From Image Generation to Infrastructure Design: a Multi-agent Pipeline for Street Design Generation

Chenguang Wang[1,3], Xiang Yan[2], Yilong Dai[2], Ziyi Wang[4], Susu Xu[1]
[1]Johns Hopkins University [2]University of Florida
[3]Stony Brook University [4]University of Maryland
chenguang.wang@stonybrook.edu susuxu@jhu.edu

Abstract

Realistic visual renderings of street-design scenarios are essential for public engagement in active transportation planning. Traditional approaches are labor-intensive, hindering collective deliberation and collaborative decision-making. While AI-assisted generative design shows transformative potential by enabling rapid creation of design scenarios, existing generative approaches typically require large amounts of domain-specific training data and struggle to enable precise spatial variations of design/configuration in complex street-view scenes. We introduce a multi-agent system that edits and redesigns bicycle facilities directly on real-world street-view imagery. The framework integrates lane localization, prompt optimization, design generation, and automated evaluation to synthesize realistic, contextually appropriate designs. Experiments across diverse urban scenarios demonstrate that the system can adapt to varying road geometries and environmental conditions, consistently yielding visually coherent and instruction-compliant results. This work establishes a foundation for applying multi-agent pipelines to transportation infrastructure planning and facility design.

1 Introduction

Cycling is an environmentally friendly mode of transportation that also offers co-benefits such as promoting personal health and reducing traffic congestion [11, 10, 5, 36]. However, bicycle infrastructure development often requires extensive stakeholder consultations, during which road users (e.g., cyclists, drivers, and pedestrians) articulate their needs and concerns. To facilitate these deliberations, visual renderings of proposed street design scenarios are widely used in practice as tools for collective reflection and collaborative decision-making. Traditionally, these visuals are created with graphic design software (e.g., Adobe Photoshop and SketchUp) [1, 4, 48] prior to an extensive user survey. While effective, these tools are time-consuming and demand specialized expertise, making it difficult to customize and dynamically adjust street design images in response to various user feedback, thereby hindering agile scenario iteration and limiting their utility in dynamic public engagement contexts that involve complex trade-offs in allocating road space [3, 37, 17, 27].

Recent advances in Generative AI (GenAI), particularly image-generation models, demonstrate significant potential to support scenario ideation and facilitate collaborative decision-making across domains such as industrial design, architecture, and site planning. [41, 9, 55, 12, 17, 27, 26]. Existing GenAI-based scenario design has leveraged post-training methods on domain-specific imagery [47, 38, 8, 15, 29, 30], which necessitates large, curated datasets and significant computational resources. More recently, the state-of-the-art models such as GPT-image-1 [42, 44] have made it feasible to apply off-the-shelf systems directly to design scenarios without task-specific retraining. These models provide strong text-to-image and image-to-image capabilities, enabling the rapid creation of immersive, design-oriented bicycle-infrastructure visualizations from urban imagery, particularly

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: UrbanAI: Harnessing Artificial Intelligence for Smart Cities.

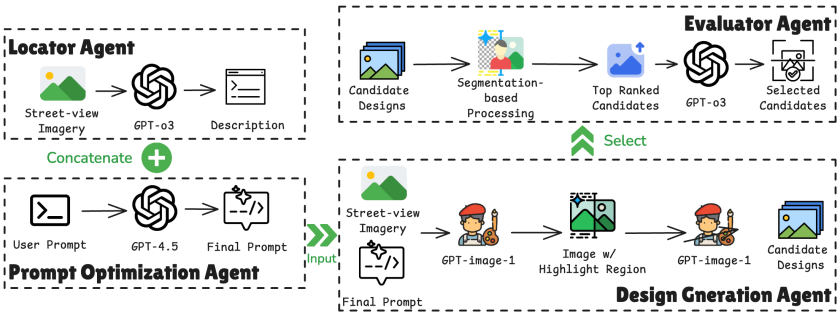

Figure 1: **Overview of our multi-agent system for bicycle-infrastructure design.** The system comprises a Locator, Prompt Optimization, Design Generation, and Evaluation agent that processes street-view imagery to generate bike-lane designs. Green arrows and plus signs denote intermediate operations on agents' output within the workflow.

street-view imagery. However, research on customizing these models for street infrastructure design, particularly at the site level, remains largely underexplored. Key limitations include: (i) inadequate reasoning about spatial and relational structure within visual inputs; (ii) semantic misinterpretation of user instructions; (iii) weak adherence to complex instructions that with multiple constraints specified; and (iv) inconsistent outputs and occasional hallucinations [44, 53, 49]. These limitations underscore that stand-alone image generation is insufficient for bicycle-infrastructure scenario design, pointing to the need for a more structured framework that situates state-of-the-art models within a more comprehensive workflow.

To leverage cutting-edge GenAI models for bicycle infrastructure scenario design while addressing the existing limitations, we propose a multi-agent system built on a state-of-the-art image generation backbone, GPT-image-1 [42]. Given a user-defined prompt and street-view imagery, the generative pipeline produces realistic bicycle facility design scenarios via four specialized agents that can tackle each of the four limitations discussed above: (1) a *Locator Agent* that generates contextually accurate descriptions of bike-lane positions using Multimodal Large Language Models (MLLMs), helping image generation model capture spatial relations; (2) a *Prompt Optimization Agent* that refines user prompts by integrating illustrative references along with the Locator's contextual descriptions, thereby reducing semantic misinterpretation. (3) a *Design Generation Agent* that decouples geometric and design-pattern constraints via a cascading generation, yielding multiple candidate scenario designs. (4) an *Evaluation Agent* that reranks candidate designs via CLIP similarity to a reference layout and conducts a binary compliance check with reasoning MLLMs, surfacing the most instruction-aligned outputs. Experimental results on street-view imagery collected from diverse road contexts show that our pipeline consistently generates realistic, instruction-aligned, and spatially coherent designs. This enables rapid creation of street-design scenarios and supports collective reflection and collaborative decision-making in bicycle infrastructure planning and design.

The contributions of this work are threefold:

- We extend the applicability of generative AI in urban planning by integrating state-of-the-art image-generation models for bicycle infrastructure design.
- We develop a multi-agent system that generates street infrastructure configurations with high spatial accuracy and contextual relevance, while ensuring compliance with planning guidelines.
- We design a pipeline that streamlines the design workflow, reducing complexity, expertise requirements, and time cost in scenario generation.

2 Method

This section presents our proposed multi-agent system for bicycle infrastructure design. We begin by describing the three agents for image editing and design generation, followed by the evaluation agent responsible for selecting the best generated designs. The multi-agent pipeline can use any street view image as input for the generation of street design scenarios. To demonstrate the scalability and

widespread applicability of the proposed approach, here we use Google Street View imagery obtained via the Google Street View API [1].

2.1 Image Editing Agents

Locator Agent Preliminary experiments and prior studies indicate that current image generation models often lack the ability to accurately interpret and render the spatial configuration of street-scene elements in street-view imagery [44]. As a result, directly prompting these models often fails to reliably locate existing bicycle infrastructure or depict it in accordance with the given instructions. To address this limitation, we design a *Locator Agent* powered by GPT-o3, a state-of-the-art reasoning MLLM [43]. The agent analyzes street-view images and generates detailed, structured descriptions of bike-lane features, including lane markings/paint, patterns, widths, and relative positions to reference objects (e.g., "adjacent to the sidewalk"). If no bike lane is present, the image is excluded from downstream processing. These precise descriptions serve as contextual anchors for subsequent generation steps, improving spatial accuracy in the synthesized images and reducing ambiguity in the location description of user-defined instructions.

Prompt Optimization Agent Prior work has demonstrated that the quality of synthesized images depends strongly on prompt formulation, and that in-context learning for prompt generation can substantially improve results [19, 56]. Motivated by these findings, we design the *Prompt Optimization Agent*. We begin by manually drafting a series of candidate prompts to identify the description format that best enables the model to understand the task and produce satisfactory results. Our Exploratory experiments indicate that a structured template, comprising an overall lane description, detailed left/right boundary specifications, and explicit constraints, yields the most reliable outputs. Based on this template, we prepare several high-quality examples to serve as in-context references, integrate user-specific instructions, and prompt a GPT-4.5 model to automatically generate the final image-generation prompt (An illustrative example is in Appendix F). The generated prompt is then concatenated with the structured bike-lane descriptions from the *Locator Agent* or other further requirements, ensuring that the image generation model receives a clear, precise, and contextually rich instruction set for modifying bicycle infrastructure.

Design Generation Agent Building on prior findings that image generation models often struggle to faithfully render all specified elements in complex, compositional prompts [13, 23, 22, 7], we seek to enhance the robustness of the generation process. To this end, the *Design Generation Agent* adopts a two-step cascading strategy. In the first step, the model is prompted solely to edit existing bike lanes or add new ones as clearly highlighted regions in the street-view image. In the second step, we apply our final optimized prompt, augmented with an explicit statement that the highlighted regions represent bike lanes, to produce diverse modifications such as standard marked lanes, buffered lanes, and colored-surface lanes for improved visibility and safety. For each scenario, the agent generates 5 to 10 candidates, forming a diverse pool for the *Evaluator Agent* to assess and select the most suitable synthesized designs.

2.2 Evaluator Agent

Recent work has reported that redundant environmental noise within images can undermine the effectiveness of CLIP embeddings [45] for similarity assessment. Such noise, originating from elements like vehicles, pedestrians, and buildings, can interfere with embedding computations, thereby reducing the accuracy of embedding-based evaluations [28]. To address this limitation, we design a segmentation-based preprocessing stage to remove environmental noise from the synthesized designs. Specifically, we manually annotated bicycle lane regions in a representative subset of synthesized designs (including unsatisfactory generations) to construct a training set. We then fine-tuned a YOLO-v11 segmentation model [25] to achieve precise bicycle infrastructure segmentation under diverse urban conditions. The resulting model generates binary masks that isolate the bicycle infrastructure regions, with all non-relevant areas masked out using a uniform color, effectively eliminating extraneous visual content.

Following segmentation, CLIP embeddings were computed on the isolated bicycle infrastructure regions, and the cosine similarity between each candidate and a designated reference design was

[1] `https://developers.google.com/streetview`

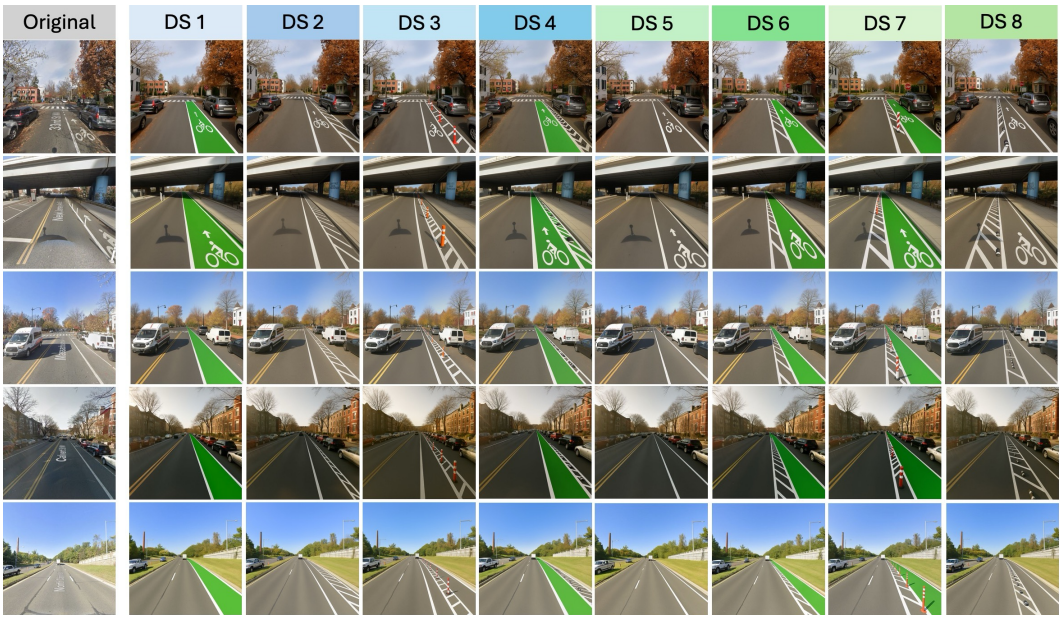

Figure 2: **Generated bicycle-lane designs across diverse urban contexts.** Each row shows the original street-view scene (left) and eight variations generated by our multi-agent pipeline, one per predefined design scenario (DS1–DS8; DS = design scenario).

calculated for re-ranking. Only the top three most similar candidates were advanced to the second-stage evaluation. In this stage, each candidate image, along with its corresponding prompt, was passed to GPT-o3 [43], which was tasked with determining whether the design complied with the specified requirements stated in the final optimized prompt. GPT-o3 produced binary suitability judgments for each candidate. This two-stage evaluation process ensured that the final selected design exhibited full adherence to the specified design criteria.

3 Main Result

Based on common configurations of bicycle facilities in real-world urban environments, we define eight representative bikeway design scenarios. These scenarios encompass a range of widely used treatments, including standard marked lanes, buffered lanes providing separation from vehicle traffic, and colored-surface lanes to improve visual saliency and safety awareness. The detailed definitions and corresponding visual examples for each design scenario are provided in Appendix B.

Figure 2 presents the AI-enabled street design scenarios generated by our proposed multi-agent pipeline. The generated designs span a variety of urban scenarios, including dense city streets, suburban roads, highways, and complex intersections. These qualitative results demonstrate that our pipeline can consistently embed different lane patterns into diverse street-view contexts while maintaining correct spatial alignment with the roadway and preserving overall scene realism. Even in challenging conditions, such as complex backgrounds or partial occlusions, the generated lanes remain visually distinct and contextually appropriate. Notably, several scenarios (including a highway context) were deliberately designed as counterfactual cases to challenge the pipeline's ability to generate bicycle infrastructure in less conventional or unfavorable settings. Despite these challenges, our approach is able to produce satisfactory and contextually coherent results. Furthermore, Appendix B reports results from cross-model comparisons and ablation studies, which justify our choice of the backbone model and demonstrate the individual contributions of each module of the multi-agent system to generation accuracy

Desgin Scenario	1	2	3	4	5	6	7	8
Eval Acc. (%)	95.5	96.5	97.0	95.5	96.0	95.5	97.0	96.5

Table 1: **Accuracy of the Evaluator Agent in correctly selecting the candidate.**

In addition, to assess the accuracy of the Evaluator Agent in determining whether a candidate design adheres to the final optimized prompt, we conducted a human evaluation on a street-view test set that was never used in model training. The results, presented in Table 1, show that the Evaluator Agent

can consistently and accurately select the most suitable candidate across all eight design scenarios, achieving accuracies above 95%. These findings further demonstrate the robustness and effectiveness of our method.

4 Conclusion

In this work, we present a multi-agent framework for bicycle infrastructure design that integrates advanced image generation models with reasoning MLLMs. The system decomposes the design process into several steps through specialized agents, enabling context-aware and spatially accurate modifications to street-view imagery. Both qualitative and quantitative evaluation results demonstrate that our approach can robustly produce accurate, contextually appropriate, and visually realistic bicycle infrastructure designs. By integrating advanced AI systems into the street design workflow, our method offers a promising tool to support bicycle infrastructure planning and facilitate design.

Acknowledgement

We are grateful for the support from the District Department of Transportation throughout this project. This project was also supported by the U.S. National Science Foundation (Award Nos. 2425029 and 2425030). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the District Department of Transportation or the National Science Foundation.

References

[1] Kheir Al-Kodmany. Using visualization techniques for enhancing public participation in planning and design: process, implementation, and evaluation. *Landscape and urban planning*, 45(1):37–45, 1999.

[2] Ahmad Arrabi, Xiaohan Zhang, Waqas Sultani, Chen Chen, and Safwan Wshah. Cross-view meets diffusion: Aerial image synthesis with geometry and text guidance. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5356–5366. IEEE, 2025.

[3] Karen Bickerstaff, Rodney Tolley, and Gordon Walker. Transport planning and participation: the rhetoric and realities of public involvement. *Journal of Transport Geography*, 10(1):61–73, 2002.

[4] Peter Bosselmann, Elizabeth Macdonald, and Thomas Kronemeyer. Livable streets revisited. *Journal of the American Planning Association*, 65(2):168–180, 1999.

[5] Christian Brand, Thomas Götschi, Evi Dons, Regine Gerike, Esther Anaya-Boig, Ione Avila-Palencia, Audrey De Nazelle, Mireia Gascon, Mailin Gaupp-Berghausen, Francesco Iacorossi, et al. The climate change mitigation impacts of active travel: Evidence from a longitudinal panel study in seven european cities. *Global environmental change*, 67:102224, 2021.

[6] Yuri Calleo, Nadia Giuffrida, and Francesco Pilla. Exploring hybrid models for identifying locations for active mobility pathways using real-time spatial delphi and gans. *European Transport Research Review*, 16(1):61, 2024.

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

[8] Boyang Deng, Richard Tucker, Zhengqi Li, Leonidas Guibas, Noah Snavely, and Gordon Wetzstein. Streetscapes: Large-scale consistent street view generation using autoregressive video diffusion. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024.

[9] Yu-Min Fang. The role of generative ai in industrial design: enhancing the design process and education. In *IET Conference Proceedings CP868*, volume 2023, pages 135–136. IET, 2023.

[10] Elliot Fishman. Cycling as transport, 2016.

[11] Elliot Fishman, Paul Schepers, and Carlijn Barbara Maria Kamphuis. Dutch cycling: quantifying the health and related economic benefits. *American journal of public health*, 105(8):e13–e15, 2015.

[12] Yanjie Fu. Towards ai urban planner in the age of genai, llms, and agentic ai. *arXiv preprint arXiv:2507.14730*, 2025.

[13] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023.

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

[15] Songen Gu, Jinming Su, Yiting Duan, Xingyue Chen, Junfeng Luo, and Hao Zhao. Text2street: Controllable text-to-image generation for street views. In *International Conference on Pattern Recognition*, pages 130–145. Springer, 2025.

[16] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14953–14962, 2023.

[17] Liam Hancock and John Parkin. The challenges of applying cycling design guidance. In *Proceedings of the Institution of Civil Engineers-Transport*, volume 177, pages 494–503. Emerald Publishing, 2024.

[18] Tiankai Hang, Shuyang Gu, Dong Chen, Xin Geng, and Baining Guo. Cca: collaborative competitive agents for image editing. *Frontiers of Computer Science*, 19(11):1–17, 2025.

[19] Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:66923–66939, 2023.

[20] Mingyi He, Yuebing Liang, Shenhao Wang, Yunhan Zheng, Qingyi Wang, Dingyi Zhuang, Li Tian, and Jinhua Zhao. Generative ai for urban design: A stepwise approach integrating human expertise with multimodal diffusion models. *arXiv preprint arXiv:2505.24260*, 2025.

[21] Chuanbo Hu, Shan Jia, and Xin Li. Ursimulator: Human-perception-driven prompt tuning for enhanced virtual urban renewal via diffusion models. *ISPRS Journal of Photogrammetry and Remote Sensing*, 228:356–369, 2025.

[22] Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[23] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

[24] Koichi Ito, Matias Quintana, Xianjing Han, Roger Zimmermann, and Filip Biljecki. Translating street view imagery to correct perspectives to enhance bikeability and walkability studies. *International Journal of Geographical Information Science*, 38(12):2514–2544, 2024.

[25] Glenn Jocher and Jing Qiu. Ultralytics yolo11, 2024.

[26] Timo Kapsalis. Urbangenai: reconstructing urban landscapes using panoptic segmentation and diffusion models. *arXiv preprint arXiv:2401.14379*, 2024.

[27] Emma R Lawlor, Kate Ellis, Jean Adams, Russell Jago, Louise Foley, Stephanie Morris, Tessa Pollard, Carolyn Summerbell, Steven Cummins, Hannah Forde, et al. Stakeholders' experiences of what works in planning and implementing environmental interventions to promote active travel: a systematic review and qualitative synthesis. *Transport reviews*, 43(3):478–501, 2023.

[28] Chuanhao Li, Chenchen Jing, Zhen Li, Mingliang Zhai, Yuwei Wu, and Yunde Jia. In-context compositional generalization for large vision-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17954–17966, 2024.

[29] Ke Li, Chenyu Zhang, Yuxin Ding, Xianbiao Hu, and Ruwen Qin. Acquiring and accumulating knowledge from diverse datasets for multi-label driving scene classification. *arXiv preprint arXiv:2506.17101*, 2025.

[30] Ke Li, Chenyu Zhang, Yuxin Ding, Xianbiao Hu, and Ruwen Qin. Multi-label scene classification for autonomous vehicles: Acquiring and accumulating knowledge from diverse datasets. *arXiv e-prints*, pages arXiv–2506, 2025.

[31] Ming Li, Pei Chen, Chenguang Wang, Hongyu Zhao, Yijun Liang, Yupeng Hou, Fuxiao Liu, and Tianyi Zhou. Mosaic-it: Cost-free compositional data synthesis for instruction tuning. *arXiv preprint arXiv:2405.13326*, 2024.

[32] Ming Li, Chenguang Wang, Yijun Liang, Xiyao Wang, Yuhang Zhou, Xiyang Wu, Yuqing Zhang, Ruiyi Zhang, and Tianyi Zhou. Caughtcheating: Is your mllm a good cheating detective? exploring the boundary of visual perception and reasoning. *arXiv preprint arXiv:2507.00045*, 2025.

[33] Ming Li, Ruiyi Zhang, Jian Chen, Jiuxiang Gu, Yufan Zhou, Franck Dernoncourt, Wanrong Zhu, Tianyi Zhou, and Tong Sun. Towards visual text grounding of multimodal large language model. *arXiv preprint arXiv:2504.04974*, 2025.

[34] Mingcheng Li, Xiaolu Hou, Ziyang Liu, Dingkang Yang, Ziyun Qian, Jiawei Chen, Jinjie Wei, Yue Jiang, Qingyao Xu, and Lihua Zhang. Mccd: Multi-agent collaboration-based compositional diffusion for complex text-to-image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13263–13272, 2025.

[35] Yijun Liang, Ming Li, Chenrui Fan, Ziyue Li, Dang Nguyen, Kwesi Cobbina, Shweta Bhardwaj, Jiuhai Chen, Fuxiao Liu, and Tianyi Zhou. Colorbench: Can vlms see and understand the colorful world? a comprehensive benchmark for color perception, reasoning, and robustness. *arXiv preprint arXiv:2504.10514*, 2025.

[36] Sheng Liu, Auyon Siddiq, and Jingwei Zhang. Planning bike lanes with data: Ridership, congestion, and path selection. *Management Science*, 2024.

[37] Sofia Löfgren. Designing with differences, cross-disciplinary collaboration in transport infrastructure planning and design. *Transportation research interdisciplinary perspectives*, 4:100106, 2020.

[38] Xiaohu Lu, Zuoyue Li, Zhaopeng Cui, Martin R Oswald, Marc Pollefeys, and Rongjun Qin. Geometry-aware satellite-to-ground image synthesis for urban areas. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 859–867, 2020.

[39] Qi Mao, Haobo Hu, Yujie He, Difei Gao, Haokun Chen, and Libiao Jin. Emoagent: Multi-agent collaboration of plan, edit, and critic, for affective image manipulation. *arXiv preprint arXiv:2503.11290*, 2025.

[40] National Association of City Transportation Officials. *Urban bikeway design guide*. Island Press, 2025.

[41] Damilola Onatayo, Adetayo Onososen, Abiola Oluwasogo Oyediran, Hafiz Oyediran, Victor Arowoiya, and Eniola Onatayo. Generative ai applications in architecture, engineering, and construction: trends, implications for practice, education & imperatives for upskilling—a review. *Architecture*, 4(4):877–902, 2024.

[42] OpenAI. Introducing 4o image generation. `https://openai.com/index/introducing-4o-image-generation/`, 2025.

[43] OpenAI. OpenAI o3 and o4-mini System Card. `https://openai.com/index/o3-o4-mini-system-card/`, April 2025.

[44] Yusu Qian, Jiasen Lu, Tsu-Jui Fu, Xinze Wang, Chen Chen, Yinfei Yang, Wenze Hu, and Zhe Gan. Gie-bench: Towards grounded evaluation for text-guided image editing. *arXiv preprint arXiv:2505.11493*, 2025.

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[46] Balaji Ganesh Rajagopal, Manish Kumar, Abdulaziz H Alshehri, Fayez Alanazi, Ahmed Farouk Deifalla, Ahmed M Yosri, and Abdelhalim Azam. A hybrid cycle gan-based lightweight road perception pipeline for road dataset generation for urban mobility. *Plos one*, 18(11):e0293978, 2023.

[47] Krishna Regmi and Ali Borji. Cross-view image synthesis using geometry-guided conditional gans. *Computer Vision and Image Understanding*, 187:102788, 2019.

[48] Stephen RJ Sheppard. Guidance for crystal ball gazers: developing a code of ethics for landscape visualization. *Landscape and urban planning*, 54(1-4):183–199, 2001.

[49] Takahiro Shirakawa and Seiichi Uchida. Noisecollage: A layout-aware text-to-image diffusion model based on noise cropping and merging. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8921–8930, 2024.

[50] Stability AI. Introducing *Stable Diffusion 3.5*. urlhttps://stability.ai/news/introducing-stable-diffusion-3-5, October 2025. Updated October 29th with release of Stable Diffusion 3.5 Medium.

[51] Jiayang Sun, Hongbo Wang, Jie Cao, Huaibo Huang, and Ran He. Marmot: Multi-agent reasoning for multi-object self-correcting in improving image-text alignment. *arXiv preprint arXiv:2504.20054*, 2025.

[52] Xie Tianyidan, Rui Ma, Qian Wang, Xiaoqian Ye, Feixuan Liu, Ying Tai, Zhenyu Zhang, Lanjun Wang, and Zili Yi. Anywhere: A multi-agent framework for user-guided, reliable, and diverse foreground-conditioned image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7410–7418, 2025.

[53] Sujith Vemishetty, Advitiya Arora, and Anupama Sharma. Towards evaluating robustness of prompt adherence in text to image models. *arXiv preprint arXiv:2507.08039*, 2025.

[54] Kavana Venkatesh, Connor Dunlop, and Pinar Yanardag. Crea: A collaborative multi-agent framework for creative content generation with diffusion models. *arXiv preprint arXiv:2504.05306*, 2025.

[55] Qingyi Wang, Yuebing Liang, Yunhan Zheng, Kaiyuan Xu, Jinhua Zhao, and Shenhao Wang. Generative ai for urban planning: Synthesizing satellite imagery via diffusion models. *arXiv preprint arXiv:2505.08833*, 2025.

[56] Zhendong Wang, Yifan Jiang, Yadong Lu, Pengcheng He, Weizhu Chen, Zhangyang Wang, Mingyuan Zhou, et al. In-context learning unlocked for diffusion models. *Advances in Neural Information Processing Systems*, 36:8542–8562, 2023.

[57] Jasper S Wijnands, Kerry A Nice, Jason Thompson, Haifeng Zhao, and Mark Stevenson. Streetscape augmentation using generative adversarial networks: Insights related to health and wellbeing. *Sustainable Cities and Society*, 49:101602, 2019.

[58] Junge Zhang, Qihang Zhang, Li Zhang, Ramana Rao Kompella, Gaowen Liu, and Bolei Zhou. Urban scene diffusion through semantic occupancy map. *arXiv preprint arXiv:2403.11697*, 2024.

[59] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023.

A Related Work

A.1 AI-assisted Generative Design in Urban Planning

Established methods have sophisticatedly integrated the image generation model into the design of urban planning. Wijnands et al. [57] used unpaired GAN-based image translation on large Google Street View datasets stratified by population health, augmenting streetscapes to reveal actionable design cues—more green space, wider footpaths/sidewalks, fewer fences, and greater frontage compactness—informing healthier street and sidewalk design. Ito et al. [24] quantifies the bias in car-mounted street-view imagery and introduces a GAN-based translation framework to convert road-center views into cyclist/sidewalk perspectives, aligning semantic indicators so bikeability/walkability assessments better inform bikeway and streetscape design. Rajagopal et al. [46] similarly use CycleGAN to convert simple annotated road layouts into lifelike street-level images, enabling rapid prototyping of bike lane designs. Calleo et al. [6] employ a Real-Time Spatial Delphi method combined with GAN-based image synthesis to produce photorealistic street redesign visuals, significantly aiding stakeholder communication.

Arrabi et al. [2] propose a two-stage, geometry-preserving ground-to-aerial synthesis pipeline (BEV layout prediction from a street photo, then text-conditioned diffusion) to generate realistic overhead imagery that can aid transportation/streetscape planning. Zhang et al. [58] introduce a 3D diffusion approach conditioned on BEV layouts to generate large, unbounded urban scenes as semantic occupancy maps (and renderable images), enabling rapid what-if exploration of road-network design options. Hu et al. [21] develop a human-perception-guided prompt-tuning framework that locally edits street-view images with Stable Diffusion to simulate urban-renewal interventions and quantitatively boost perceived safety/beauty/liveliness—useful for previewing streetscape or bike-corridor upgrades.

Collectively, these studies demonstrate the promise of generative AI for urban and streetscape visualization, but they are ill-suited for bicycle-infrastructure design. In particular, they still face notable limitations, such as insufficient spatial immersion and fine-grained environmental cues to convey the on-road cyclist experience, or high computational costs and labor-intensive data preparation for generative models. In addition, the overall design workflow remains cumbersome, requiring extensive manual adjustments and cross-disciplinary coordination among human experts.

A.2 Multi-agent system for Image Editing

Recent developments favor multi-agent systems to enhance complex image editing capabilities. Gupta et al. [16] introduce VisProg, which demonstrated how an LLM-based single-agent planner could decompose intricate editing tasks, highlighting the benefits of agent-driven task segmentation but constrained by reliance on a single controller. Hang et al. [18] proposed Collaborative Competitive Agents (CCA), employing two generators and a discriminator in an iterative feedback loop, where generators compete yet collaboratively improve via shared feedback, significantly enhancing robustness for complex multi-step edits. Venkatesh et al. introduce CREA [54], which employed distinct role-based agents (e.g., Creative Director, Art Critic) to iteratively refine images creatively, significantly outperforming single-prompt diffusion models by achieving greater output diversity and semantic alignment through a collaborative, human-like creative process. Xie et al. [52] adopted a modular multi-agent approach for foreground-aware editing tasks, assigning dedicated agents for foreground semantics, object integrity, and background consistency. This system notably enhanced image quality and control compared to end-to-end models. EmoAgent [39] tackled affective image manipulation by emulating cognitive painting workflows through planning, execution, and critique agents, significantly enhancing emotional expression and editing interpretability. Marmot [51] employed specialized agents post-generation to correct object count, attributes, and spatial arrangement, substantially improving alignment with textual descriptions. Multi-Agent Collaboration-based Compositional Diffusion (MCCD) [34] similarly used multimodal LLMs to parse prompts into object-specific agents, integrating regional outputs through hierarchical diffusion to achieve precise control and improved compositional consistency.

Collectively, these works establish multi-agent image editing as a mature and effective paradigm: role-specialized agents, LLM planners, and collaborative/competitive feedback reliably decompose complex edits, improve semantic alignment, and enhance robustness over single-prompt baselines. Building on this consensus, we tailor a coordinated agentic architecture to bicycle infrastructure scenario design, aligning agents with domain needs, so that modified scenario design to street-view imagery can be produced consistently.

B Street View Image Collection

To ensure that our proposed approach is widely applicable across diverse road environments, we sampled 150 road segments and intersections by considering several main factors (e.g., type of bike facility, speed limit, and neighborhood environment). The sampled locations include both locations that already contain visible bicycle infrastructure and locations without it, but with road layouts suitable for adding new bike lanes. This selection covers a diverse range of urban contexts, including residential neighborhoods, suburban corridors, central business districts, highways, and intersections.

Based on the coordinates of the sampled locations, we use the Google Street View API [2] to retrieve static street-view images from multiple horizontal viewing angles, while keeping both pitch and field of view fixed to capture an eye-level perspective. This ensures coverage of different forward-looking viewpoints even when the street orientation and traffic direction are uncertain. All images are obtained at a resolution of 1024×1024 pixels to align with the input requirements of the downstream image generation backbone. We further perform a quality control stage in which human experts review all captured views for each location and select the most suitable image. The chosen images are those that provide a clear view of the carriageway, sidewalk, or curb reference lines, and any existing bicycle lane markings when present. This curated set of high-quality street-view images forms the foundation for all subsequent stages of our pipeline, ensuring consistent and contextually rich inputs for the Locator, Prompt Optimization, and Design Generation agents.

C Model Evaluation and Validation

In this section, we justify our selection of the backbone model and assess the effectiveness of each component in the multi-agent generation pipeline through corresponding qualitative evaluations.

C.1 Generation Backbone Comparison

In this section, we compare our image generation backbone, GPT-image-1, with the state-of-the-art open-source image generation model, Stable Diffusion 3.5 [50], demonstrating the rationale of backbone selection. We focus on three design scenarios, Design Scenario 1, Design Scenario 6, and Design Scenario 7 (Please see Table 2), because these designs are among the most frequently implemented in contemporary urban cycling infrastructure and are explicitly emphasized in widely used design guidelines such as the NACTO Urban Bikeway Design Guide [40]. To ensure that observed performance differences are attributable solely to the generative models rather than content variability, we standardize the experimental conditions: for each design scenario, all models receive the exact same input street-view photograph and identical textual prompt. The only variable is the generative backbone, GPT-image-1 versus the three latest Stable Diffusion models, Stable Diffusion 3.5 variants (Stable Diffusion 3.5 Large, Stable Diffusion 3.5 Large Turbo, and Stable Diffusion 3.5 Medium), allowing us to directly attribute output differences to model-specific handling of spatial and semantic constraints.

Figures 3,4, and 5 present qualitative comparisons under identical input photographs and textual prompts, covering the three most commonly implemented bikeway typologies as emphasized in widely used design guidelines [40]. Furthermore, to ensure that observed performance differences are attributable solely to the generative models rather than content variability, we standardize the experimental conditions: for each design scenario, all models receive the exact same input street-view photograph and identical textual prompt. From the qualitative comparison, we find that GPT-image-1 consistently adheres to the prompt with high fidelity, preserving all background elements while making precise, localized edits to the bicycle infrastructure. In Figure 3, the model modifies the current bike lane into a green-painted bike lane that follows road geometry and perspective, with bicycle glyphs correctly scaled, oriented, and positioned. In Figure 4, it generates a buffered lane of appropriate width and perspective without altering adjacent lane markings or roadside objects. In Figure 5, it accurately places a protected lane with bollards while maintaining occlusion relationships with visual elements. These results demonstrate strong semantic alignment between the textual description and visual output, with minor changes to unrelated scene content.

[2] `https://developers.google.com/streetview`

Figure 3: **Qualitative comparison of Desgin Scenario 1 outputs** generated by GPT-image-1 and three Stable Diffusion 3.5 variants (Stable Diffusion 3.5 Large (SD3.5-L), Stable Diffusion 3.5 Large Turbo(SD3.5-L-T), and Stable Diffusion 3.5 Medium(SD3.5-M).

In contrast, the Stable Diffusion 3.5 variants (Large, Large Turbo, and Medium; SD3.5-L/SD3.5-L-T/SD3.5-M) exhibit recurrent failure modes despite visually realistic textures. The core limitation is twofold: these models neither reliably internalize where bicycle infrastructure must be placed relative to road elements, nor do they strictly follow our prompt to modify highly specialized, position-dependent details. Concretely, we observe (i) *topological errors*: lanes crossing the centerline, appearing on the parking side, or breaking at junctions; (ii) *metric/perspectival inconsistencies*: lanes rendered with an approximately constant image-space width rather than shrinking with depth, and buffers whose width varies along the road direction; (iii) *symbolography mistakes*: bicycle glyphs and arrows with incorrect orientation, spacing, or lane offset; (iv) *occlusion/layering violations*: painted lanes rendered over vehicles or barriers instead of behind them; and (v) *unintended scene rewriting*: hallucinated buildings or altered road geometry, i.e., background replacement (cf. Figs.3, 4, and5). These design scenarios suggest weak inductive bias for structured layout and scene topology: the models optimize global realism but lack mechanisms to enforce local, prompt-governed geometric constraints and strict background preservation demanded by infrastructure editing. Mitigation via Stable-diffusion-based fine-tuning is pragmatic but typically requires large, curated, domain-specific

Figure 4: **Qualitative comparison of Desgin Scenario 6 outputs** generated by GPT-image-1 and three Stable Diffusion 3.5 variants (Stable Diffusion 3.5 Large (SD3.5-L), Stable Diffusion 3.5 Large Turbo(SD3.5-L-T), and Stable Diffusion 3.5 Medium(SD3.5-M).

datasets and task-specialized conditioning signals, which are resource-intensive and, for this narrowly scoped editing task, have conflict over our research motivation [59]. Accordingly, we center our evaluation on GPT-image-1 and do not extend to additional Stable-diffusion-based variants.

C.2 Ablation Study

In this section, we perform an ablation study by applying our pipeline to selected examples while deliberately removing specific agents to assess the individual contribution and effectiveness of each component.

Locator Agent Removal Figure 6 compares generation results with and without the Locator Agent. When the Locator Agent is removed, the system loses the ability to correctly identify the spatial position of the bicycle lane. As shown in the right column, the model frequently misinterprets traffic lanes as bike lanes, leading to misplaced green surfacing or lane markings in the center of the road rather than adjacent to the curb. In several cases, the generation process also introduces unintended

Figure 5: **Qualitative comparison of Desgin Scenario 7 outputs** generated by GPT-image-1 and three Stable Diffusion 3.5 variants (Stable Diffusion 3.5 Large (SD3.5-L), Stable Diffusion 3.5 Large Turbo(SD3.5-L-T), and Stable Diffusion 3.5 Medium(SD3.5-M).

alterations to non-bike-lane areas, such as partially removing or replacing parking zones, which changes road elements unrelated to the intended design. In contrast, the full pipeline (left column) produces bike lanes that are correctly aligned with the curb and preserve other roadway features, demonstrating the crucial role of accurate localization in ensuring spatially correct and contextually consistent designs.

Prompt Optimization Removal Figure 7 shows that removing the Prompt Optimization Agent causes the system to rely on raw user prompts, which are often ambiguous or too terse. Without structured disambiguation, the generator hallucinates details and fails to realize the intended design scenario: in the first case, the simplest "two parallel lines" lane is mis-rendered with double lines as its boundary, and in another design, color spill appears outside the lane region (e.g., green paint bleeding into the travel lane or shoulder). We also observe missing or incorrect elements: buffers/hatching omitted or placed on the wrong side, symbols misaligned with the roadway direction, and inconsistent continuity along the lane. These errors stem from the model having to infer side, width, buffer type, and coloring from under-specified text, which typically demands substantial manual prompt tweaking

Figure 6: **Effect of removing the Locator Agent.** Each row shows an original street-view image (left), the generation result with the full pipeline (middle), and the result after removing the Locator Agent (right).

and trial-and-error. The Prompt Optimization Agent mitigates these failures by converting user intent into a structured, context-aware prompt (with explicit side, width/buffer specification, coloring rules, and "edit-only within highlighted region" constraints) and by injecting in-context exemplars. This reduces ambiguity, stabilizes design scenario realization, and keeps the output aligned with both the guideline and the user's intention.

Image Generation Auxiliary Step Removal Figure 8 contrasts the full pipeline (middle) with a variant that omits the first step that generates a highlight region first, where we normally convert the target lane area into a highlighted pre-edit region and then synthesize the final bike lane from that highlight. Without this visual scaffold, the generator must infer both localization and styling from text alone. As shown in the right column, this leads to systematic loss of width control (lanes become too wide/narrow or taper inconsistently), over-stretching along the curb that spills into parking or shoulders, and perspective-inconsistent strips that look unrealistic. We also observe design scenario non-compliance, such as diagonal hatching or buffers being omitted or flipped. The highlight-first step acts as an explicit spatial prior, a geometry-aware, human-interpretable cue (not a hard mask) that fixes the lane's extent and approximate width before style details are rendered. Removing it forces the model to resolve location, width, and elements directly from ambiguous prompts, amplifying hallucinations and causing the final design to deviate from the guidelines.

Necessity of Evaluator Agent Figure 9 shows five independent generations from the same street-view image and prompt. Because the generator is stochastic and lacks hard spatial constraints, first-pass results are often unreliable and vary substantially across attempts: lane width drifts, curb alignment shifts, buffers/hatching appear or disappear, bollard count and spacing fluctuate, and green surfacing occasionally bleeds beyond the intended region. Some attempts violate the prescribed

Figure 7: **Effect of removing the Prompt Optimization Agent.** Each row shows an original street-view image (left), the generation result with the full pipeline (middle), and the result after removing the Prompt Optimization Agent (right).

design scenario or roadway context. This sample-level variance is an inherent limitation of current image generators, not a prompt formatting issue, and makes "one-shot" use impractical.

Our pipeline, therefore, produces a candidate pool and relies on the Evaluator Agent to mitigate this variance, first ranking candidates on lane-region similarity to the reference design, then verifying guideline and instruction compliance with multimodal reasoning. Removing the evaluator would propagate inconsistent or noncompliant designs; with it, the system consistently selects the most suitable outcome from noisy generations.

D Human Evaluation Framework

While the use of MLLMs to assess generation correctness is well established [14, 35, 32, 33, 31], we evaluate our multi-agent bicycle infrastructure design pipeline via a stepwise, human-in-the-loop protocol and a design-quality assessment framework informed by prior work advocating staged generation with expert oversight and end-outcome metrics (realism, instruction adherence) [20], adapted to the specific requirements of bicycle-lane design and background preservation.

Human-in-the-loop Our workflow comprises four stages. First, during location description generation, we verify that the automatically produced textual description covers existing bicycle infrastructure while respecting established road layout(e.g., parking zone and traffic lane), and we edit the description where necessary. Second, in prompt optimization, we translate design objectives and constraints into verifiable clauses (e.g., "do not alter the background") and iterate until both the generation is effective and approval is obtained from human experts. Third, when converting the existing bicycle infrastructure into a highlight region, we check the spatial precision of the region

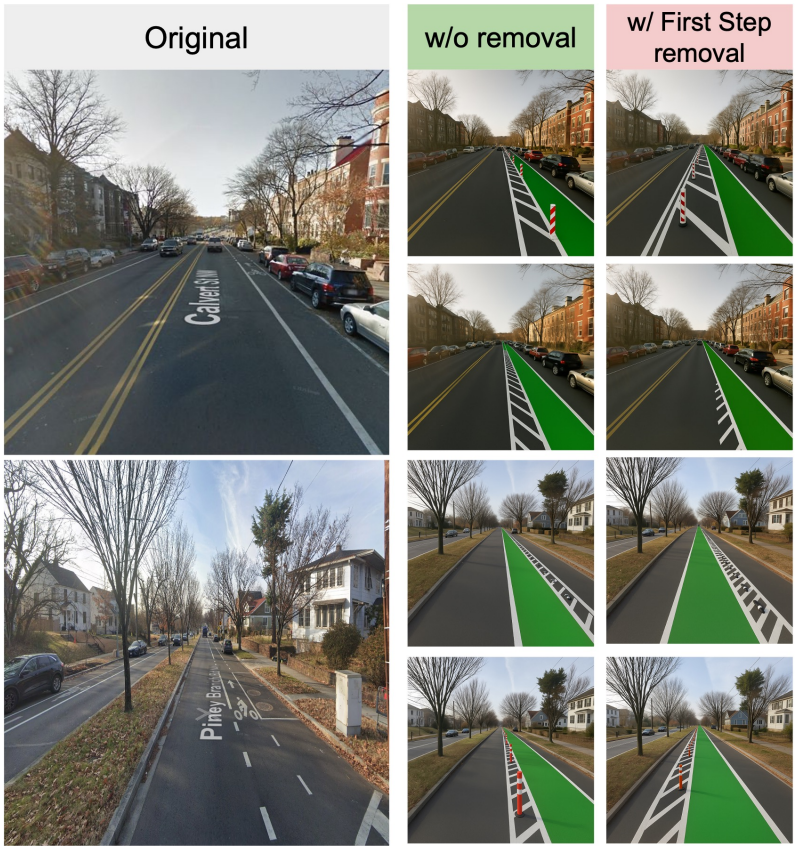

Figure 8: **Effect of removing the auxiliary step in Image Generation Agent.** Each row shows an original street-view image (left), the generation result with the full pipeline (middle), and the result after removing the auxiliary step (right).

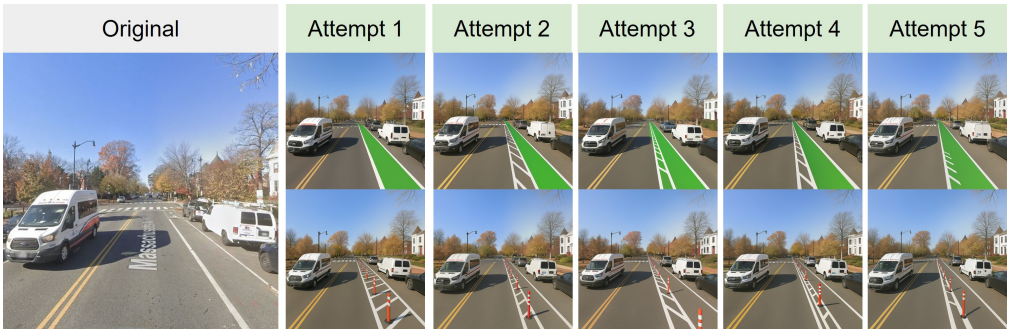

Figure 9: **Effect of removing the auxiliary step in Image Generation Agent.** Each row shows an original street-view image (left), the generation result with the full pipeline (middle), and the result after removing the auxiliary step (right).

and adjust boundaries to prevent leakage into the base scene if needed by re-generation. Fourth, in evaluator-agent selection, an evaluator agent nominates the best design from a candidate pool, while experts independently select their own top choice; disagreements trigger targeted upstream revisions (stages 1–3) before re-selection, ensuring expert control at the critical decision points. This step-by-step review-and-iterate design implements human guidance throughout the pipeline to avoid misalignment with the established design guidelines.

Final design quality assessment To complement the Evaluator Agent accuracy reported in Table 1, we conduct a human evaluation of the final bicycle-infrastructure design (the agent-selected, expert-approved output) along two axes: Visual fidelity and Instruction compliance. These axes explicitly target quality aspects that generic image metrics miss, especially hallucinations and subtle lane–background inconsistencies.

(i) Visual fidelity. We assess whether the edited bicycle infrastructure is realistic and seamlessly integrated into its context without unintended modifications outside the designated highlight region. Raters assign 1–5 Likert scores to three criteria—(i) lane plausibility (appropriate width, curvature continuity, geometric smoothness and connectivity), (ii) scene integration (consistent perspective, shading/shadows, reflections, and material/texture blending), and (iii) background preservation (absence of edits beyond the highlight, no semantic drift in roads, sidewalks, or street furniture). We aggregate these into a weighted composite score. The rubric is designed to capture common generative failure modes, including but not limited to: geometric artifacts (jagged edges, broken continuity), photometric/texture inconsistencies (incorrect shadows, tiling, color cast), perspective mismatch, improper occlusions (painted markings over vehicles/pedestrians), and background drift (unintended changes to non-highlighted regions).

(ii) Instruction compliance. We translate the optimized prompt into a requirement checklist comprising hard constraints (e.g., lane width, marking style) and soft constraints (e.g., "do not alter non-highlight background"). For each item, raters mark satisfied/unsatisfied, yielding a binary compliance vector and an overall compliance rate. We also collect a global 1–5 Likert judgment of prompt adherence. Any hallucinated or out-of-spec elements, such as spurious crosswalks/barriers, invented curb geometry, or contradictory markings, are counted as violations. Omissions of required features are likewise treated as non-compliance.

Furthermore, we treat expert judgment as the gold standard; we collapse the two-axis rubric to a single accuracy metric: a case is counted as correct iff it (i) meets the pre-specified Visual-Fidelity acceptance threshold (e.g., composite $\geq$ 4/5 with no background-change flag), and (ii) satisfies all hard instruction constraints (and at least the minimum soft-constraint level defined in the checklist). Accuracy is then reported as the proportion of test cases that satisfy these acceptance criteria.

In conclusion, our evaluation framework centers expert oversight at critical stages and evaluates the final design against human-defined standards of visual fidelity and instruction compliance, adapting stepwise, human-guided practices from generative urban design to the bicycle-lane setting.

E Limitations

Despite the effectiveness of the proposed multi-agent framework, several limitations remain. First, the current system still cannot fully guarantee pixel-level accuracy in representing spatial relationships within the generated designs. While the generated bike lanes are generally aligned with the intended roadway regions, fine-grained positional accuracy is not always achieved, particularly in complex street layouts. Second, the correctness rate of a single generation pass remains relatively low, requiring multiple candidate generations before a satisfactory result is obtained. This increases computational cost and latency in the design workflow. Finally, the pipeline still involves a substantial degree of human intervention, especially manual image selection during data preparation. Reducing this reliance on human involvement is essential for improving automation and scalability in future work.

Design Scenario	Left boundary	Right boundary
1	No buffer; direct adjacency to moving lane	No buffer; direct adjacency to parked cars
2	No buffer; direct adjacency to moving lane	3 ft white-painted buffer
3	No buffer; direct adjacency to moving lane	1.5 ft buffer with bollards
4	No buffer; direct adjacency to moving lane	1.5 ft buffer with armadillo lane dividers
5	No buffer; direct adjacency to moving lane	No buffer; direct edge (no separator)
6	3 ft white-painted buffer	No buffer; direct edge (no separator)
7	1.5 ft buffer with bollards	No buffer; direct edge (no separator)
8	1.5 ft buffer with armadillo lane dividers	No buffer; direct edge (no separator)

Table 2: **Design Scenario specifications for our experiments.**

F Prompts

Prompt of Locator Agent

System Prompt
You are a helpful vision assistant to identify the bike lane from the image and describe its location accurately.

User Prompt
Your task is to describe precisely the physical location and boundaries of the primary bike lane shown on the right side of the roadway in the provided image in sentences. Typically, bike lanes are defined clearly by white lines on both sides; however, boundary variations exist, and it is possible that one side of the bike lane may instead be marked by a buffer zone (e.g., a painted area with diagonal stripes), a curb, sidewalk edge, or physical separators like bollards or raised barriers. If either the left or right boundary is a buffer zone or another physical separator, treat that separator as part of the bike lane in your description. In this task, you should specifically detail that, in most cases, the primary bike lane is bordered by two parallel white lines: one white line forming the left boundary separating it from motor-vehicle lanes, and another white line forming the right boundary separating it from parking cars, sidewalks, or curbs. Clearly note the exact placement relative to these adjacent roadway features.

Figure 10: Prompt used in Locator Agent to request **GPT-o3** to describe the physical location and boundaries of the bicycle infrastructure from an image.

Prompt of Prompt Optimization Agent

System Prompt

You are an expert prompt optimizer. Your job is to transform a draft prompt about depicting or updating bicycle infrastructure in roadway images into a precise, unambiguous, self-contained instruction for an image-generation or editing model. Preserve the user's intent exactly; remove ambiguity; standardize terminology; and keep constraints measurable and consistent. Explicitly specify: (a) lane position on the right-hand side of the road; (b) lane width with units; (c) surface color policy (green vs. standard road surface); (d) left and right boundaries. Define the left boundary as the continuous solid white line separating the bike lane from motor-vehicle lanes; define the right boundary as either a continuous solid white line or a clearly marked buffer zone (diagonal white stripes) and, if present, any physical separators (e.g., bollards) should be described as part of the boundary, not within the lane. If the user forbids green paint, clearly say "No green paint." If the user requires a fully green lane, require that the green area is strictly contained between two continuous solid white lines. Do not invent parameters not present in the user input; when information is missing, keep statements general while matching the style of the examples. Write in clear imperative voice. Output only the optimized prompt—no commentary, headings, or quotes.

User Prompt

##Example 1##
The area currently painted green with two white boundary lines represents the existing bike lane. Your task is to clearly depict an updated bike lane that is approximately 6 feet wide, fully painted green, strictly contained between two prominent, continuous, solid white boundary lines. Ensure these white boundary lines clearly define the left and right edges of the green bike lane area, positioned along the right-hand side of the road. Do not allow any green paint to extend beyond the white boundary lines.

##Example 2##
The area currently painted green with two white boundary lines represents the existing bike lane. Clearly depict an updated bike lane approximately 6 feet wide, located along the right-hand side of the road. Do not paint the updated bike lane green; use the standard road surface color only. Clearly mark both boundaries of the bike lane: 1) Left boundary: a prominent, continuous solid white line. 2) Right boundary: a clearly marked narrow buffer zone adjacent to the bike lane, filled with prominent diagonal white stripes, and bounded on both sides by solid white lines. Ensure the updated bike lane is clearly defined by the solid white lines on both sides, distinctly separate from the striped buffer zone on its right side. No green paint should be applied.

##Example 3##
The area currently painted green with two white boundary lines represents the existing bike lane. Clearly depict an updated bike lane approximately 4 feet wide, located along the right-hand side of the road. Do not paint the updated bike lane green; use the standard road surface color only. Clearly mark both boundaries of the updated bike lane: 1) Left boundary: a prominent, continuous solid white line. 2) Right boundary: a clearly marked narrow buffer zone adjacent to the bike lane, filled with prominent diagonal white stripes, bounded on both sides by solid white lines, and distinctly featuring vertical red-and-white striped bollards placed at regular intervals. Ensure the updated bike lane is clearly defined by the solid white lines on both sides, distinctly separate from the striped buffer zone and bollards on its right side. No green paint should be applied.

Rewrite the *{USER_PROMPT}* into a single optimized prompt that follows the System Prompt and mirrors the style of the In-Context Examples. Preserve all stated constraints; make boundary definitions explicit (left vs. right); avoid contradictions; and keep the output concise (preferably $\leq$ 130 words). Output only the optimized prompt.

Figure 11: In-context learning prompt of Prompt Optimization Agent.

Prompt of auxiliary step in Design Generation Agent

System Prompt

You are a helpful assistant for precise roadway image editing. Follow instructions exactly, maintain visual realism (perspective, lighting, shadows), and avoid adding elements not requested. Preserve the legibility of traffic-control devices and do not alter objects unless explicitly instructed.

User Prompt

Edit the entire existing bike-lane corridor on the right side of the road into a {COLOR}-painted lane. Treat the corridor as the current bike-lane surface plus any immediately adjacent buffer zones or physical separators that belong to the lane configuration. Keep the lane strictly contained between two continuous solid white boundary lines (left and right). Apply the following:

– Boundaries: ensure both left and right boundaries are prominent, continuous solid white lines that follow roadway curvature. Do not let any {COLOR} paint cross, touch, or bleed over these lines.
– Buffer zones: if a narrow striped buffer exists adjacent to the lane, replace its interior stripes with a uniform {COLOR} infill, but retain the solid white lines that bound the buffer as the lane's outer edges.
– Physical separators: keep bollards, armadillos, curbs, or raised barriers intact and unpainted; they should remain visually above the {COLOR} base and aligned along the boundary. Do not recolor or remove them.
– Exclusions: exclude painted street names, arrows, lane labels, and crosswalk markings from recoloring. Do not extend the {COLOR} paint into motor-vehicle lanes, parking spaces, sidewalks, or curbs.
– Consistency: preserve road texture and lighting, respect occlusions (vehicles, pedestrians), and keep the lane's original footprint (do not widen or narrow).
– Output: deliver a clean, continuous {COLOR}-painted lane on the right side of the road, strictly bounded by solid white lines, with all exclusions observed.

Figure 12: Prompt of auxiliary step in Design Generation Agent

Prompt of Evaluator Agent

System Prompt
You are a strict binary evaluator for roadway images. Examine the candidate image (and the provided reference image, if any) and decide whether it shows a bike lane located on the right side of the road that satisfies *all* features listed in the User Prompt. Minor visual variations (e.g., perspective, lighting, small width deviations) are acceptable only if each required feature is clearly present and recognizable. If any required feature is missing, ambiguous, occluded, or contradicted, respond `no`.

Output format: a single lowercase word, exactly `yes` or `no`. Do not add punctuation, spaces, explanations, or any other text.

User Prompt
Answer ONLY `yes` or `no`:

Does the image show a bike lane on the right side of the road with the following key features? Minor variations are allowed, but all features should be clearly recognizable:

1. Left Boundary:
– Narrow buffer zone adjacent to the bike lane.
– Buffer zone bounded by solid white lines on both sides.
– Prominent diagonal white stripes filling the buffer zone.
– Rounded, semi-flexible rubber lane dividers (“armadillos”) placed centrally and evenly spaced within the buffer zone. Dividers should be dome-shaped, black with white reflective stripes.

2. Right Boundary:
– Prominent continuous solid white line marking the right-hand edge of the bike lane.

The image should closely match the reference image provided, clearly depicting both boundary conditions. Answer `no` if these conditions are not sufficiently met.

Figure 13: An example of a prompt for evaluation used in Evaluator Agent.