Suel, Samir Bhatt, Michael Brauer, Seth Flaxman, and Majid Ezzati. 2021. Multimodal deep learning from satellite and street-level imagery for measuring income, overcrowding, and environmental deprivation in urban areas. *Remote Sensing of Environment*, 257:112339.

Chenkai Sun, Ke Yang, Revanth Gangi Reddy, Yi R. Fung, Hou Pong Chan, Kevin Small, ChengXiang Zhai, and Heng Ji. 2025. Persona-db: Efficient large language model personalization for response prediction with collaborative data refinement. *Preprint*, arXiv:2402.11060.

Rul von Stülpnagel, Chayenne Petinaud, and Sven Lißner. 2022. Crash risk and subjective risk perception during urban cycling: Accounting for cycling volume. *Accident Analysis & Prevention*, 164:106470.

Chenguang Wang, Xiang Yan, Yilong Dai, Ziyi Wang, and Susu Xu. 2025a. From image generation to infrastructure design: a multi-agent pipeline for street design generation. *arXiv preprint arXiv:2509.05469*.

Jiawei Wang, Renhe Jiang, Chuang Yang, Zengqing Wu, Makoto Onizuka, Ryosuke Shibasaki, Noboru Koshizuka, and Chuan Xiao. 2024. Large language models as urban residents: An llm agent framework for personal mobility generation. *Preprint*, arXiv:2402.14744.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Zihan Wang, Yaohui Zhu, Gim Hee Lee, and Yachun Fan. 2025b. Navrag: Generating user demand instructions for embodied navigation through retrieval-augmented llm. *Preprint*, arXiv:2502.11142.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Timothy D. Wilson. 2002. *Strangers to Ourselves: Discovering the Adaptive Unconscious*. Belknap Press of Harvard University Press, Cambridge, MA.

Meghan Winters, Gavin Davidson, Diana Kao, and Kay Teschke. 2011. Motivators and deterrents of bicycling: Comparing influences on decisions to ride. *Transportation*, 38(1):153–168.

Fengyi Wu, Yifei Dong, Zhi-Qi Cheng, Yilong Dai, Guangyu Chen, Hang Wang, Qi Dai, and Alexander G Hauptmann. 2025. Govig: Goal-conditioned visual navigation instruction generation. *arXiv preprint arXiv:2508.09547*.

Wenyi Xiao, Zechuan Wang, Leilei Gan, Shuai Zhao, Zongrui Li, Ruirui Lei, Wanggui He, Luu Anh Tuan, Long Chen, Hao Jiang, Zhou Zhao, and Fei Wu. 2025. A comprehensive survey of direct preference optimization: Datasets, theories, variants, and applications. *Preprint*, arXiv:2410.15595.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering. *Preprint*, arXiv:2305.18201.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Pengfei Yang, Ngai-Man Cheung, and Xinda Ma. 2025b. Text to image generation and editing: A survey. *Preprint*, arXiv:2505.02527.

Qisheng Zeng, Zheng Gong, Songtai Wu, Caigang Zhuang, and Shaoying Li. 2024. Measuring cyclists' subjective perceptions of the street riding environment using k-means smote-rf model and street view imagery. *International Journal of Applied Earth Observation and Geoinformation*, 128:103739.

Lihong Zhang, Scott N Lieske, Dorina Pojani, Richard J Buning, and Jonathan Corcoran. 2025. Hybrid machine learning-based approaches for modeling bikeability. *Journal of Transport Geography*, 123:104150.

Tianhong Zhao, Xiucheng Liang, Wei Tu, Zhengdong Huang, and Filip Biljecki. 2023. Sensing urban soundscapes from street view imagery. *Computers, Environment and Urban Systems*, 99:101915.

Xilei Zhao, Xiang Yan, Alan Yu, and Pascal Van Hentenryck. 2020. Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models. *Travel behaviour and society*, 20:22–35.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Appendix

A.1 Recruitment And Payment

We recruited participants through social media platforms to complete anonymous surveys evaluating bicycle lane quality. A total of 427 participants completed the survey. Each survey took approximately 5 minutes to complete. To incentivize participation, we implemented a lottery-based compensation system: every 100th participant received a $100 gift card (4 participants total received compensation). This resulted in an average expected compensation of approximately $0.93 per participant, or approximately $11.16 per hour based on the estimated completion time. Participation was voluntary and anonymous, with no personally identifiable information collected. with no personally identifiable information collected.

Three graduate students were recruited from a research university to perform data annotation tasks. Annotators were paid $20 per hour for approximately 8 hours of annotation work each, totaling $160 per annotator. This compensation rate exceeds local minimum wage standards and is consistent with standard research assistant rates at our institution.

A.2 Survey Interface

Figures 5 and 6 show screenshots from our crowdsourcing survey platform. Figure 5 displays the immersive 360-degree Google Street View interface used for bikeability assessment, allowing participants to explore road environments interactively. Figure 6 shows the infrastructure preference assessment interface where participants rate their comfort levels for different cycling facility types.

A.3 Prompt Templates

We provide the prompt templates used for model inference and baseline evaluation.

Figure 5: Survey interface 1: immersive 360-degree Google Street View for bikeability assessment.

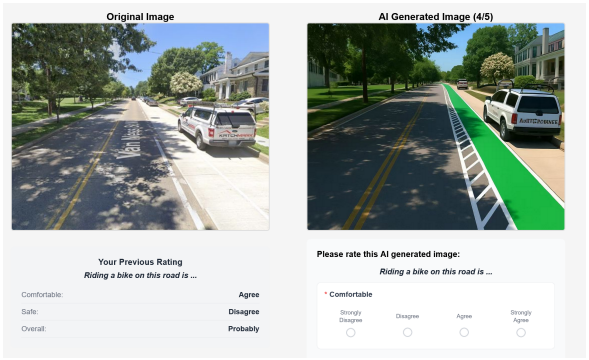

Figure 6: Survey interface 2: rating for augmented image.

A.3.1 Type 1: Full CoT Reasoning

```
As a {persona} cyclist ({persona_detailed_desc}),
analyze this street image for bikeability.

Provide a brief assessment covering:
- Key observations about the street
- Factors affecting your cycling experience
- Your comfort and safety evaluation

Rate the following on a scale of 1-4:
- Comfortable: How comfortable would you feel
  cycling here?
- Safe: How safe would you perceive this road?
- Overall: Your overall willingness to cycle
  on this road

End with:
STRUCTURED OUTPUT:
Factors: [list specific factors]
Ratings: comfortable: X, safe: Y, overall: Z
```

A.3.2 Type 2: Simplified CoT Prediction

```
As a {persona} cyclist ({persona_brief_desc}),
assess this street for bikeability.

Identify the most important factors affecting
bikeability for someone with your cycling
```

```
preferences, then rate the street.

Format your response as:
Factors: [list key factors]
Ratings: comfortable: X, safe: Y, overall: Z

Use a 1-4 scale for ratings.
```

A.3.3 Type 3: Direct Rating

```
As a {persona} cyclist ({persona_brief_desc}),
rate this street's bikeability.

Provide ratings (1-4 scale):
Ratings: comfortable: X, safe: Y, overall: Z
```

A.3.4 GPT-4o Zero-shot Baseline

```
You are an expert in urban cycling infrastructure
assessment. Analyze the provided street view image
and assess its bikeability from the perspective
of a specific cyclist persona.

Cyclist Persona: {persona}
Persona Description: {persona_desc}
{osm_text}

Task: Perform a bikeability assessment:
1. Analyze the street environment and OSM
   attributes
2. Consider how a "{persona}" cyclist would
   perceive this environment
3. Provide ratings (1-4) for: Comfortable,
   Safe, Overall
4. Output JSON with influencing_factors and
   ratings
```

A.3.5 Persona Descriptions

The four cyclist personas used in prompts:

```
Strong and Fearless: Comfortable with all
 infrastructure types, showing little preference
 between protected and unprotected facilities.

Enthused and Confident: Regular cyclists who
 prefer bike lanes but will ride in mixed
  traffic when necessary.

Interested but Concerned: Would cycle more if
 separated from traffic; requires protected
  infrastructure to feel safe.

No Way No How: Non-cyclists who find cycling
 too dangerous regardless of infrastructure.
```