arXiv:2601.03534v1 [cs.CL] 7 Jan 2026

具备人物画像感知能力且可解释的骑行适宜性评估：一种视觉语言模型方法

Yilong Dai[1]*、Ziyi Wang[2]*、Chenguang Wang[3],KexinZhou[4], YihengQian[4],Susu Xu[5],XiangYan[4][1]阿拉巴马大学， [2]马里兰大学帕克分校，[3]石溪大学， [4]佛罗里达大学， [5]约翰斯·霍普金斯大学 ydai17@crimson.ua.edu, zoewang@umd.edu, chenguangwang@stonybrook.edu, zhoukexin@ufl.edu, yihengqian@ufl.edu, susuxu@jhu.edu, xiangyan@ufl.edu

摘要

骑行适宜性评估对于推动可持续城市交通和建设骑行者友好型城市至关重要，同时需要纳入用户对安全性和舒适度的感知。然而，现有基于感知的骑行适宜性评估方法在捕捉道路环境的复杂性以及充分考量用户主观感知的异质性方面存在关键局限。本文提出一种具备人物画像感知能力的视觉语言模型框架，用于骑行适宜性评估，并作出三项创新贡献：（i）基于成熟的骑行者类型学，以理论为基础进行画像条件化，并通过思维链推理生成特定画像的解释；（ii）采用多粒度监督微调，结合稀缺的专家标注推理和丰富的用户评分，实现联合预测与可解释评估；（iii）采用人工智能赋能的数据增强，创建受控配对数据，以分离基础设施变量的影响。为测试和验证该框架，我们开发了一个基于全景图像的众包系统，并从427名骑行者处收集了12,400条基于画像的评估。实验结果表明，所提出的框架在骑行适宜性评分预测方面具有竞争力，同时独具可解释的因素归因能力。

代码与数据：
https://github.com/Dyloong1/Bikeability.git

1 引言

骑行适宜性评估是指对支持骑行的道路基础设施进行系统评估，这对于推动可持续城市交通和建设骑行者友好型城市至关重要。传统方法侧重于交通容量和几何设计等客观基础设施指标（Dowlingand Reinke，2008年；Manual，2000年），但越来越多的证据表明

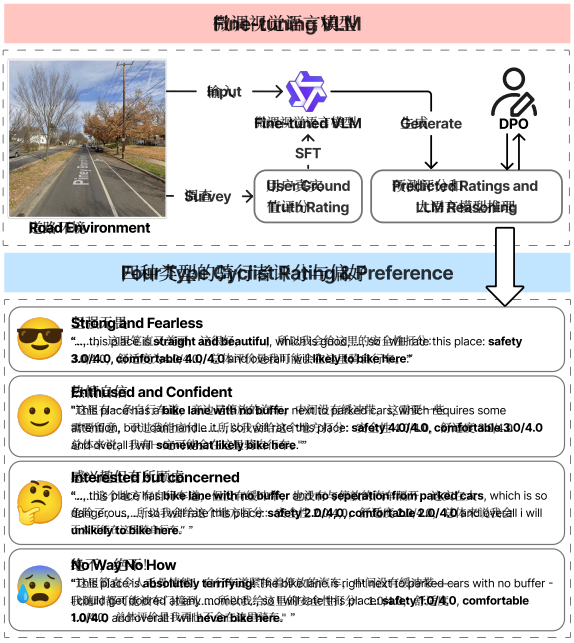


图1：我们提出的角色画像感知、可解释的骑行适宜性评估框架概览。

人们在某个地点骑行的意愿，与其说取决于这些客观测量的特征，不如说更多取决于他们对环境的感知（Guoand He，2021年；Ma 等人，2014年）。与此同时，人们对特定地点安全性的感知能够预测潜在的安全风险，往往能在事故发生前识别出危险状况（Prajap-ati and Tiwari，2013年；von Stülpnagel 等人，2022年）。因此，越来越多的呼声倡导将用户对安全性和舒适度的感知纳入骑行适宜性评估（Kellstedt 等人，2021年；Gössling and McRae，2022年；Rodriguez-Valencia 等人，2022年；Griswold 等人，2018年）。然而，现有的基于感知的评估方法存在两个主要局限：（i）无法全面衡量具有多维属性和相互作用因素的复杂道路环境，而这些因素都会影响骑行适宜性；（ii）未能恰当考虑骑行者的异质性，因为不同骑行

经历、偏好和态度的骑行者，对同一环境的感知可能有所不同。

视觉语言模型（VLMs）的最新进展为基于感知的骑行适宜性评估带来了新机遇，使得能够大规模生成高保真度的用户安全性与舒适度感知数据。视觉语言模型在捕捉和模拟人类感知及认知过程方面展现出强大能力，因而可应用于环境评估（Ito and Biljecki，2021年）、街道设计可视化（Wang 等人，2025a）以及人类感知建模（Danish 等人，2025年）等领域。这些模型能够处理结合视觉与结构化数据的多模态输入，并生成有助于提升可解释性的自然语言输出。与此同时，基于角色画像的方法通过大语言模型条件化，在捕捉不同用户行为方面展现出潜力（Wang 等人，2024年；Chen 等人，2024年）。

然而，将视觉语言模型应用于基于感知的骑行适宜性评估面临若干关键技术挑战。首先，视觉语言模型可能出现幻觉和预测不稳定等问题，从而限制其可靠性与泛化能力。其次，要有效训练视觉语言模型，既实现较高的推理准确率，又保持较强的可解释性，需要大规模收集来自不同用户的自行车友好度评分数据，以及能够记录每项评分依据的高质量“推理链”数据。然而，大规模获取这两类数据成本高昂，往往难以实现。第三，现实环境中的各种因素高度相关（例如，设有自行车道的道路往往绿化也更好）；如果校准仅依赖观测数据，这可能会限制视觉语言模型的推理能力，使其难以分离各项基础设施变量对感知的影响。

本文提出了一种基于骑行者画像的骑行适宜性评估框架（图1），通过以下创新方法应对这三项挑战：(i) 一种以理论为基础的画像条件化方法，首先依据既有类型学对骑行者进行分类（Kimand Mokhtarian，2023年；Dill and McNeil，2013年），再通过思维链推理生成针对特定画像的解释；(ii) 一种多粒度监督微调策略，结合标注深度各异的数据，共同实现评分预测、因素识别和可解释推理生成；(iii) 一种由人工智能驱动的数据增强方法，在各种自行车设施设计场景下引出不同的用户反馈，从而使模型能够更有效地分离各项基础设施变量对自行车友好度评分的感知影响。为测试和验证这一框架，我们开发了一个基于全景图像的众包调查系统，收集华盛顿特区不同道路环境中多个路段的数百名骑行者的骑行适宜性评分。

2 相关工作

2.1 骑行适宜性评估

骑行适宜性评估需要评估骑行环境的多维属性，同时考虑用户的主观感知（Kellstedt 等人，2021年；Gössling and McRae，2022年）。早期方法包括交通压力等级（LTS）框架（Dill and McNeil，2016；Mekuria 等人，2012年），但该框架过度简化了用户异质性（Damant-Sirois 等人，2014年）。传统统计模型更重视因果关系，而非预测准确率（Breiman，2001；Shmueli，2010；Zhao 等人，2020年）。近年来，数据驱动方法利用多样化的数据来源：Paranga and Oda（Paranga and Oda，2025年）对调查指标应用了主成分分析／验证性因子分析，Zhang 等人（Zhang 等人，2025年）将复杂比例评估（COPRAS）与机器学习相结合，Ito and Biljecki（Ito and Biljecki，2021年）基于计算机视觉特征训练 LightGBM，Zeng 等人（Zeng 等人，2024年）采用随机森林，而 AutoLTS（Lin etal.，2024年）则采用了对比学习。然而，这些方法要么将骑行者视为同质群体，要么依赖可能无法反映实际用户偏好的专家预先确定的权重；此外，基于视觉的方法通常需要多阶段流程，从而限制了其泛化能力。

2.2 用于城市分析的视觉语言模型

融合街景图像、遥感和地理空间数据的多模态方法，已支持多种城市分析（Gebru 等人，2017年；Dai 等人，2025年；Zhao 等人，2023年；Danish 等人，2025年；Suel 等人，2021年；Albert 等人，2017年；Cao 等人，2020年）。通义千问-VL（Yang 等人，2025a）和 LLaVA（Liu 等人，2024年）等开源视觉语言模型，如今具备强大的视觉理解能力，并支持参数高效微调（Hu 等，2021），从而能够针对专业任务进行领域自适应。这些模型通过思维链推理（Wei 等人，2022年；Kong 等人，2024年）和多粒度指令

微调（Liu 等人，2024年；Deng 等人，2025年；Wang 等人，2022年；Sanh 等人，2021年；Chung 等人，2024年），在导航（Wang 等人，2025b；Wu 等人，2025年）、基础设施评估（Ito and Biljecki，2021年）、街道设计（Wang 等人，2025a）和决策（Chen 等人，2025年）方面展现出应用前景。与此同时，图像生成模型的最新进展（Yang 等人，2025b）促进了高保真合成数据创建，为配对观测稀缺领域中的可控数据增强带来了新的可能性。然而，领域应用需要具备可解释性，以便利益相关方进行验证（Arrieta 等人，2019年；Javed 等人，2023年；Dong 等人，2025年）；现有方法则侧重于合成数据或网络规模的数据，而非专家标注稀缺的特定领域场景。对于此类主观感知任务，使模型输出与人类偏好保持一致至关重要，因为真实值反映的是个人判断，而非客观标签。基于人类反馈的强化学习（RLHF）（Ouyang 等人，2022年）和直接偏好优化（DPO）（Rafailov 等人，2024年；Bai 等人，2022年；Meng 等人，2024年；Xiao 等人，2025年）等偏好优化方法，为实现这种对齐提供了有效机制。

2.3 角色感知建模

用户异质性是基于感知的评估的基础，因为专家制定的标准往往偏离实际用户体验（Gössling andMcRae，2022年）。“四类骑行者”类型学（Dill and McNeil，2013年，2016）证实，不同骑行者类型在基础设施需求方面存在显著差异。角色条件化增强了大语言模型的行为模拟能力（Wang 等人，2024年；Li 等人，2024年；Chen 等人，2024年；Sun 等人，2025年），但现有方法依赖预定义类别或聚合数据（Chuang 等人，2024年；Park 等人，2024年），缺乏对街景差异化视觉解读进行建模的机制（Ito and Bil-jecki，2021年）。

3 方法

我们提出了一种考虑角色画像的骑行适宜性评估框架，既能应对客观道路环境评估的复杂性，也能处理用户主观感知的异质性。图2 展示了我们的流程，包含三个阶段：(1) 多源数据收集：我们通过众包调查，收集针对华盛顿特区多种道路环境的角色画像专属骑行适宜性评分及其影响因素，并通过基于人工智能的图像增强来控制基础设施变化。(2) 多粒度监督微调：我们创建三类训练数据（专家标注的完整推理链、用户提供的结构化因素—评分对以及直接评分），并按照固定采样比例联合使用这三类数据进行训练，使模型在保持稳健预测能力的同时学会可解释推理。(3) 基于偏好的推理优化：我们根据人类反馈，通过直接偏好优化（DPO）优化推理输出。

3.1 数据收集与预处理

3.1.1 研究设计与抽样

我们从华盛顿特区的道路网络中随机抽取了 200 个道路路段，并优先考虑空间分散性，以确保涵盖多样化的骑行环境。所选路段涵盖不同的机动车道配置、骑行基础设施类型（从无基础设施到完全隔离保护的设施）以及城市环境。我们将 GSV 全景 API 集成到调查系统中，使参与者能够进行详细的 360 度观察，而不局限于静态二维图像。

3.1.2 图像增强流程

在道路路段抽样过程中，我们发现，自然道路网络中很少存在理想的配对比较，即路段之间仅有极少数基础设施属性不同；即便是相邻道路，通常也会同时在多个维度上存在差异。这使模型难以从有限的调查数据中学习细粒度的属性重要性。为解决这一问题，我们开发了一套基于 AI 的图像编辑流程，使用 GPT-image-1 系统性地修改客观定义的基础设施变量，同时避免引入开发者的主观判断。这些变量包括：自行车道有无、车道宽度（窄/标准/宽）、车道颜色（绿色/无标线）、缓冲带类型（无缓冲带/标准/防撞柱/犰狳式隔离设施）以及缓冲带位置（邻近行驶中的汽车/邻近停放的汽车）。通过改变这些变量，同时保留所有其他环境因素，该流程能够支持有效的配对比较，从而隔离基础设施的影响。原始图像样本和图像编辑结果如图3所示。

3.1.3 众包调查

我们的调查收集以下信息：

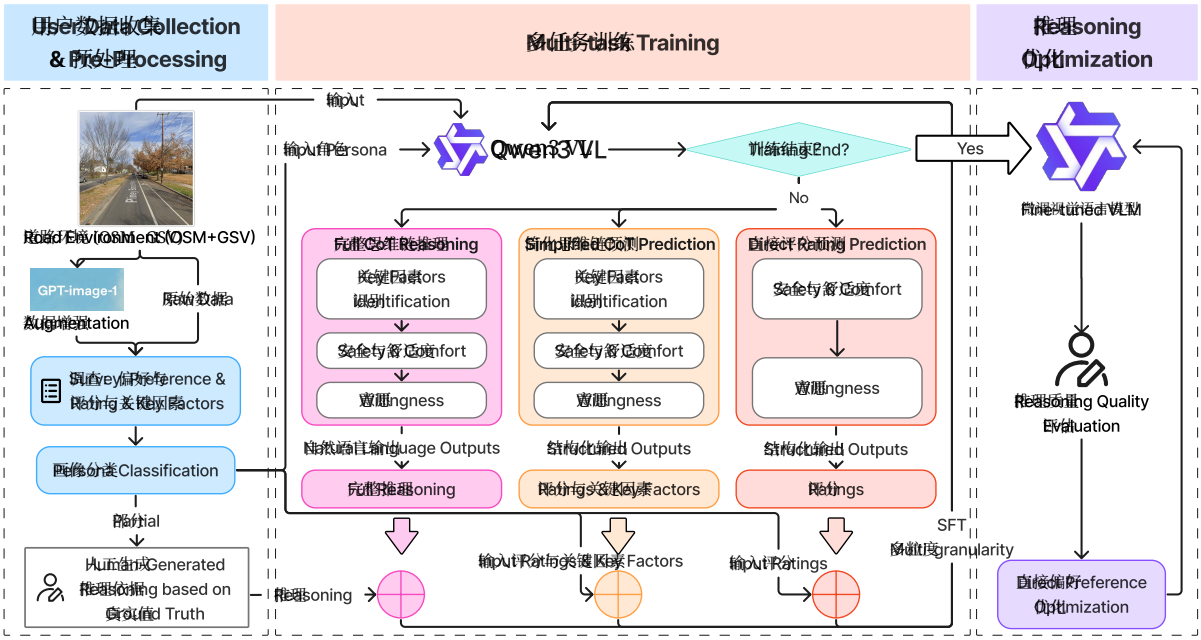


图 2：三阶段模型架构：(1) 通过众包调查和基于人工智能的图像增强进行多源数据收集；(2) 使用三种数据类型进行多粒度监督微调（Type 1：完整推理，Type 2：因素-评分对，Type 3：仅评分）；(3) 通过直接偏好优化进行基于偏好的推理优化。

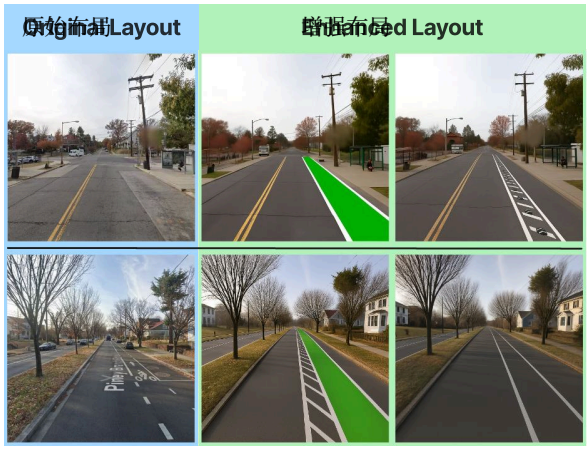


图 3：基于人工智能的图像增强：原始街景图像（左）以及通过控制基础设施变化而系统修改后的版本（右）。

人口统计特征：性别、年龄组、种族/族裔和教育水平，以确保具有多样性的代表性。

骑行偏好：使用参考图像对八种基础设施类型进行舒适度评分（5 分制： 1=非常不舒适至 5=非常舒适）：(1) 无自行车道，(2) 道路路肩，(3) 路外多用途步道，(4) 带有共享车道标记的混行车道，(5) 人行道，(6) 标线自行车道，(7) 缓冲式自行车道，以及 (8) 隔离式自行车道。这些评分有助于识别骑行者画像，体现个人的风险承受度和舒适度

阈值。

路段评估：针对随机分配给每位受访者的 15 个地点（经数据增强后为 20 个），我们收集了：(1) 安全性评分 和 (2)舒适度评分，采用 4 点李克特量表（1=非常不同意至 4=非常同意）；(3) 骑行意愿，采用 4 点量表（1=从不至 4=绝对愿意）；以及 (4) 通过多选标签界面收集的影响因素，该界面提供预设选项（自行车道状况、交通状况、隔离设施、环境）和开放式填写栏。

3.1.4 画像分类

我们的调查收集了 427 名参与者的答复。每位参与者首先对八种基础设施类型的舒适度进行评分（1–5 分），随后至少对 20 张街景图像的骑行适宜性进行评估，最终获得 12,400 项基于画像的评估（均值：每位参与者 29.0 张图像）。

为验证角色画像感知建模的必要性，我们分析了参与者内部的评分方差。图4 显示出显著的参与者间异质性：街道路段评分的方差介于 0 至 2.17 之间（中位数 = 0.872），且与平均评分水平正交（r = −0.10），这证实了这些差异源于真实的个体差异，而非量表使用伪差。

我们按照“四类骑行者”

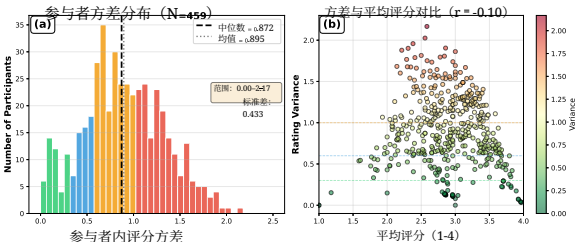


图4：427名参与者的参与者内评分方差分布。(a) 直方图显示出显著的异质性（中位数 = 0.872，范围：0–2.17）。(b) 散点图证实方差与平均评分正交（r = −0.10）。

类型划分法，将参与者分为四类骑行者原型（Dill andMcNeil，2016）。我们计算了三种防护等级下的平均舒适度评分：低（无自行车道、混合车道）、中（划线车道、路肩、非道路自行车道）和高（缓冲/受保护车道），并考察了从低到高防护等级的变化幅度。根据这些指标进行聚类后，发现了四个群体：总体舒适度较低且变化幅度最小的参与者归为完全拒绝型（NWNH）；舒适度较高且变化平缓的参与者归为强势无畏型（S&F）；变化幅度较大的参与者构成感兴趣但担忧型（IbC）群体；而舒适度和变化幅度均适中的参与者则归为热情且自信型（E&C）。表1 展示了各群体的分布。

表1：角色画像分布及基础设施偏好特征（N=427）。均值和梯度（gradient）根据八种基础设施类型的舒适度评分（1-5分制）计算得出。

角色画像	Pct.	Mean	梯度。
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 任务定义

我们将角色画像感知的骑行适宜性评估建模为条件生成任务，视觉语言模型根据骑行者画像和多模态输入生成主观评估。通过将街道层面的视觉观察与结构化基础设施属性结合起来作为输入，我们确保了全面覆盖，同时避免特征工程中的人为选择偏差。

形式上，给定骑行者画像 P，模型会生成一段主观评价文本 $X = \{x_1, x_2, ..., x_S\}$，由 S 个词组成，用于从该类角色的视角描述骑行适宜性评估。模型的输入包括：

- 一种骑行者画像 P，分为四种类型：S&F、E&C、IbC 或 NWNH
- O：一张捕捉骑行者第一人称视角的街景图像，提供丰富的视觉背景信息
- $A = \{a_1, a_2, ..., a_I\}$：来自 OpenStreetMap 的一组 I 基础设施属性（例如道路类型、车道配置），确保对非视觉因素的全面覆盖

给定来自调查数据的用户真实评价 X^*，我们优化模型参数 θ，以最大化生成正确评价的条件概率：

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

遵循既有的《交通》理论（DillandMcNeil，2016；Winterset al.，2011年）以及我们对系统性参与者间异质性的实证验证（第3.1.4节，图4），我们将特定画像评估视为一项基本设计要求，而非可选功能。

3.2.1 多格式训练数据

由于经过人工标注的完整思维链（CoT）推理数据十分稀缺且耗时耗力，我们设计了三种具有不同监督粒度的训练数据格式。我们的多粒度策略发挥了互补优势：Type 1 支持生成可解释的推理，Type 2 培养因素识别能力，Type 3 确保稳健的评分预测。通过以固定采样比例（第4.1节）联合训练这三种类型，模型能够学习到一种统一表征，在支持多种输出格式的同时，保持不同解释深度之间的一致性。这三种格式共享相同的多模态输入（P、O、A），但其提示指令和监督信号有所不同。

Type 1：完整思维链推理。这种格式旨在训练模型生成可解释的自然语言推理。然而，直接从众包调查中获取完整的推理链颇具挑战，原因包括：(i)认知负担导致完成率降低（Krosnick，1991年；

Deutskens 等人，2004年）；(ii) 表达方式不一致 会产生噪声，增加训练难度；以及 (iii)隐性推理，即感知判断在潜意识中发生（Nisbett 和 Wilson，1977年；Wilson，2002年）。

鉴于这些限制，我们通过交通研究人员的专家标注构建完整的推理链。关键在于，这与传统的专家主导方法不同：专家并非预先定义哪些因素重要，而是根据调查回复中的用户提供因素和评分，综合形成连贯的推理链。这种数据驱动方法确保推理模式源自真实的用户偏好。模型接收：

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

其中，提示$_{\text{reason}}$ 指示模型：（i）分析图像中的环境条件；（ii）识别影响骑行者感知的关键因素；（iii）预测安全性和舒适度评分；（iv）综合得出总体意愿评估。监督信号为：

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

其中，交通研究人员使用自然语言撰写推理链，将真实因素 T^*和评分 R^* 联系起来。尽管大型教师模型（例如 GPT-4）可以自动完成标注，但这类方法会带来幻觉风险和不一致的模式（Bang 等人，2023年；Xu 等人，2023年）。我们的专家标注确保监督信号具有高质量，能够忠实呈现参与者视角。

Type 2：简化思维链预测。与 Type 1 不同，这种格式不包含人工撰写的推理叙述，而是直接根据调查回答得出。模型接收：

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

其中，prompt$_{\text{struct}}$ 指示模型执行简化的思维链过程，省略客观条件分析步骤：(i) 识别对骑行者感知影响最大的关键因素，(ii) 预测安全性和舒适度评分，以及 (iii) 提供总体使用意愿评估。监督信号由参与者的配对标注构成：

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

其中，$T^* = \{t_1, t_2, ..., t_n\}$ 表示参与者识别出的影响因素集合，而 R^* 包含相应的三维评分。这种格式通过明确的因素识别，在感知与评价之间建立联系，而无需完整的解释性叙述。

Type 3：直接评分预测。这种格式进一步简化了任务，省去因素识别步骤，并利用所有可用的调查评分以最大限度提高数据效率。模型接收：

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

其中，prompt$^{\text{rating}}$ 要求直接进行评分预测：模型预测安全性和舒适度评分，随后给出总体使用意愿评估，无需进行中间的因素识别或推理。监督信号由三维评分向量构成：

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

这种格式体现了基本的刺激—反应映射，提供了最直接的监督信号，并确保具备稳健的端到端预测能力。

3.2.2 模型训练

为在调查数据有限的情况下高效适配模型并防止过拟合，我们在预训练的 Qwen3-VL-8B-Instruct 基础模型上采用 LoRA（Hu 等，2021年）进行参数高效微调。

多粒度指令微调。 由于人工标注的完整思维链推理（Type 1）十分稀缺且耗时费力，我们借鉴 LLaVA-1.5 的方法，采用多粒度指令微调方案（Liu 等人，2024年）。我们的训练数据呈现出自然的粒度差异：虽然全部 12,400 条调查回复都包含评分（Type 3）和用户识别的因素（Type 2），但只有约 2,000 个样本包含专家标注的推理链（Type 1）。这一数据限制使多粒度训练成为必要：仅依赖 Type 1 会导致样本不足，难以学习稳健的视觉语言表征；反之，仅使用 Type 3 则会牺牲所有可解释性。我们的多粒度策略发挥了不同类型的互补优势：Type 1 支持生成可解释的推理，Type 2 培养因素识别能力，Type 3 则确保评分预测的稳健性。通过按照固定采样比例（第4.1节）联合使用这三种类型进行训练，模型能够学习统一表征，支持多种输出格式，同时在不同解释深度之间保持一致性。

直接偏好优化。 在监督微调之后，我们采用直接偏好优化（DPO）（Rafailov 等人，2024年）对 Type 1 输出进行优化，以提升推理质量。为收集偏好数据，我们从训练集中抽取实例，并使用监督微调模型为每个实例生成两条 Type 1 推理解释，采用不同的采样温度。三位交通领域专家分别对两条解释进行独立评估，并根据以下标准选出偏好项：(i) 事实准确性，(ii) 逻辑连贯性，以及 (iii) 角色一致性。通过多数投票确定最终偏好对。按照 DPO 框架，我们直接优化策略，提高偏好响应的可能性，而无需单独训练奖励模型。DPO 目标采用参考模型正则化，防止模型偏离监督微调初始化，同时保留其在 Type 2 和 Type 3 格式上的性能。

4 实验

4.1 实验设置

实现细节。 所有实验均在一块英伟达 A100 80GB 图形处理器上进行。对于监督式微调，我们采用秩为$r = 32$ 、缩放因子为 $\alpha = 64$的 LoRA，并以查询、键、值和输出投影矩阵（q_proj、k_proj、v_proj、o_proj）为目标。我们使用 AdamW 优化器，初始学习率为2×10^{-4}，训练 5 个 epoch，并采用余弦退火学习率调度。学习率遵循衰减调度：前 3 个 epoch 温和衰减（每个 epoch $0.8\times$ ），其余 epoch 快速衰减（每个 epoch $0.5\times$ ）。批量大小设为 4，并进行 4 步梯度累积，使有效批量大小达到 16。我们采用梯度检查点来管理内存使用，并设置 10 % 的预热步数。为提高效率，采用半精度浮点数（FP16）进行混合精度训练。

多粒度采样比例。 我们在所有训练轮次中按照固定比例采样训练数据：15% 为 Type 1（完整推理），40% 为 Type 2（结构化因素-评分对），45% 为 Type 3（仅评分）。这一分布与 LLaVA-1.5 的研究发现一致：详细推理样本占比为 15% 时，性能最佳（ Liu 等人，2024 年）。固定比例设计确保：(i) 高效利用稀缺的专家标注，避免过度采样（15% 的比例意味着每轮有 ∼1,860 个 Type 1 样本，接近我们 2,000 个样本的完整预算），(ii) 在各类监督级别之间实现均衡的多任务学习，以及 (iii) 通过避免灾难性遗忘来保障训练稳定性（Kirkpatrick 等人，2017年）。我们通过明确的提示指令和后处理纠正，严格执行评分约束（1–4 分制）。

直接偏好优化训练细节。在偏好数据收集过程中，我们随机抽取 500 个实例，并以 0.7 和 1.0 的采样温度为每个实例生成两条解释，经专家多数投票后得到 500 对偏好数据。直接偏好优化训练采用正则化系数 $\beta = 0.1$、学习率 5×10^{-6}、批量大小 8，并训练 3 个轮次。

基线。我们与三个基线进行比较：(1)*GPT-4o* 零样本：向 GPT-4o 提供角色描述和多模态输入，不进行特定任务训练；(2)*Kmeans-SMOTE* 随机森林 (Zeng 等人，2024年)：为确保比较公平，我们复现了 Zeng 等人基于预处理的方法，使用 YOLOv8 全面检测可观察到的道路基础设施要素（自行车道、缓冲带、交通信号灯、街道设施、绿化等），将这些计算机视觉特征与 OpenStreetMap 属性（机动车道数量、限速、自行车道类型）及图像提取的潜在表示（通过预训练的 ResNet-50 编码）相结合。随后，按照其 Kmeans-SMOTE 平衡策略，将这些多源特征输入随机森林分类器，并加入经大语言模型归一化的因素标签池，以扩展因素识别能力；(3) *Kmeans-SMOTE* 随机森林（仅评分）：采用相同的预处理和特征提取流程，但仅针对评分预测进行训练，以单独考察联合标签—评分学习的作用。

评估指标。 对于评分预测，我们报告三个维度（安全性、舒适度、使用意愿）的平均指标：平均绝对误差（MAE）、精确匹配率（EM）、一分以内准确率（W1，在 ±1以内的百分比）以及皮尔逊相关系数。对于因素识别，我们使用句子嵌入（all-MiniLM-L6-v2）和阈值为 0.7 的贪心匹配，计算语义精确率、召回率和 F1 分数。对于推理质量，我们使用 GPT-4o 作为自动评审（Zheng 等人，2023年），评估：(1)事实准确性，即对基础设施特征的正确描述；(2)逻辑连贯性，即表述清晰且逻辑流畅；(3)角色一致性，即与角色画像特定关注点的一致程度。

表 2：比较评分预测和因素识别性能的主要实验结果。MAE：平均绝对误差，EM：完全匹配，W1：一分以内准确率，Corr：皮尔逊相关系数，Prec：精确率，Rec：召回率，F1：F1分数（阈值为=0.7）。

方法	评分预测				因素识别		
	MAE↓	EM↑	W1↑	皮尔逊相关系数↑	精确率↑	Rec↑	F1↑
GPT-4o 零样本	1.00	0.30	0.70	0.25	0.12	0.08	0.10
KS-RF（仅评分）	**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 主要结果

表 2 展示了我们的主要结果。对于评分预测，我们的方法取得了具有竞争力的性能（MAE：0.71，EM：0.41，皮尔逊相关系数：0.48），与专门的回归方法 KS-RF（仅评分）基本相当，同时实现了最高的一分以内准确率（W1：0.87）。

在因素识别方面，我们的方法达到了 0.52 的精确率、0.46 的召回率和 0.49 的 F1 分数，显著优于 KS-RF（F1：0.31）和 GPT-4o 零样本（F1：0.10）。这表明，尽管生成因素与真实因素之间存在措辞差异，领域特定训练仍能准确识别骑行适宜性相关因素。

4.3 人类偏好对齐

表 3：人类偏好对齐结果。Acc.：事实准确性，Coh.：逻辑连贯性，Cons.：角色一致性。

方法	Acc.↑	Coh.↑	Cons.↑
GPT-4o	0.25	0.694	0.995
我们的（监督微调）	0.58	0.580	0.920
我们的+直接偏好优化	**0.59**	**0.610**	**0.950**
提升	+1.7%	+5.2%	+3.3%

表3 显示，直接偏好优化提升了所有维度的解释质量。我们的监督微调模型事实准确率显著高于 GPT-4o（0.58 对 0.25），体现了领域特定训练的效果；而 GPT-4o 凭借更大的规模，在连贯性和一致性方面表现更强。直接偏好优化缩小了这一差距，其中连贯性的提升最大（+5.2%），这验证了偏好优化能够改善解释风格和逻辑流畅性。

4.4 消融研究

多粒度数据消融。 我们评估了不同监督粒度的贡献。仅使用 Type 3（仅评分）进行训练时，平均绝对误差为 0.75，一分以内准确率为 0.85，但无法进行因素识别或提供解释。加入 Type 2（因素-评分对）后，预测性能提升至平均绝对误差 0.73、一分以内准确率 0.86，但仍然缺乏推理能力。

采用 Type 1（思维链推理）并按 15/40/45 的比例配置的完整方法取得了最佳准确率（平均绝对误差：0.71，一分以内准确率：0.87），同时能够识别因素（F1 分数：0.49）并进行可解释推理。这表明，思维链为联合优化提供了最丰富的监督信号。固定比例策略确保了训练稳定性，因为逐轮大幅调整比例会导致收敛不稳定和灾难性遗忘。

5 结论

我们提出了一种具备角色画像感知能力的视觉语言模型方法，用于评估骑行适宜性，以应对基础设施评估的复杂性和用户感知的异质性。研究结果表明，可解释性并不必然以牺牲准确性为代价：我们的框架在实现有竞争力的评分预测的同时，还具备现有方法所不具备的独特能力，包括因素识别和可解释推理。

对于研究人员而言，我们的框架在适配新的城市环境时，无需重复执行预处理流程。对于规划者而言，针对不同角色画像的评估能够揭示“平均”骑行者之外多元化的社区需求，为制定更具包容性的基础设施决策提供依据。我们希望这项工作能推动视觉语言模型在交通研究中的广泛应用，支持以人为本、兼顾用户多样性的基础设施评估。

6 局限性

我们的研究存在若干局限。首先，尽管该框架具有可迁移性，但基于华盛顿特区数据训练的模型可能需要针对基础设施风格不同的城市进行微调。其次，

我们的四类别骑行者类型学是对连续偏好谱的离散近似。第三，静态街景图像无法捕捉交通流量、天气或时间变化等动态因素。最后，与 GPT-4o 等更大型模型相比，我们的模型在通用语言能力方面仍存在差距。

7 致谢

我们感谢交通部（District Department of Transportation）对本项目的支持。本项目获得美国国家科学基金会的支持（资助编号：2425029）。

参考文献

Adrian Albert、Jasleen Kaur 和 Marta C Gonzalez。2017年。利用卷积网络和卫星图像大规模识别城市环境中的模式。载于 第*23*届 *ACM SIGKDD* 国际知识发现与数据挖掘会议论文集，第1357–1366 页。

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 和 Francisco Herrera。2019年。 可解释人工智能（XAI）：概念、分类体系、机遇，以及迈向负责任人工智能的挑战。 预印本，arXiv:1910.10045。

Yuntao Bai、Saurav Kadavath、Sandipan Kundu、A manda Askell、Jackson Kernion、Andy Jones、Anna Chen、Anna Goldie、Azalia Mirhoseini、Cameron McKinnon、Carol Chen、Catherine Olsson、Christo- pher Olah、Danny Hernandez、Dawn Drain、Deep Ganguli、Dustin Li、Eli Tran-Johnson、Ethan Perez 及其他 32 位作者。2022年。 宪法式人工智能：来自人工智能反馈的无害性。 预印本，arXiv:2212.08073。

Yejin Bang、Samuel Cahyawijaya、Nayeon Lee、Wen- liang Dai、Dan Su、Bryan Wilie、Holy Lovenia、Zi- wei Ji、Tiezheng Yu、Willy Chung、Quyet V. Do、Yan Xu 和 Pascale Fung。2023年。 对 ChatGPT 在推理、幻觉和交互性方面进行的多任务、多语言、多模态评测。 预印本，arXiv:2302.04023。

Leo Breiman. 2001. 统计建模：两种文化（附作者的评论和回应）。*Statistical Science*，16(3):199–231。

Rui Cao、Wei Tu、Cuixin Yang、Qing Li、Jun Liu、Jia- song Zhu、Qian Zhang、Qingquan Li 和 Guoping Qiu。2020年。基于深度学习的遥感与社会感知数据融合，用于城市区域功能识别。*I SPRS*摄影测量与遥感杂志，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 和 Yanghua Xiao。2024年。从角色画像到个性化：角色扮演语言智能体调查。预印本，arXiv:2404.18231。

Ruxiao Chen、Chenguang Wang、Yuran Sun、Xilei Zhao 和 Susu Xu。2025年。从感知到决策：基于行为理论的大语言模型进行野火疏散决策预测。预印本，arXiv:2502.17701。

Yun-Shiuan Chuang、Krirk Nirunwiroj、Zach Studdi- ford、Agam Goyal、Vincent V. Frigo、Sijia Yang、Dha- van Shah、Junjie Hu 和 Timothy T. Rogers。2024年。超越人口统计特征：利用人类信念网络对齐角色扮演型大语言模型智能体。预印本，arXiv:2406.17232。

Hyung Won Chung、Le Hou、Shayne Longpre、Barret Zoph、Yi Tay、William Fedus、Yunxuan Li、Xuezhi Wang、MostafaDehghani、Siddhartha Brahma 等另外 1 位作者。2024年。指令微调语言模型的规模化。机器学习研究期刊，25(70):1–53。

Yilong Dai、Luyu Liu、Kaiyue Wang、Meiqing Li 和 Xiang Yan。2025年。利用计算机视觉和街景图像评估公交车站设施。《计算机、环境与城市系统》，117:102254。

Gabriel Damant-Sirois、Michael Grimsrud 和 Ahmed M El-Geneidy。2014年。你属于哪种类型：多维骑行者类型学。《交通》，41(6):1153–1169。

Matthew Danish、SM Labib、Britta Ricker 和 Marco Helbich。2025年。利用开放街景图像收集人类对城市环境感知的公民科学工具包。《计算机、环境与城市系统》，116:102207。

Huilin Deng、Ding Zou、Rui Ma、Hongchen Luo、Yang Cao 和 Yu Kang。2025年。通过课程强化学习提升视觉语言模型的泛化能力和推理能力。预印本，arXiv:2503.07065。

Elisabeth Deutskens、ko de ruyter、Martin Wetzels 和 Paul Oosterveld。2004年。互联网调查的响应率与响应质量：一项实验研究。*MarketingLetters*，15：21–36。

Jennifer Dill 和 Nathan McNeil。2013年。四类骑行者？检视骑行者类型学，以更好地理解骑行行为及其潜力。交通研究记录，2387(1)：129–138。

Jennifer Dill 和 Nathan McNeil。2016。《重新审视四类骑行者：来自一项全国调查的发现》。《交通研究记录》，2587(1)：90–99。

Yifei Dong、Fengyi Wu、Kunlin Zhang、Yilong Dai、Sanjian Zhang、Wanghao Ye、Sihan Chen 和 Zhi- Qi Cheng。2025年。金融领域的大语言模型智能体：连接研究、实践与现实部署的调查。载于计算语言学协会 *EMNLP 2025* 年会议论文集，第 17889–17907 页，中国苏州。计算语言学协会。

Richard Dowling 和 David Reinke。2008年。城市街道多模式服务水平分析：用户指南。《交通研究委员会》，美国华盛顿哥伦比亚特区。

Timnit Gebru、Jonathan Krause、Yilun Wang、Duyun Chen、Jia Deng、Erez Lieberman Aiden 和 Li Fei-Fei。2017年。利用深度学习和谷歌街景估算美国各地社区的人口构成。美国国家科学院院刊，114(50)：13108–13113。Stefan Gössling 和 Sophia McRae。2022年。主观安全的骑行基础设施：对城市设计的新见解。《交通地理学杂志》，101：103340。Julia B Griswold、Mengqiao Yu、Victoria Filingeri、Offer Grembek 和 Joan L Walker。2018年。一种用于评估自行车服务水平的行为建模方法。《交通研究*A*辑：政策与实践》，116：166–177。Yuanyuan Guo 和 Sylvia Y He。2021年。客观与感知建成环境在影响无桩共享单车作为地铁通勤接驳方式选择中的作用。《交通研究*A*辑：政策与实践》，149：377–396。Edward J. Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang 和 Weizhu Chen。2021年。LoRA：大语言模型的低秩适配。预印本，arXiv:2106.09685。Koichi Ito 和 Filip Biljecki。2021年。利用街景图像和计算机视觉评估骑行适宜性。《交通研究*C*辑：新兴技术》，132：103371。Abdul Rehman Javed、Waqas Ahmed、Sharnil Pandya、Praveen Kumar Reddy Maddikunta、Mamoun Alazab 和 Thippa Reddy Gadekallu。2023年。智能城市可解释人工智能调查。《电子学》，12(4)。Debra K Kellstedt、John O Spengler、Margaret Foster、Chanam Lee 和 Jay E Maddock。2021年。骑行适宜性评估方法范围综述。《社区健康杂志》，46(1)：211–224。Sung Hoo Kim 和 Patricia L Mokhtarian。2023年。交通领域的有限混合（或潜在类别）建模：趋势、应用、潜力与未来方向。《交通研究*B*辑：方法论》，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 和 Raia Hadsell。2017年。克服神经网络中的灾难性遗忘。美国国家科学院院刊，114(13)：3521–3526。

Xianghao Kong、Jinyu Chen、Wenguan Wang、Hang Su、Xiaolin Hu、Yi Yang 和 Si Liu。2024年。基于思维链提示的可控导航指令生成，第 37–54 页。施普林格·自然瑞士。

Jon A. Krosnick。1991年。调查问卷中应对态度测量认知负荷的应答策略。应用认知心理学，5(3)：213–236。Xuchuan Li、Fei Huang、Jianrong Lv、Zhixiong Xiao、Guolong Li 和 Yang Yue。2024年。更加真实：使用大语言模型智能体和个体画像生成旅行日记。预印本，arXiv:2407.18932。Bo Lin、Shoshanna Saxe 和 Timothy CY Chan。2024年。通过对比学习和空间后处理实现自动化骑行压力评估。在人工智能促进协会年会论文集，第38卷，22222–22230页。Haotian Liu、Chunyuan Li、Yuheng Li 和 Yong Jae Lee。2024年。通过视觉指令微调改进基线模型。预印本，arXiv:2310.03744。Liang Ma、Jennifer Dill 和 Cynthia Mohr。2014年。客观环境与感知环境：对于骑行而言，什么更重要？《交通》，41(6)：1135–1152。《公路通行能力手册》。2000年。《公路通行能力手册》。华盛顿哥伦比亚特区，2(1)：1。Maaza C Mekuria、Peter G Furth 和 Hilary Nixon。2012年。低压力骑行与网络连通性。技术报告 CA-MTI-12-1005，Mineta 交通运输研究所。Yu Meng、Mengzhou Xia 和 Danqi Chen。2024年。Simpo：使用无参考奖励进行简单偏好优化。预印本，arXiv:2405.14734。Richard Nisbett 和 Timothy Wilson。1977年。言不尽意：关于心理过程的口头报告。《心理学评论》，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 和 Ryan Lowe。2022年。训练语言模型遵循人类反馈中的指令。预印本，arXiv:2203.02155。

Norbert Anthony Gerome Paranga 和 Takanori Oda。2025年。在步行适宜性和骑行适宜性评估中关联客观特征（道路清单）与主观特征（用户感知）。《*TransportationResearch Procedia*》，82：1452–1471。

Joon Sung Park、Carolyn Q. Zou、Aaron Shaw、Benjamin Mako Hill、Carrie Cai、Meredith RingelMorris、Robb Willer、Percy Liang 和 Michael S. Bernstein。2024年。对 1,000 人进行生成式智能体模拟。预印本，arXiv:2411.10109。

Pankaj Prajapati 和 Geetam Tiwari。2013年。实际事故风险与感知事故风险之间关系的研究。*Procedia-Social and Behavioral Sciences*，104：1095–1104。

Rafael Rafailov、Archit Sharma、Eric Mitchell、Stefano Ermon、Christopher D. Manning 和 Chelsea Finn。2024年。直接偏好优化：你的语言模型实际上是一个奖励模型。预印本，arXiv:2305.18290。

Alvaro Rodriguez-Valencia、Jose Agustin Vallejo-Borda、German A Barrero 和 Hernan Alberto Ortiz-Ramirez。2022年。构建面向行人和骑行者基础设施的服务评估丰富框架：承认用户感知的力量。《交通》，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及另外1位作者。2021年。多任务提示训练实现零样本任务泛化。*arXiv*预印本 *arXiv:2110.08207*。Galit Sh mueli。2010年。解释还是预测？统计科学，第289–310页。Esra Suel、Samir Bhatt、Michael Brauer、Seth Flaxman和Majid Ezzati。2021年。利用卫星和街景图像进行多模态深度学习，以测量城市地区的收入、过度拥挤程度和环境匮乏程度。*RemoteSensing ofEnvironment*，257:112339。Chenkai Sun、Ke Yang、Revanth Gangi Reddy、Yi R. Fung、Hou Pong Chan、Kevin Small、ChengXiang Zhai和Heng Ji。2025年。Persona-db：通过协作式数据精炼实现高效的大语言模型个性化响应预测。预印本，arXiv:2402.11060。Rul von Stülpnagel、Chayenne Petinaud和Sven Lißner。2022年。城市骑行期间的事故风险与主观风险感知：考虑骑行量。*Accident Analysis& Prevention*，164:106470。Chenguang Wang、Xiang Yan、Yilong Dai、Ziyi Wang和Susu Xu。2025a。从图像生成到基础设施设计：用于街道设计生成的多智能体流程。*arXiv*预印本 *arXiv:2509.05469*。

Jiawei Wang、Renhe Jiang、Chuang Yang、Zengqing Wu、Makoto Onizuka、Ryosuke Shibasaki、Noboru Koshizuka和Chuan Xiao。2024年。作为城市居民的大语言模型：用于个人出行生成的大语言模型智能体框架。预印本，arXiv:2402.14744。

Xuezhi Wang、Jason Wei、Dale Schuurmans、Quoc Le、Ed Chi、Sharan Narang、Aakanksha Chowdhery 和 Denny Zhou。2022年。自一致性提升了语言模型中的思维链推理。*arXiv* 预印本 *arXiv:2203.11171*。

Zihan Wang、Yaohui Zhu、Gim Hee Lee 和 Yachun Fan。2025b。Navrag：通过检索增强型大语言模型为具身导航生成用户需求指令。预印本，arXiv:2502.11142。

Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Fei Xia、Ed Chi、Quoc V Le、Denny Zhou 和另外 1 位作者。2022年。思维链提示能够激发大语言模型的推理能力。神经信息处理系统进展，35:24824–24837。

Timothy D. Wilson. 2002年。《陌生的自己：发现适应性无意识》。哈佛大学出版社贝尔纳普出版社，马萨诸塞州剑桥。

Meghan Winters、Gavin Davidson、Diana Kao 和 Kay Teschke。2011年。骑行的激励因素与阻碍因素：比较影响骑行决策的因素。《交通》，38(1):153–168。

Fengyi Wu、Yifei Dong、Zhi-Qi Cheng、Yilong Dai、Guangyu Chen、Hang Wang、Qi Dai 和 Alexan- der G Hauptmann。2025年。Govig：目标条件视觉导航指令生成。*arXiv* 预印本 *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 和 Fei Wu。2025年。直接偏好优化综合综述：数据集、理论、变体与应用。预印本，arXiv:2410.15595。Fangyuan Xu、Yixiao Song、Mohit Iyyer 和 Eun- sol Choi。2023年。长篇问答评估的批判性评价。预印本，arXiv:2305.18201。An Yang、Anfeng Li、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chang Gao、Chengen Huang、Chenxu Lv 等。2025a。通义千问3技术报告。*arXiv* 预印本 *arXiv:2505.09388*。Pengfei Yang、Ngai-Man Cheung 和 Xinda Ma。2025b。文生图生成与编辑：一项调查。预印本，arXiv:2505.02527。Qisheng Zeng、Zheng Gong、Songtai Wu、Caigang Zhuang 和 Shaoying Li。2024年。使用K均值SMOTE-RF模型和街景测量骑行者对街道骑行环境的主观感知

图像。《国际应用地球观测与地理信息学杂志》，128:103739。

Lihong Zhang、Scott NLieske、DorinaPojani、Richard J Buning 和 Jonathan Corcoran。2025年。用于骑行适宜性建模的混合机器学习方法。《交通地理学杂志》，123:104150。

Tianhong Zhao、Xiucheng Liang、WeiTu、Zhengdong Huang 和 Filip Biljecki。2023年。利用街景图像感知城市声景。《计算机、环境与城市系统》，99:101915。

Xilei Zhao、Xiang Yan、Alan Yu 和 Pascal Van Hent enryck。2020年。出行方式选择的预测与行为分析：机器学习与逻辑回归模型的比较。《出行行为与社会》，20:22–35。

Lianmin Zheng、Wei-LinChiang、Ying Sheng、Siyuan Zhuang、Zhanghao Wu、Yonghao Zhuang、Zi Lin、Zhuohan Li、Dacheng Li、Eric P. Xing、Hao Zhang、Joseph E. Gonzalez 和 Ion Stoica。2023年。使用 MT-Bench 和 Chatbot Arena 评判 LLM 评审。预印本，arXiv:2306.05685。

A 附录

A.1 招募与报酬

我们通过社交媒体平台招募参与者完成匿名调查，以评估自行车道质量。共有 427 名参与者完成了调查问卷。每份调查问卷约需 5 分钟完成。为鼓励参与，我们采用了基于抽奖的补偿机制：每第 100 名参与者可获得一张价值 100 美元的礼品卡（共有 4 名参与者获得补偿）。据此估算，每位参与者的平均预期补偿约为 0.93 美元；按估计的完成时间计算，时薪约为 11.16 美元。参与完全自愿且匿名，不收集任何个人身份信息。未收集任何个人身份信息。

我们从一所研究型大学招募了三名研究生，负责执行数据标注任务。标注人员的时薪为 20 美元，每人约完成 8 小时的标注工作，因此每位标注人员共获得 160 美元报酬。该薪酬高于当地最低工资标准，也符合本机构的研究助理薪酬标准。

A.2 调查界面

图5 和6 展示了我们的众包调查平台截图。图5展示了用于骑行适宜性评估的沉浸式 360 度谷歌街景界面，该界面允许参与者交互式探索道路环境。图6 展示了基础设施偏好评估界面，参与者可在其中对不同骑行设施类型的舒适度进行评分。



图 5：调查界面 1：用于骑行适宜性评估的沉浸式 360 度谷歌街景。

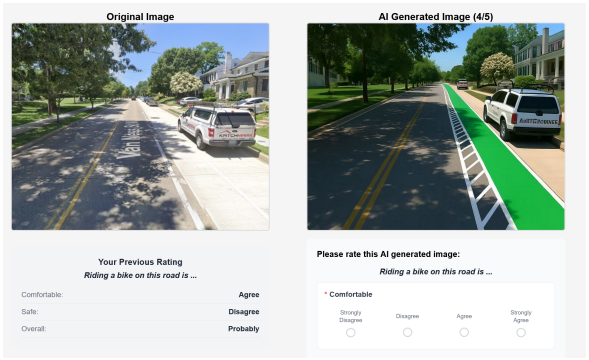


图 6：调查界面 2：对增强图像进行评分。

A.3 提示模板

我们提供了用于模型推理和基线评估的提示模板。

A.3.1 类型 1：完整思维链推理

作为一名{persona}骑行者（{persona_详细_描述}），请分析这张街道图像的骑行适宜性。

请提供一份简要评估，涵盖：- 关于街道的主要观察结果 - 影响您骑行体验的因素 - 您对舒适度和安全性的评估

请按 1-4 分进行评分：- 舒适度：您觉得在这里骑行有多舒适？- 安全性：您觉得这条道路有多安全？- 总体评价：您在这条道路上骑行的总体意愿

结尾：
结构化输出：
因素：［列出具体因素］
评分：舒适度：X，安全性：Y，总体评价：Z

A.3.2 类型2：简化思维链预测

作为一名{persona}骑行者（{persona_简要_描述}），请评估这条街道的骑行适宜性。

确定影响具有您的骑行偏好的人员骑行适宜性的最重要因素，然后对街道进行评分。

请按以下格式回复：因素：［列出关键因素］评分：舒适度：X，安全性：Y，总体评价：Z

评分采用1-4分制。

A.3.3 类型3：直接评分

作为一名{persona}骑行者（{persona_简要_描述}），请评价这条街道的骑行适宜性。

请提供评分（1-4分制）：评分：舒适度：X，安全性：Y，总体评价：Z

A.3.4 GPT-4o零样本基线

你是城市骑行基础设施评估方面的专家。请分析所提供的街景图像，并从特定骑行者画像的角度评估其骑行适宜性。

骑行者画像：{persona} 角色描述：{persona_desc} {osm_text}

任务：进行骑行适宜性评估：1. 分析街道环境和 OSM 属性 2. 考虑一位「{persona}」骑行者会如何看待这一环境 3. 对以下项目进行评分（1-4）：舒适度、安全性、总体评价 4. 输出包含影响因素和评分的 JSON

A.3.5 角色描述

提示中使用的四种骑行者画像：
强势无畏型：能够适应所有基础设施类型，对受保护设施和无保护设施几乎没有偏好差异。

热情而自信：经常骑行的骑行者，他们更喜欢自行车道，但必要时也会在混合交通中骑行。

感兴趣但担忧型：如果与交通分离，就会更多骑行；需要受保护的基础设施才会感到安全。

No 绝不骑行型：认为骑行太危险的非骑行者无论基础设施如何都如此。