From Pixels to Policy: A Multi-Agent System for Intervention and Geo-Spatial Decision Support

Anonymous authors
Paper under double-blind review

Abstract

Urban environments are shaped by design choices with long-term implications for health, safety, and quality of life, yet evaluating proposed interventions remains costly, time-consuming, and often impractical. Existing geospatial vision methods largely focus on monitoring urban indicators from aerial and street-view imagery, rather than proposing interventions and estimating their effects on such indicators. Moving beyond recognition, we introduce the problem of discovering interventions that improve target indicators for a given aerial or street-view image. We argue that a black-box indicator model, combined with a generative editing model, can serve as an *implicit digital twin* for testing intervention hypotheses. We present VIDA-Geo , a multi-agent system that explores this intervention space by coordinating segmentation, diffusion-based inpainting, and indicator scoring models to produce interventions that are both perceptually realistic and aligned with real-world policies. We evaluate our system on 8 indicators across aerial and street-view imagery, measuring changes in factors such as perceived safety and greenery. Our approach outperforms existing baselines in many cases, achieving up to 2× higher perceptual quality and policy alignment scores. Finally, our model provides users with multiple candidate interventions, supporting an expert city-planner-in-the-loop workflow. Code is available at `https://anonymous.4open.science/r/VIDA-GEO-2558/`

1 Introduction

Visual geospatial data is used in several real-world applications for monitoring and decision-making. Urban planners make use of satellite images to monitor various indicators such as road safety (Cha et al., 2025; Elallaf et al., 2026) and greenery (Rahaman et al., 2024). Similarly, climate scientists make use of satellite images to understand pollution (Guo et al., 2025; Porcheddu et al., 2025), perceived temperatures (Ke et al., 2025; Briegel et al., 2024) and thermal exposure (Zaerpour et al., 2025; Dousset et al., 2011). Beyond satellite images, experts also make use of other visual information, such as street view imagery or surveillance footage, to monitor perceived safety (Zhang et al., 2024; Liu et al., 2023), road-safety (Yu et al., 2024; Wei et al., 2025), or walkability (Berriel et al., 2017; Nadai et al., 2016).

Computer vision research on such visual geospatial data has primarily focused on improving recognition, and consequently, we see highly performing models for many of the aforementioned tasks. However, monitoring via recognition systems is only half of the story. The goal of such planners and scientists is not just to monitor indicators but to intervene when necessary and make changes in accordance with policies. For example, an urban planner might suggest that a particular street is perceived as unsafe due to the presence of scaffolding, and therefore, based on the city's policy, the scaffolding must be replaced with safer scaffolding. While monitoring is well explored, identifying such recommendations remains underexplored.

Moving beyond visual monitoring, we propose a new framework for *visual intervention for decision support*. Given an input geospatial image of a region, such as an aerial or street view image, and a black-box indicator prediction model, such as road safety, our framework

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

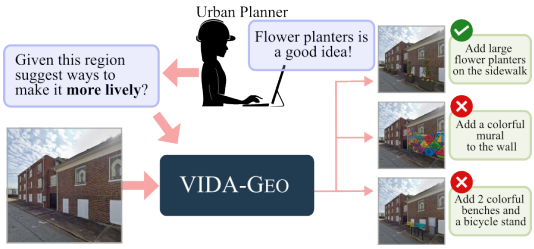


Figure 1: Given a geospatial image (street view or aerial), a user can ask VIDA-Geo to suggest interventions that improve a target indicator, such as liveliness. Our system proposes **multiple** realistic, policy-grounded interventions to support decision-making. The user can then choose the intervention most suitable for their use case.

outputs realistic interventions to the image that improve the indicator score. Our framework could aid policymakers in domains such as urban planning or climate science by providing a diverse set of interventions before implementing interventions in the real world (see fig. 1).

The current paradigm for finding interventions in urban planning is to run simulations in digital twins of a city (Herzog et al., 2025; Ghaffarian, 2025). However, this has several challenges: 1) they require regular and expensive data collection to maintain the digital twin, 2) the verification of interventions requires expensive and time-consuming simulations, and 3) they could only model specific types of interventions depending on the simulation model.

We instead posit that, rather than creating expensive digital twins, *a combination of an indicator monitoring model (recognition) and an image editing model (generative model) can serve as an implicit digital twin (IDT) for creating such interventions.* Indicator recognition/monitoring models (Naik et al., 2014; Larkin et al., 2021) have been shown to be highly accurate. Moreover, as we show in the results (see fig. 8), they are also good at handling counterfactuals, which is an important requirement to be treated as a component of an implicit digital twin. However, a prediction model alone cannot serve as an IDT, as it cannot generate interventions. Modern generative models have both realistic generation capabilities and controllable editing (Rombach et al., 2022; Wang et al., 2025), including in domains such as satellite imagery (Cher et al., 2025; Sastry et al., 2024). Therefore, we use generative editing models to generate realistic interventions.

However, one key question remains: how do we propose realistic interventions that can actually be deployed in the real world, rather than simply optimizing the indicator score? To ensure that interventions are grounded from a policymaker's perspective, we use large language models (LLMs) for intervention proposals. Specifically, we introduce a multi-agent system with several sub-agents. We start with a high-level *planner* agent that, given an image, identifies regions of interest (ROIs) that can potentially be intervened upon. The ROIs, along with the image, are passed to a *policy sub-agent*, which uses real-world knowledge to suggest realistic interventions in the ROIs. These ROIs are also passed to a *segmentation agent*, which uses several foundation models to find relevant masks for intervention. The *generation sub-agent* then uses these prompts to generate realistic edits. The edited images are evaluated both for improvement in the black-box indicator score and for whether they remain grounded in the goals of the *policy agent*. All trials are kept in memory and passed back to the *planner agent*, which decides the next set of ROIs. Framing this as a multi-agent system allows interventions to be regionally and policy-grounded while remaining modular and understandable to experts. We call this a multi-agent system **VIDA-Geo** (**V**isual **I**ntervention and **D**ecision-support **A**gent for **Geo**spatial data).

To evaluate our method and future works on this problem, we introduce a new benchmark with evaluation metrics for these 8 tasks. We show that our agentic method can find realistic and grounded interventions by testing it on satellite and street view images across 8 indicator tasks. Our main contributions are as follows.

- We introduce a new problem setup that moves beyond geospatial monitoring towards intervention for policy-based decision-making.
- We provide a new benchmark for evaluating this task, with evaluation metrics that measure the realism of interventions from both image-based and policy-oriented perspectives.
- We introduce a multi-agent system VIDA-Geo, that proposes, executes, and quantitatively evaluates geospatial interventions. We frame this system as an Implicit Digital Twin and show that it outperforms methods related to this problem setting.

2 Related Works

Urban Indicator Estimation Overhead-images have been used for safety and mobility indicators, including crash-risk signals (Najjar et al., 2017; He et al., 2021; Elallaf et al., 2026) and pedestrian-network extraction through sidewalks (Hosseini et al., 2023; Nadai et al., 2016). Segmentation methods (Xia et al., 2023; Demir et al., 2018) have also been used to measure vegetation (Li et al., 2015; Wolch et al., 2014). Street-view imagery has been widely used to study perceived urban qualities such as safety and liveliness (Dubey et al., 2016; Hou et al., 2024; Naik et al., 2014). Recent work also uses it to model human perceptions (Danish et al., 2025; Hamim & Ukkusuri, 2024) and for multimodal LLM-based evaluation (Zhou et al., 2025). While such models are well established, we use them to measure and guide visual interventions that improve an indicator score at a given place.

Intervention and Decision Making Digital twins combine city data and simulations to support evidence-based decision making (Ghaffarian, 2025; Cecere et al., 2025). However, such interventions are domain-specific and expensive to simulate. Prior works use counterfactual models to understand the inputs' effect on the model's prediction (Jacob et al., 2022; Zemni et al., 2023). Others use diffusion models to generate counterfactual explanations for black-box models (Farid et al., 2023; Chiquier et al., 2025; Prabhu et al., 2023). While these works aim to understand model behavior, our goal is instead to use them as real-world actions.

Agents for Intervention and Interpretability Remote-sensing VLMs extend instruction-following models to overhead imagery (Kuckreja et al., 2024; Zhan et al., 2025), with recent work exploring agentic systems as well (Xu et al., 2024; Shabbir et al., 2025; Kao et al., 2025). Other works show how agents can design experiments to explain model behavior and discover failure modes (Shaham et al., 2025; Schwettmann et al., 2023). Related works use generative models to simulate street-view interventions and changes in perceived urban quality (Hu et al., 2025; Zhao et al., 2026). Unlike these works, VIDA-Geo formulates urban intervention as a multimodal decision problem. Moreover, the proposed actions are optimized not only for target urban indicators but also for physical realism and policy feasibility.

3 Methodology

VIDA-Geo is a multi-agent system that takes an image and an improvement goal instruction to produce grounded intervention proposals consisting of *a region, an edit, and a textual description*. Instead of editing the whole image in one step, it decomposes the problem into specialized agents that reason about where to intervene, what to change, which changes are permissible, and how to generate plausible edits. This creates changes that are localized and atomic and therefore much easier to implement in the real world. The agentic system includes a multi-epoch feedback loop, where summaries of prior epochs inform the subsequent planning and steering away from repeated failure modes.

The VIDA-Geo pipeline (see fig. 2) consists of four specialized agents followed by a scoring and multi-epoch orchestration. The **Planning Agent** decides where and what to change; the **Policy Agent** determines which changes are permissible and which regions are constrained; the **Segmentation Agent** identifies which pixels correspond to each region of interest; and the **Generation Agent** creates the visual interventions through prompted inpainting. Candidate interventions are evaluated by the black-box scorer, and the resulting outcomes are used by the **Orchestration** mechanism to guide later epochs. Each epoch corresponds to one complete planning-to-evaluation cycle. We next present these agents.

3.1 Planning Agent

Directly editing the entire image is highly unconstrained and may produce unrealistic or infeasible changes. Therefore, the **Planning Agent** first identifies and ranks multiple ROIs in the image as candidate intervention locations. For each ROI, the planner produces ranked spatial coordinates, bounding boxes, and semantic descriptions or keywords required for the downstream segmentation task in section 3.3. In later epochs, the planner additionally receives a summary of previous attempts; we describe this feedback mechanism in section 3.5.

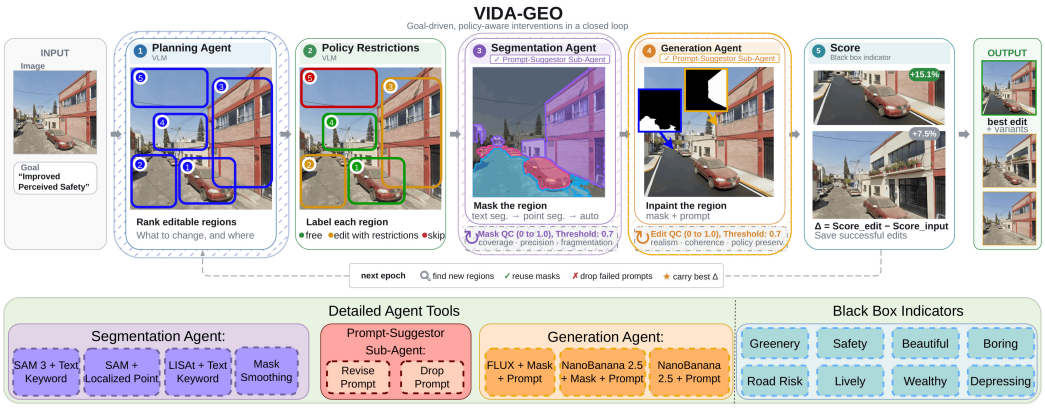


Figure 2: Overview of the VIDA-Geo pipeline. Given an image and a target urban-improvement instruction, a group of agents identifies ROIs, segments, and edits with task-specific constraints to produce candidate intervention edits. Quality checks assess realism and policy preservation. Candidates are finally evaluated by a black-box indicator scorer.

3.2 Policy Agent

The **Policy Agent** then determines the physical and semantic constraints governing each ROI. These constraints fall into three categories: *free* regions, where any modification is permissible; *constrained* regions, where edits are allowed but must respect specific rules; and *skip* regions, which contain critical infrastructure that must remain unaltered. These constraints determine how each ROI may be modified and are also passed to the Generation Agent in section 3.4. For street-view images, the policy agent also enforces general street-scene plausibility, ensuring that buildings, entrances, and other functional elements remain coherent in any edited output. An example of such policy violations can be seen in fig. 5a.

3.3 Segmentation Agent

The **Segmentation Agent** receives each ROI from the planner and segments each region using one of several segmentation tools (fig. 2 (bottom)). These include text-referred segmentation driven by the planner's suggested prompt and open-vocabulary models using the planner's keyword. When both fail, the agent falls back to point-prompted segmentation models from the planner's estimated ROI center. Examples of segmentation failures can be seen in fig. 5b and the fallback for segmentation is provided in section A.

Segmentation QC Sub-Agent. Because different segmentation methods can produce inaccurate, overly broad, or fragmented masks, each candidate mask is evaluated by an LLM-based quality-control (QC) step that compares the original image against the binary mask. This evaluation considers three criteria. *Coverage* checks whether the mask spans the correct target area. *Precision* estimates how much unrelated regions are excluded. *Fragmentation* measures how spatially connected the mask is, penalizing, e.g., floating islands.

Feedback loop. Failed masks and their QC descriptions are returned to the Segmentation Agent to adjust the mask selection strategy. The Prompt Suggestion sub-agent revises the segmentation keyword or prompt. This inner closed-loop mechanism improves poor-quality masks before presenting them to downstream intervention stages.

3.4 Generation Agent

The **Generation Agent** receives the input image and all accepted masks to produce edited images intended to improve the black-box indicator. The images are edited through a mask-and-text-prompted diffusion process, with the *Policy Agent*'s constraint annotations injected into the text prompt. These constraints ensure that edits remain consistent with the

physical and functional structure of the scene. When mask-based inpainting is not feasible, the agent falls back on mask-free text-guided generation models (see examples in section A).

Generation QC Sub-Agent. Because edits may contain unrealistic, inconsistent, or semantically incorrect modifications, the images undergo a QC step that evaluates three properties. *Realism* checks whether the edited region can plausibly exist in the real world, flagging, e.g., unnatural textures. *Coherence* verifies that the edit is contextually consistent with the rest of the image. Finally, *prompt adherence* confirms that the generation agent followed the suggested editing prompt. The complete prompts are present in section G.

Policy compliance check. As previously mentioned in section 3.2, each ROI is associated with policy constraints, verified using the Generation QC to ensure the physical coherence of edits. For example, a residential cluster replacing a retail parking lot is implausible policywise. Violations are fed back to the generation agent so that it can suggest better-aligned prompts. In this work, the regulations come from the LLM's knowledge; however, *a future direction would be to improve policy compliance by incorporating city regulations rulebooks.*

Feedback loop. Failed outputs, including examples shown in fig. 5c, and their corresponding QC feedback are returned to the Generation Agent. The agent analyzes whether the edits lacked realism, broke coherence, or deviated from the prompt. The Prompt Suggestion sub-agent then produces a revised editing prompt for the next generation attempt.

3.5 Scoring and Multi-Epoch Orchestration

Finally, for each ROI, multiple edited images may pass quality control. All accepted candidates are scored using the black-box indicator and compared against the original image. Candidates are ranked based on the difference between the two scores (delta score). This process is repeated over multiple cycles, with the results from each cycle passed back to the planning agent. To encourage the agents to identify new regions, produce new masks, and generate alternative edits, all previous candidates and their corresponding delta scores are retained and carried forward to the subsequent cycle.

This accumulated history discourages the system from repeating unsuccessful strategies while encouraging the exploration of diverse editing directions. As a result, this approach yields multiple edited variants from a single input image, each of which adheres to the policy constraints while improving the black-box indicator score.

4 Results

In this section, we discuss the implementation details of our agent, present the benchmark and baselines, and show that our multi-agent system provides realistic and grounded interventions. The implementation details are presented in the appendix (section F).

4.1 Benchmark

To perform a comprehensive evaluation with different types of visual geospatial modalities, we sample both aerial images and Street View images. More specifically, we source high-resolution MapBox aerial images and Google Street View images to evaluate interventions.

For intervention and decision-making, we use black-box indicator prediction models. For aerial images, we select two models: First is BetaRisk model (Elallaf et al., 2026), trained to predict fatal-crash risk values. The second is OpenEarthMap (Xia et al., 2023) used for greenery prediction. For Street View images, we use the models released with Global Streetscapes (Hou et al., 2024), which were pretrained on MIT Place Pulse 2.0 (Dubey et al., 2016), and predict six standard urban perception attributes: safety, lively, beautiful, wealthy, less boring, and less depressing, on a 0–10 scale which we rescale to 0–100.

Because the evaluation requires repeated multi-agent and generative model interactions, we prioritize careful assessment over dataset size. For each of the 8 indicators, we score randomly sampled images with the corresponding black-box model and filter out those already scoring

above 90. From the remaining pool, a VLM selects 100 images per indicator, ensuring diversity in score, geography, and appearance. See section K for benchmark statistics.

4.2 Evaluation Metrics

Evaluating interventions is a challenging problem, as VIDA-Geo and baselines can generate a wide range of possible interventions, and it is not directly observable whether these interventions would be effective in the real world. Therefore, we evaluate such interventions across two complementary axes. First, the ***perceptual quality*** of the generated interventions; by determining whether they appear perceptually realistic. Second, the ***policy alignment***; by determining whether the interventions are realistic and actionable from a policy perspective.

FID-Proxy Through Mahalanobis Distance is used to assess if generated edits remain within the distribution of real images (Mahalanobis, 2018). FID requires a set of generated images; however, since we have only one image at a time, we use a proxy (see section B).

LLM-as-a-Judge Evaluation. Since human evaluation at scale is expensive, we primarily rely on LLM judges (Zheng et al., 2023), using GPT 5.2 and Qwen 3VL (Singh et al., 2025; Bai et al., 2025) (results in section J). Each edited image is scored independently (more information in section G) on a discrete 1-10 (later scaled to 1-100) scale across three criteria:

- **Visual quality**: The technical quality of the edit and how natural the image looks. This metric also measures the *perceptual quality*.
- **Realism**: Assess whether the edited world is physically and semantically plausible. Objects should have plausible function, material, scale, geometry, and spatial relationships.
- **Policy preservation**: How well the edit preserves the original scene's structure, layout, function, and geographic identity, while limiting changes to the intended local intervention.

A composite judge score is computed as the mean of these 3 LLM-as-a-judge scores. The latter two metrics assess whether the intervention is plausible and *policy-aligned*.

Human Expert Evaluation. We additionally evaluate a subset of our data with seven graduate students and five experts recruited through Prolific (Prolific, 2026), specializing in urban planning and policy design (see section C.1). Evaluators make pairwise comparisons between VIDA-Geo and a selected baseline, providing human validation of perceptual realism and policy compatibility beyond LLM-based evaluation. We also report high agreement between LLM and expert decisions, suggesting that LLM judges can serve as a scalable and reproducible proxy for human judgments (see section C.2).

4.3 Baselines

Since we propose a new problem of discovering interventions, there are not many relevant strong baselines. The most closely related method is URSimulator (Hu et al., 2025), which performs perception-guided iterative editing of street-view images with a diffusion model. URSimulator tunes text prompts according to predicted human-perception scores but does not adhere to realism and policy. We additionally adapt methods that generate counterfactuals for interpretability to compare with VIDA-Geo. DIFFusion (Chiquier et al., 2025) interprets differences between 2 classes by using a diffusion model to transform an image's class. It is a differentiable method that edits via gradient optimization instead of a grounded discrete search like VIDA-Geo. LANCE (Prabhu et al., 2023) utilizes discrete search to find good counterfactuals for a vision classifier. We use it to find edits maximizing indicator scores. We also compare our method against NanoBanana 2.5 in the zero-shot setting (NB2.5 (ZS)), where we directly prompt it to edit an image to improve the black-box score. For fair comparison, all models get the same number of calls to generative models.

4.4 Results and Discussion

We first test VIDA-Geo quantitatively against the aforementioned baselines. We observe that it achieves the highest overall performance on both performance axes. Additionally, we qualitatively show that VIDA-Geo produces multiple plausible interventions that increase the indicator scores and can be used to provide multiple interventions sequentially.

Table 1: Task performance, perceptual quality, and policy alignment for edited images judged with **GPT5.2**. Green and red tasks favor higher and lower task scores respectively. Delta Avg. is signed toward the task objective. FID-Proxy and judge-score columns report average *95% confidence-intervals*. The best judge score for each task and evaluation is boldfaced.

Task	Method	Output Avg. (%)	Delta Avg. (%)	Perceptual Quality		Policy Alignment		
				FID-Proxy (↓) ± 1.00	Visual Quality (%) ± 0.70	Realism (%) ± 0.87	Policy Pres. (%) ± 1.25	LLM Judge Avg. (%) ± 0.76
Greenery	**VIDA-Geo**	35.6	6.3	44.8	72.2	67.2	64.1	67.8
	LANCE	25.3	-4.1	57.3	37.5	28.6	16.9	27.7
	DIFFusion	59.5	30.1	59.6	26.1	25.5	26.0	25.9
	NB2.5 (ZS)	43.0	13.6	44.7	71.9	66.5	66.1	68.2
Road Risk	**VIDA-Geo**	82.1	4.0	**47.5**	**72.8**	**67.4**	57.8	**66.0**
	LANCE	68.7	17.4	58.2	38.6	31.5	19.0	29.7
	DIFFusion	13.3	72.8	60.2	24.6	26.8	26.0	25.8
	NB2.5 (ZS)	81.6	4.5	49.8	68.1	61.9	57.5	62.5
Safety	**VIDA-Geo**	20.9	3.8	56.5	**66.5**	**73.5**	**70.3**	**70.1**
	LANCE	33.8	16.6	61.0	58.5	55.8	32.8	49.0
	DIFFusion	50.5	33.4	70.3	28.6	35.0	32.1	31.9
	URSimulator	33.0	15.9	57.4	57.9	50.8	40.0	49.6
	NB2.5 (ZS)	41.3	24.2	61.4	62.3	61.8	40.1	54.7
Lively	**VIDA-Geo**	25.1	11.4	61.5	**61.4**	**63.6**	**65.6**	**63.5**
	LANCE	23.0	9.4	62.6	59.4	53.9	29.3	47.6
	DIFFusion	39.5	25.8	66.2	34.9	42.2	36.5	37.8
	URSimulator	29.6	16.0	**58.8**	52.7	44.7	35.4	44.3
	NB2.5 (ZS)	52.0	38.3	68.2	54.2	53.0	46.8	51.3
Beautiful	**VIDA-Geo**	26.1	8.4	54.5	**65.7**	**74.7**	**72.4**	**70.9**
	LANCE	42.1	24.4	58.2	56.4	52.3	29.9	46.2
	DIFFusion	53.0	35.3	71.8	25.9	29.9	27.6	27.8
	URSimulator	32.8	15.1	54.9	56.9	52.5	41.3	50.2
	NB2.5 (ZS)	43.7	25.9	61.9	53.8	66.3	53.9	58.0
Wealthy	**VIDA-Geo**	27.0	9.1	58.0	**65.6**	**70.8**	**62.8**	**66.4**
	LANCE	35.4	17.4	61.0	58.9	55.2	31.0	48.4
	DIFFusion	54.9	37.0	69.8	31.0	38.7	35.7	35.1
	URSimulator	35.5	17.5	57.9	56.3	47.9	35.9	46.7
	NB2.5 (ZS)	68.8	50.8	62.5	59.8	38.4	15.8	38.0
Boring	**VIDA-Geo**	74.6	6.0	58.2	**65.7**	**72.2**	**68.1**	**68.7**
	LANCE	62.8	17.8	59.5	59.8	54.3	30.1	48.1
	DIFFusion	55.2	25.4	62.7	35.1	42.2	37.4	38.2
	URSimulator	64.4	16.2	60.5	53.1	43.2	33.4	43.2
	NB2.5 (ZS)	61.1	19.4	74.6	42.8	35.8	31.9	36.9
Depressing	**VIDA-Geo**	75.5	4.6	57.6	**65.6**	**73.5**	**70.2**	**69.7**
	LANCE	58.9	21.2	61.0	57.3	51.6	28.1	45.7
	DIFFusion	52.3	27.8	69.8	30.2	37.7	33.2	33.7
	URSimulator	71.2	8.9	57.9	54.9	48.8	38.3	47.3
	NB2.5 (ZS)	57.9	22.2	63.4	60.1	61.3	46.2	55.9

Can VIDA-Geo produce good interventions? We compare VIDA-Geo against other baselines on all 8 tasks in table 1. VIDA-Geo consistently obtains the highest LLM-judge scores on 7 tasks, often scoring 15pp higher than the other models, indicating that its edits are judged to be more policy-aligned and visually coherent compared to the baselines.

In several tasks, DIFFusion produces the largest changes, such as reducing road risk by 72.8%, but these edits receive much lower perceptual-quality and policy-alignment scores. As for LANCE, it may change the target indicator in the unintended direction and receive lower LLM-judge scores than VIDA-Geo, suggesting that its interventions are less actionable. The overall results suggest that VIDA-Geo can produce good interventions in terms of perceptual quality and can adhere much better to real-world intervention policies.

URSimulator only applies to street view images, where it tends to be more conservative in making edits; moreover, it shows a noticeable gap in policy alignment compared to VIDA-Geo. The zero-shot baseline NB2.5 (ZS) which, unlike VIDA-Geo, is not explicitly designed to search for interventions that satisfy real-world policy constraints. Nevertheless, it achieves a larger task-directed change than VIDA-Geo. For example, *Lively* and *Wealthy* scores are improved by 39.1% and 50.8%, respectively, compared to 12.7% and 10.2% for VIDA-Geo. *However, these larger changes do not consistently translate into high-quality or realistic interventions.* The exception is the *Greenery* task, where NB2.5 (ZS) performs comparably to VIDA-Geo in perceptual and policy-alignment quality. However, this task is

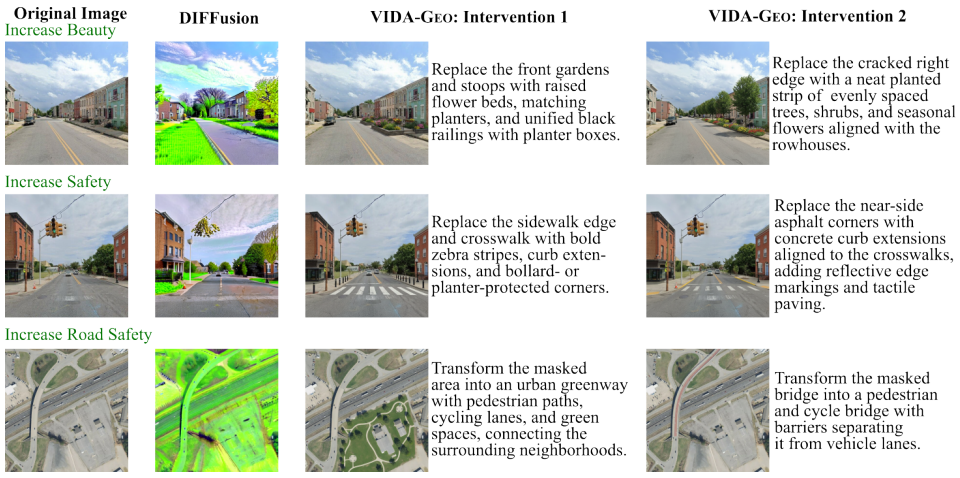


Figure 3: Comparison of intervention outputs from VIDA-Geo vs. DIFFusion for street view (first 2 rows) and aerial inputs (last row). The interventions provided by VIDA-Geo are more practical and realizable in the real world, e.g., adding bollards and tactile pavements around the intersection in the second. Moreover, VIDA-Geo can provide text instructions along with the images, unlike DIFFusion, which is more useful to a user.

relatively simple, as increasing greenery can often be achieved through simple visual edits such as adding vegetation in empty spaces without requiring complex reasoning.

Do VIDA-Geo interventions make sense to experts? We also asked domain experts to pairwise compare the VIDA-Geo against DIFFusion and LANCE for the same inputs. The evaluators were asked, "Which image better reflects the target attribute, which appears more realistic, and which edit would be more feasible in urban planning?". The results reveal a discrepancy between black-box models and human evaluation. Against DIFFusion, experts prefer our agent's outputs 61.3% of the time on perceptual attribute alignment, compared to only 10% of the black-box models favoring VIDA-Geo. The same trends held against LANCE, where experts preferred VIDA-Geo 86.7% of the time on perceptual attributes and 79.3% of the time on feasibility. Overall, VIDA-Geo achieves the strongest overall human-aligned performance (See section C for more results).

While experts disagree with the black-box model, they have a high level of agreement with LLM judges, averaging 88.5% on the *realism* and 91.6% on the *policy-preservation* metric (random baselines is 50%). This indicates that LLMs provide evaluations closely aligned with expert judges and can serve as a scalable proxy for more expensive human evaluations.

Finally, VIDA-Geo's interventions can resemble real-world changes. Such changes increase black-box beauty and safety scores by an average of 24.3% and 12.46%, respectively, through interventions such as extended curb geometry, tactile surfaces, and added vegetation. Similar interventions emerge from VIDA-Geo (see section M).

How do interventions look qualitatively? Figure 3 presents example interventions for 3 tasks. In most cases, the interventions suggested by DIFFusion are harder to interpret, perceptually unrealistic, and not policy-realizable. In the second row, it removes the traffic light, making the region less safe. In the aerial image, it mainly increases the saturation of the green channel, which is both policy-wise impractical and perceptually unrealistic.

VIDA-Geo, on the other hand, provides multiple interventions (along with their text descriptions) that look realistic and are easier to implement. For aerial images, it suggests replacing large parking lots with greener parks, replacing a road bridge with a cycling bridge, and making roads safer, especially for pedestrians (refer to section L for more examples).

Since aerial images represent a larger physical area, even small interventions, such as replacing the bridge, would be a much bigger undertaking for an urban planner. This explains the relatively lower scores for aerial image tasks in table 1 compared to street-view tasks.

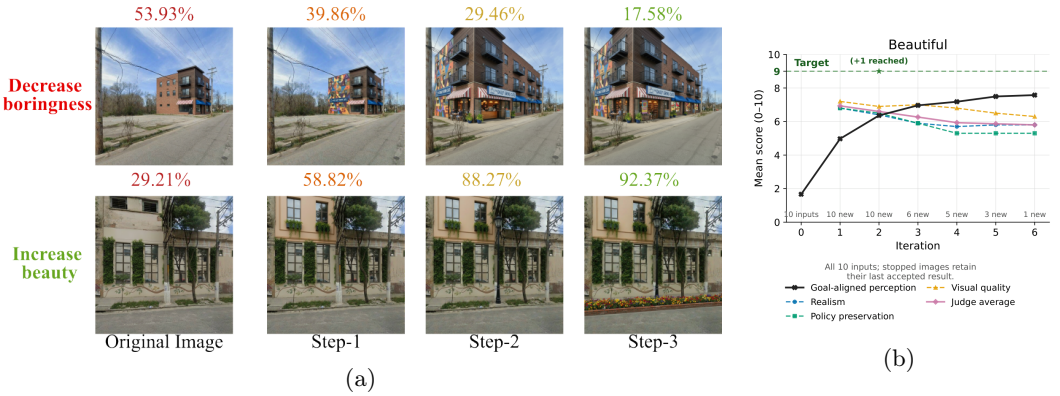


Figure 4: Using VIDA-Geo with multiple sequential edits. (a) It suggests multiple edits in different regions and increases the indicator score. For example, in the bottom row, it improves the building facade, then adds a lamppost, followed by roadside flowers, increasing the beauty score from 29.2% to 92.4%. (b) Quantitative progression over successive iterations.

Can VIDA-Geo do sequential edits? In fig. 4, we show the results of applying VIDA-Geo sequentially to outputs from a prior run. These sequential edits produce interventions that align with policy and increase the indicator score more than a single edit. For example, in the top row, VIDA-Geo first adds a mural, then expands the buildings, and finally adds texture to the window. The boring score decreases from 53% to 17%, while with single-edit we get the score to 40% at best. This suggests that when urban planners seek larger changes, VIDA-Geo can propose sequential edits that progressively improve the target indicator while largely preserving quality and policy alignment. As shown in fig. 4b, goal-aligned scores increase with iterations, while LLM-judges metric decrease slightly, suggesting a reasonable trade-off. Additional sequential analysis can be found in section D.

Ablations: Full ablations of VIDA-Geo can be found in section N. We show that policy constraints are the largest contributor to policy-aligned outputs. Edit QC and Mask QC also provide substantial gains, and the suggestor offers smaller but consistent improvements.

5 Conclusion, Limitations, and Future Work

Conclusion: We presented a new problem setup moving beyond monitoring towards testing interventions and decision-making. Testing and evaluating successful intervention is a challenge; therefore, we proposed an evaluation benchmark of 8 different tasks across aerial and street view images, consisting of 800 images. We also proposed new metrics to evaluate the quality of such interventions, both perceptually and policywise. To tackle the problem, we introduced VIDA-Geo a multi-agentic method that suggests good grounded interventions. We believe such intervention can be used as decision support to real-life urban planners and policymakers when deciding to change a physical space in the real world.

Limitations and Future Work: Since our work introduces the new problem of visual intervention for geospatial support, we present the first method as a solution. We believe there is substantial room for improvement. For example, in the current system, edits generated by FLUX, especially for aerial images, frequently fail quality checks, leading to over-reliance on NanoBanana. Additionally, replacing closed-source models as agents with open-source counterparts is possible but degrades quality significantly. Furthermore, the current method relies on multiple mask/edit quality checks to ensure good outputs, which could be reduced with better segmentation and editing tools. Finally, this work does not directly encode full zoning law, engineering standards, or construction budgets. However, it is intended as an early-stage decision support for urban planners and policymakers, rather than an implementation system. The generated edits and their indicator scores provide visual proxies that can help experts explore and compare candidate interventions, and we hope that future work will include such constraints to further improve the intervention framework.

6 Ethics Statement

We believe our method follows the conference ethics guidelines. We do not redistribute Google Street View or Mapbox images. We only release independently curated coordinates and camera orientation for reproducibility. There are no class labels to be included. Google Street View, Mapbox, and all original creators of the discussed indicator models are credited and cited.

The appendix includes screenshots of the instructions given to the evaluators. The subsection "Human Evaluation" under the Evaluation Metrics section also adds details about the human evaluation process.

The study protocol was reviewed and **approved** by the university's Institutional Review Board (IRB). The study consisted of a minimal-risk annotation task in which adult participants evaluated AI-generated outputs. No identifiable, sensitive, personal, or demographic information was collected, and responses were analyzed only in aggregate. The evaluators were graduate-student volunteers and urban planning and design experts. During recruitment, student participants were informed about the task and provided informed (verbal) consent for their responses to be used in the study. Participation was voluntary, and no names, contact details, or other personally identifying information were collected. The experts evaluated the outputs through Prolific and were paid $29.00/hr.

7 Reproducibility statement

We have shared the anonymized code repository and will share the dataset metadata needed to reproduce the experiments, including the location coordinates and camera angles associated with the Street View imagery (without redistributing the images themselves). Section F details the experimental setup and implementation details. Our experiments relied on API access to GPT-5.1, GPT-5.2, Gemini 2.5 Flash Image (NanoBanana 2.5) and Qwen3-VL-32B.

The anonymous repository for VIDA-Geo is available at `https://anonymous.4open.science/r/VIDA-GEO-2558/`. It includes configuration and setup instructions for the agentic tools, along with the complete set of prompts used for intervention generation and the LLM-as-judge evaluations.

8 AI Use Statement

In this work, generative AI tools were used as part of the agentic pipeline. Our multi-agent modules require LLMs to make decisions, evaluate outputs, and provide suggestions for the next stages as detailed in the architecture of this work. We also generate and inpaint images using generative models. Finally, our evaluation uses two different LLMs as judges.

Separately from their role in the method, we use AI tools to assist with editing grammar and fixing spelling mistakes. We also used agentic tools to help implement code to create charts and prepare figures included in the paper. All AI-assisted outputs were reviewed and verified by authors for correctness.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Rodrigo F. Berriel, André Teixeira Lopes, Alberto F. de Souza, and Thiago Oliveira-Santos. Deep learning-based large-scale automatic satellite crosswalk classification. *IEEE Geoscience and Remote Sensing Letters*, 14(9):1513–1517, 2017. doi: 10.1109/LGRS.2017.2719863.

Black Forest Labs. FLUX.1 Fill [dev]. `https://huggingface.co/black-forest-labs/FLUX.1-Fill-dev`, 2024. Hugging Face model card; accessed 2026-05-07.

Ferdinand Briegel, Jonas Wehrle, Dirk Schindler, and Andreas Christen. High-resolution multi-scaling of outdoor human thermal comfort and its intra-urban variability based on machine learning. *Geoscientific model development*, 17(4):1667–1688, 2024.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2026. URL https://arxiv.org/abs/2511.16719.

Liliana Cecere, Michele Grimaldi, Angelo Lorusso, Alessandra Marra, and Federica Stoia. Immersive urban planning: Evaluating park safety perception with digital twins and metaverse simulation. *Sustainability*, 17(17):7608, 2025.

Jungeun Cha, Seunghyeok Lee, and Hoe-Kyoung Kim. Deep learning-based detection and assessment of road damage caused by disaster with satellite imagery. *Applied Sciences*, 15 (14), 2025. ISSN 2076-3417. doi: 10.3390/app15147669. URL https://www.mdpi.com/2076-3417/15/14/7669.

Daniel Cher, Brian Wei, Srikumar Sastry, and Nathan Jacobs. Vectorsynth: Fine-grained satellite image synthesis with structured semantics, 2025. URL https://arxiv.org/abs/2511.07744. Accessed: 2026-01-15.

Mia Chiquier, Orr Avrech, Yossi Gandelsman, Berthy Feng, Katherine Bouman, and Carl Vondrick. Teaching humans subtle differences with diffusion, 2025. URL https://arxiv.org/abs/2504.08046.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

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

Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 172–181, 2018.

Bénédicte Dousset, Françoise Gourmelon, Karine Laaidi, Abdelkrim Zeghnoun, Emmanuel Giraudet, Philippe Bretin, Elena Mauri, and Stéphanie Vandentorren. Satellite monitoring of summer heat waves in the paris metropolitan area. *International Journal of Climatology*, 31(2):313–323, 2011.

Abhimanyu Dubey, Nikhil Naik, Devi Parikh, Ramesh Raskar, and César A. Hidalgo. Deep learning the city : Quantifying urban perception at a global scale, 2016. URL https://arxiv.org/abs/1608.01769.

Ahmad Elallaf, Nathan Jacobs, Xinyue Ye, Mei Chen, and Gongbo Liang. Beta distribution learning for reliable roadway crash risk assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Karim Farid, Simon Schrodi, Max Argus, and Thomas Brox. Latent diffusion counterfactual explanations, 2023. URL https://arxiv.org/abs/2310.06668.

Saman Ghaffarian. Rethinking digital twin: Introducing digital risk twin for disaster risk management. *npj Natural Hazards*, 2(1):79, 2025.

Wenxuan Guo, Zhaoping Hu, Ling Jin, Yanyan Xu, and Marta C Gonzalez. Deepair: deep learning and satellite imagery to estimate high-resolution pm2. 5 at scale. *Machine Learning: Science and Technology*, 6(1):015057, 2025.

Omar Faruqe Hamim and Satish V. Ukkusuri. Towards safer streets: A framework for unveiling pedestrians' perceived road safety using street view imagery. *Accident Analysis & Prevention*, 195:107400, 2024. ISSN 0001-4575. doi: https://doi.org/10.1016/j.aap.2023.107400. URL `https://www.sciencedirect.com/science/article/pii/S0001457523004475`.

Songtao He, Mohammad Amin Sadeghi, Sanjay Chawla, Mohammad Alizadeh, Hari Balakrishnan, and Samuel Madden. Inferring high-resolution traffic accident risk maps based on satellite imagery and gps trajectories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11977–11985, 2021.

Rico H Herzog, Till Degkwitz, and Trivik Verma. The urban model platform: A public backbone for modeling and simulation in urban digital twins. *arXiv preprint arXiv:2506.10964*, 2025.

Maryam Hosseini, Andres Sevtsuk, Fabio Miranda, Roberto M Cesar Jr, and Claudio T Silva. Mapping the walk: A scalable computer vision approach for generating sidewalk network datasets from aerial imagery. *Computers, Environment and Urban Systems*, 101: 101950, 2023.

Yujun Hou, Matias Quintana, Maxim Khomiakov, Winston Yap, Jiani Ouyang, Koichi Ito, Zeyu Wang, Tianhong Zhao, and Filip Biljecki. Global streetscapes – a comprehensive dataset of 10 million street-level images across 688 cities for urban science and analytics. *ISPRS Journal of Photogrammetry and Remote Sensing*, 215:216–238, 2024. doi: 10.1016/j.isprsjprs.2024.06.023.

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.

Paul Jacob, Éloi Zablocki, Hédi Ben-Younes, Mickaël Chen, Patrick Pérez, and Matthieu Cord. Steex: Steering counterfactual explanations with semantics, 2022. URL `https://arxiv.org/abs/2111.09094`.

Chia Hsiang Kao, Wenting Zhao, Shreelekha Revankar, Samuel Speas, Snehal Bhagat, Rajeev Datta, Cheng Perng Phoo, Utkarsh Mall, Carl Vondrick, Kavita Bala, and Bharath Hariharan. Towards llm agents for earth observation: The univearth dataset. *ACL Findings*, 2025.

Huabing Ke, Zhongyuan Li, Zhaohua Liu, and Zhaoliang Zeng. Satellite-ground data fusion for hourly 5-km gridded human-perceived temperature estimation in the yangtze river basin, china. *Remote Sensing*, 17(18):3260, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. URL `https://arxiv.org/abs/2304.02643`.

Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 27831–27840, 2024.

Andrew Larkin, Xiang Gu, Lizhong Chen, and Perry Hystad. Predicting perceptions of the built environment using gis, satellite and street view image approaches. *Landscape and urban planning*, 216:104257, 2021.

Xiaojiang Li, Chuanrong Zhang, Weidong Li, Robert Ricard, Qingyan Meng, and Weixing Zhang. Assessing street-level urban greenery using google street view and a modified green view index. *Urban forestry and urban greening*, 14(3):675–685, 2015.

Yunzhe Liu, Meixu Chen, Meihui Wang, Jing Huang, Fisher Thomas, Kazem Rahimi, and Mohammad Mamouei. An interpretable machine learning framework for measuring urban perceptions from panoramic street view images. *Iscience*, 26(3), 2023.

Prasanta Chandra Mahalanobis. On the generalized distance in statistics. *Sankhyā: The Indian Journal of Statistics, Series A (2008-)*, 80:S1–S7, 2018.

Marco De Nadai, Radu L. Vieriu, Gloria Zen, Stefan Dragicevic, Nikhil Naik, Michele Caraviello, Cesar A. Hidalgo, Nicu Sebe, and Bruno Lepri. Are safer looking neighborhoods more lively? a multimodal investigation into urban life, 2016. URL https://arxiv.org/abs/1608.00462.

Nikhil Naik, Jade Philipoom, Ramesh Raskar, and César Hidalgo. Streetscore-predicting the perceived safety of one million streetscapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 779–785, 2014.

Alameen Najjar, Shun'ichi Kaneko, and Yoshikazu Miyanaga. Combining satellite imagery and open data to map road safety. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Andrea Porcheddu, Ville Kolehmainen, Timo Lähivaara, and Antti Lipponen. Machine learning data fusion for high spatio-temporal resolution pm2. 5. *Atmospheric Measurement Techniques*, 18(18):4771–4789, 2025.

Viraj Prabhu, Sriram Yenamandra, Prithvijit Chattopadhyay, and Judy Hoffman. Lance: Stress-testing visual models by generating language-guided counterfactual images, 2023. URL https://arxiv.org/abs/2305.19164.

Prolific. Prolific. https://www.prolific.com, 2026. Online participant recruitment platform.

Jerome Quenum, Wen-Han Hsieh, Tsung-Han Wu, Ritwik Gupta, Trevor Darrell, and David M. Chan. Lisat: Language-instructed segmentation assistant for satellite imagery. *ArXiv*, abs/2505.02829, 2025. URL https://api.semanticscholar.org/CorpusID:278338943.

G M Atiqur Rahaman, Martin Längkvist, and Amy Loutfi. Deep learning based automated estimation of urban green space index from satellite image: A case study. *Urban Forestry and Urban Greening*, 97:128373, 2024. ISSN 1618-8667. doi: https://doi.org/10.1016/j.ufug.2024.128373. URL https://www.sciencedirect.com/science/article/pii/S1618866724001717.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

Srikumar Sastry, Subash Khanal, Aayush Dhakal, and Nathan Jacobs. Geosynth: Contextually-aware high-resolution satellite image synthesis. In *IEEE/ISPRS Workshop: Large Scale Computer Vision for Remote Sensing (EARTHVISION)*, 2024.

Sarah Schwettmann, Tamar Shaham, Joanna Materzynska, Neil Chowdhury, Shuang Li, Jacob Andreas, David Bau, and Antonio Torralba. Find: A function description benchmark for evaluating interpretability methods. *Advances in Neural Information Processing Systems*, 36:75688–75715, 2023.

Akashah Shabbir, Muhammad Akhtar Munir, Akshay Dudhane, Muhammad Umer Sheikh, Muhammad Haris Khan, Paolo Fraccaro, Juan Bernabe Moreno, Fahad Shahbaz Khan, and Salman Khan. Thinkgeo: Evaluating tool-augmented agents for remote sensing tasks. *arXiv preprint arXiv:2505.23752*, 2025.

Tamar Rott Shaham, Sarah Schwettmann, Franklin Wang, Achyuta Rajaram, Evan Hernandez, Jacob Andreas, and Antonio Torralba. A multimodal automated interpretability agent, 2025. URL https://arxiv.org/abs/2404.14394.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Weicheng Wang, Guoli Jia, Zhongqi Zhang, Liang Lin, and Jufeng Yang. Ps-diffusion: Photorealistic subject-driven image editing with disentangled control and attention. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18302–18312. IEEE Computer Society, 2025.

Chen Wei, Yiwei Zhang, Xi Guan, Qianheng Zhang, Yanbing Chen, Yuhan Ji, Yibo Zhao, Maoping Wang, Ying Nie, Hanchen Zhuang, and Song Gao. Geoai for driving risk assessment via vision-language models: A knowledge guided rag system and dual evaluation. In *Proceedings of the 8th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, GeoAI '25, pp. 161–173, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400721793. doi: 10.1145/3764912.3770832. URL https://doi.org/10.1145/3764912.3770832.

Jennifer R. Wolch, Jason Byrne, and Joshua P. Newell. Urban green space, public health, and environmental justice: The challenge of making cities 'just green enough'. *Landscape and Urban Planning*, 125:234–244, 2014. ISSN 0169-2046. doi: https://doi.org/10.1016/j.landurbplan.2014.01.017. URL https://www.sciencedirect.com/science/article/pii/S0169204614000310.

Junshi Xia, Naoto Yokoya, Bruno Adriano, and Clifford Broni-Bediako. Openearthmap: A benchmark dataset for global high-resolution land cover mapping. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6254–6264, 2023.

Wenjia Xu, Zijian Yu, Boyang Mu, Zhiwei Wei, Yuanben Zhang, Guangzuo Li, Jiuniu Wang, and Mugen Peng. Rs-agent: Automating remote sensing tasks through intelligent agent. *arXiv preprint arXiv:2406.07089*, 2024.

Xujing Yu, Jun Ma, Yihong Tang, Tianren Yang, and Feifeng Jiang. Can we trust our eyes? interpreting the misperception of road safety from street view images and deep learning. *Accident Analysis and Prevention*, 197:107455, 2024. ISSN 0001-4575. doi: https://doi.org/10.1016/j.aap.2023.107455. URL https://www.sciencedirect.com/science/article/pii/S000145752300502X.

Masoud Zaerpour, Simon Michael Papalexiou, and Alain Pietroniro. Increasing tree canopy lowers urban air temperature by up to 1.5° c in heat-prone areas. *npj Urban Sustainability*, 5(1):92, 2025.

Mehdi Zemni, Mickaël Chen, Éloi Zablocki, Hédi Ben-Younes, Patrick Pérez, and Matthieu Cord. Octet: Object-aware counterfactual explanations, 2023. URL https://arxiv.org/abs/2211.12380.

Yang Zhan, Zhitong Xiong, and Yuan Yuan. Skyeyegpt: Unifying remote sensing vision-language tasks via instruction tuning with large language model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 221:64–77, 2025.

Jiaxin Zhang, Yunqin Li, Tomohiro Fukuda, and Bowen Wang. Revolutionizing urban safety perception assessments: Integrating multimodal large language models with street view images. *Available at SSRN 4960800*, 2024.

Chenbo Zhao, Yoshiki Ogawa, Shenglong Chen, Takuya Oki, and Yoshihide Sekimoto. Street space quality improvement: Fusion of subjective perception in street view image generation. *Information Fusion*, 125:103467, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, 2023.

Qianyu Zhou, Jiaxin Zhang, and Zehong Zhu. Evaluating urban visual attractiveness perception using multimodal large language model and street view images. *Buildings*, 15 (16):2970, 2025.

Appendix Contents

A VIDA-Geo Algorithm and Fallbacks **16**

B Details on FID-Proxy **17**

C Human Evaluation Analysis **18**

D Sequential Editing Artifacts **19**

E Using Generative Models + Black Box Indicator Predictions as IDTs **21**

F Implementation and Experimental Setup Details **21**

G Prompts Used Within VIDA-Geo **21**

H Handling of Failure Cases **23**

I Analysis of the Best Output for Each Model **23**

J Analysis of Different LLM Judge **25**

K Benchmark Data Statistics **26**

L More Qualitative Examples **27**

M Comparison with Real-World Interventions **30**

N Ablation Studies Details **30**

O More Details on the User Interface for Human Evaluation **30**

P Broader Impact of this Work **31**

Q Licenses for Images **32**

A VIDA-Geo Algorithm and Fallbacks

Algorithm 1: VIDA-Geo Agent Procedures

Function `Planner`$(I, G, \mathcal{H})$**:**
  $\mathcal{R} \leftarrow \text{RankROIs}(I, G, \mathcal{H})$; `// identify and rank candidate ROIs`
  return $\mathcal{R}$;

Function `PolicyAgent`(I, r, G)**:**
  $C_r \leftarrow \text{AssignConstraint}(I, r, G)$; `// free, constrained, or skip`
  return C_r;

Function `SegmentationAgent`(I, r)**:**
  $M \leftarrow \text{Segment}(I, r)$; `// select SAM, SAM3, or LISA`
  $(q_m, f_m) \leftarrow \text{MaskQC}(I, M, r)$;
  if f_m *indicates fragmentation* **then**
    $M \leftarrow \text{SmoothMask}(M)$; `// smooth fragmented regions`
    $(q_m, f_m) \leftarrow \text{MaskQC}(I, M, r)$;
  while $q_m = \text{fail}$ *and retries remain* **do**
    $P_s \leftarrow \text{PromptSuggestor}(r, f_m)$; `// revise segmentation prompt`
    $M \leftarrow \text{Segment}(I, r, P_s)$; `// revise prompt and/or switch segmenter`
    $(q_m, f_m) \leftarrow \text{MaskQC}(I, M, r)$;
  return M, q_m;

Function `GenerationAgent`(I, M, r, G, C_r)**:**
  $P_g \leftarrow \text{PromptSuggestor}(r, G, C_r)$; `// propose editing prompt`
  repeat
    $I' \leftarrow \text{Generate}(I, M, P_g)$; `// select available generation tool`
    $(q_g, f_g) \leftarrow \text{GenerationQC}(I, I', M, P_g, C_r)$;
    if $q_g = \text{fail}$ **then**
      $P_g \leftarrow \text{PromptSuggestor}(P_g, f_g)$; `// revise editing prompt`
  until $q_g = \text{pass}$ *or retries exhausted*;
  return I', q_g;

The goal of VIDA-Geo is to avoid full-image-level interventions that could lead to artifacts and policy violations shown in fig. 5a. The algorithm shown in algorithm 2 summarizes the overall flow of the VIDA-Geo pipeline that operates on the region-level; including the interaction between planning, policy assignment, segmentation, generation, quality control, and multi-epoch feedback.

In the segmentation stage, failure cases can be addressed by revising the segmentation prompt, switching to an alternative segmentation tool, or applying mask smoothing when fragmentation is detected. In fig. 5b, the intended region of interest is the building on the right, but the SAM3 prompt "mid-rise building" would segment every building in the image. Therefore, using the location point from the Planner Agent with SAM could select the correct building in the image as shown in fig. 6a.

Similarly, generation failures such as those shown in fig. 5c can trigger a revised editing prompt before the intervention is regenerated. These fallback mechanisms allow the system to adapt its strategy without restarting the full pipeline, while preserving the same ROI and intervention objective. In the first example, the pavement originally looked unrealistic and was corrected in the next editing stage. In the bottom example, the first edit was made in the wrong region of the image, outside the masked area. It was rejected and corrected in the second iteration as illustrated in fig. 6b.

Algorithm 2: VIDA-Geo Multi-Agent Geo-Editing Pipeline

Input: Image I, goal G, scorer S
Output: Accepted interventions $\mathcal{O}$

$\mathcal{H} \leftarrow \emptyset$, $\mathcal{O} \leftarrow \emptyset$;
$s_0 \leftarrow S(I)$;
for $e = 1, \ldots, E$ **do**
  $\mathcal{R} \leftarrow \texttt{Planner}(I, G, \mathcal{H})$;
  foreach $r \in \mathcal{R}$ **do**
    $C_r \leftarrow \texttt{PolicyAgent}(I, r, G)$;
    if $C_r = \text{skip}$ **then**
      continue;
    $(M, q_m) \leftarrow \texttt{SegmentationAgent}(I, r)$;
    if $q_m = \text{fail}$ **then**
      continue;
    $(I', q_g) \leftarrow \texttt{GenerationAgent}(I, M, r, G, C_r)$;
    if $q_g = \text{pass}$ **then**
      $\Delta \leftarrow S(I') - s_0$;
      $\mathcal{O} \leftarrow \mathcal{O} \cup \{(r, M, I', \Delta)\}$;
  $\mathcal{H} \leftarrow \texttt{UpdateHistory}(\mathcal{H}, \mathcal{O})$; // inform the next epoch
return $\mathcal{O}$;



(a) Policy violations.



(b) Bad masks.



(c) Bad edits.

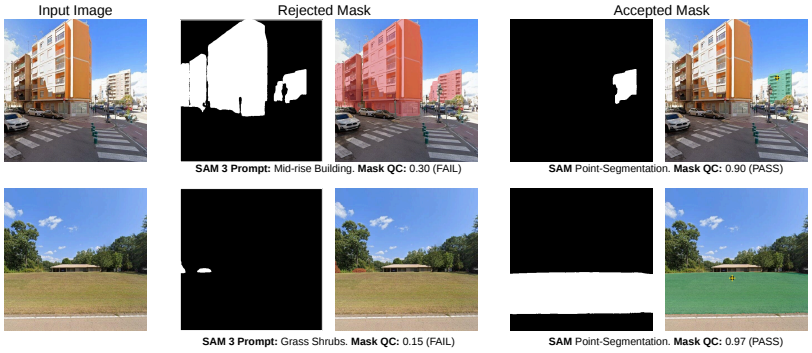
Figure 5: Examples of failure cases. (a) shows policy violations that our system steers away from. (b) shows segmentation masks rejected by the QC sub-agent, with the top image segmenting "medium rise building" and the bottom "grass shrubs". Whereas subfigure (c) illustrates rejected edits, the top edit includes a low-res and unrealistic edit, and the bottom edit inpainting the wrong region. Fallbacks and fixes are explained in section A.

B Details on FID-Proxy

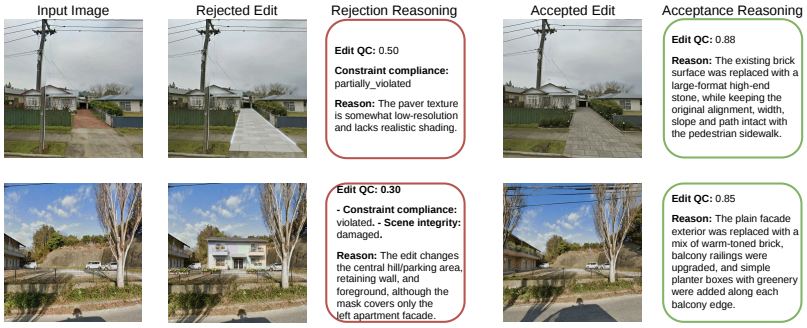
A reference Gaussian $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is estimated from a held-out set of real satellite/street-view images using 2048-dimensional feature vectors extracted from a pretrained InceptionV3 network, with the final classifier replaced by an identity layer. For each generated image x, the domain typicality score is:

$$d(x) = \sqrt{(f(x) - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (f(x) - \boldsymbol{\mu})} \tag{1}$$

where $f(x) \in \mathbb{R}^{2048}$ is the embedded feature vector of image x, and $\boldsymbol{\Sigma}^{-1}$ is computed via Cholesky decomposition with a small regularization term $\epsilon \mathbf{I}$ for numerical stability. Unlike set-level FID, this formulation yields a per-image score, enabling image-level comparison



(a) Workaround example for a rejected mask after QC. Another tool was used to generate the successful mask.



(b) Workaround example for a rejected edit after QC. Prompts were revised to generate better, and more realistic interventions.

Figure 6: Qualitative examples from the Segmentation and Generation Agents outputs, indicating originally failed outputs, that were later corrected to pass the QC.

across individual edits and models. A lower value $d(x)$ indicates that the edited image is more typical of the real aerial or street-view domain.

C Human Evaluation Analysis

C.1 Evaluators Background

This section dives deeper into the analysis of the human evaluation results. The results are broken down by question type and metric. Seven graduate student evaluators participated in the study, each reviewing 30 image pairs comparing outputs from VIDA-Geo and DIFFusion or LANCE separately, for a total of 210 pairwise trials per comparison. Similarly, five users with expertise in Urban Planning and Policy Design from Prolific participated in the study. The left–right placement of the two methods was randomized.

Note that the graduate students were from the university department and were not members of the paper's author team. During recruitment, they were informed about the evaluation and provided verbal consent for their responses to be used in the study. Participation was voluntary, and no names, contact details, or other personal information were collected.

C.2 Human Evaluation Results

Table 2 summarizes the graduate-student evaluation of VIDA-Geo against DIFFusion and LANCE across perceptual attribute alignment, realism, and urban-planning feasibility. The table additionally provides attribute-level results for the perceptual evaluation.

Table 2: Graduate Student preference for VIDA-Geo over DIFFusion and LANCE. Bold rows report the overall result for each evaluation question. For Q1, we additionally report preference by targeted perceptual attribute. Each entry gives the number and percentage of judgments favoring VIDA-Geo.

Evaluation	VIDA-Geo vs. DIFFusion	VIDA-Geo vs. LANCE
Q1 – Metric Preference	**123/210 (58.6%)**	**189/210 (90.0%)**
↪ Lively	28/35 (80.0%)	34/35 (97.1%)
↪ Beautiful	25/35 (71.4%)	34/35 (97.1%)
↪ Less Boring	25/35 (71.4%)	29/35 (82.9%)
↪ Less Depressing	18/35 (51.4%)	32/35 (91.4%)
↪ Safe	17/35 (48.6%)	34/35 (97.1%)
↪ Wealthy	10/35 (28.6%)	26/35 (74.3%)
Q2 – Realism	**201/210 (95.7%)**	**160/210 (76.2%)**
Q3 – Urban Planning	**183/210 (87.1%)**	**161/210 (76.7%)**

Table 3: Expert preference for VIDA-Geo over DIFFusion and LANCE. Bold rows report the overall result for each evaluation question. For Q1, we additionally report preference by targeted perceptual attribute. Each entry gives the number and percentage of judgments favoring VIDA-Geo.

Evaluation	VIDA-Geo vs. DIFFusion	VIDA-Geo vs. LANCE
Q1 – Metric Preference	**92/150 (61.3%)**	**130/150 (86.7%)**
↪ Lively	21/25 (84.0%)	22/25 (88.0%)
↪ Beautiful	18/25 (72.0%)	24/25 (96.0%)
↪ Less Boring	15/25 (60.0%)	19/25 (76.0%)
↪ Less Depressing	10/25 (40.0%)	23/25 (92.0%)
↪ Safe	15/25 (60.0%)	24/25 (96.0%)
↪ Wealthy	13/25 (52.0%)	18/25 (72.0%)
Q2 – Realism	**142/150 (94.7%)**	**100/150 (66.7%)**
Q3 – Urban Planning	**137/150 (91.3%)**	**119/150 (79.3%)**

Expert Evaluation. We further evaluate VIDA-Geo against DIFFusion and LANCE using domain experts. Table 3 summarizes their preferences across the same three evaluation questions, with a detailed breakdown of the perceptual attributes.

Agreement with LLM Judges. We further examine whether human decisions align with the LLM judges' evaluations. Specifically, we use the LLM judges' scores for VIDA-Geo and each baseline to make pairwise decisions analogous to those made by the human evaluators. Table 4 reports agreement with both student and expert majority decisions across DIFFusion and LANCE comparisons. Agreement on realism and urban-planning feasibility is consistently above the 50% random baseline, showing that LLM judges can serve as a scalable proxy for more expensive human evaluation.

D Sequential Editing Artifacts

One concern with sequential editing is the accumulation of artifacts from one iteration to the next, as some acceptable edits may already receive lower LLM-Judge scores after the first iteration and continue to degrade with additional edits. To further analyze this behavior, we iteratively edit 10 inputs per Street View metric and stop when no acceptable edit passes QC or if the edited image reaches the desired goal for the metric (9.0 for increasing metrics and 1.0 for decreasing metrics). We then evaluate the outputs using the LLM-Judge metrics discussed in section 4.2, and the results can be seen in fig. 7. This showcases the trade-off between continual editing and the LLM-as-a-Judge scoring. Across tasks, the first iteration

Table 4: Agreement between LLM judges and human majority preferences. For each image pair and evaluation question, agreement indicates whether the LLM judge selected the same preferred output as the corresponding human majority. Human ties are excluded.

Judge	Human Group	Q1 Metric	Q2 Realism	Q3 Planning
VIDA-Geo vs. DIFFusion				
Qwen3-VL	Students	46.4%	93.1%	89.3%
	Experts	50.0%	96.6%	92.9%
GPT-5.2	Students	48.3%	92.9%	96.6%
	Experts	51.7%	96.4%	100.0%
VIDA-Geo vs. LANCE				
Qwen3-VL	Students	58.6%	77.3%	80.8%
	Experts	65.5%	81.8%	80.8%
GPT-5.2	Students	55.2%	75.0%	85.2%
	Experts	82.8%	79.2%	92.6%

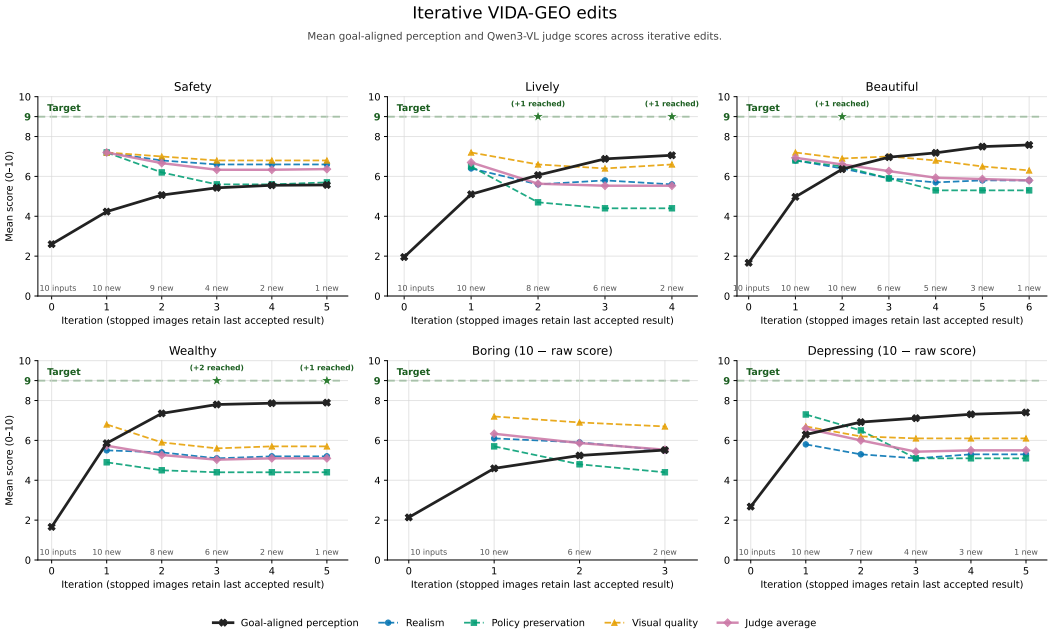


Figure 7: Sequential editing behavior of VIDA-Geo across six Street View perception tasks. Each panel reports the mean goal-aligned perception score together with Qwen3-VL realism, policy-preservation, visual-quality, and average LLM-Judge scores over sequential edits of 10 inputs. For *Boring* and *Depressing*, the raw perception score is inverted (10 − raw score) so that higher values consistently indicate progress toward the target. Editing stops for an input when no acceptable candidate passes QC or when the target is reached; stopped inputs retain their last accepted result in subsequent cohort averages.

produces the largest improvement in the target indicator and generally the highest average LLM-Judge scores, but the following iterations indicate a slight drop in perceptual quality and policy alignment scores. Despite this decline, the average LLM-Judge scores at the final iteration remain above 5.0, suggesting that the resulting edits remain reasonably realistic and policy-aligned.

E Using Generative Models + Black Box Indicator Predictions as IDTs

In section 1, we mention that a model can be considered an *Implicit Digital Twin* if it has the ability to not only make correct predictions on the in-distribution test data similar to the training data but also on counterfactuals that may be out-of-distribution. Here we present brief qualitative evidence that the black-box models we have chosen perform fairly well on counterfactual edits. Figure 8, shows changes to the score from these black box model when certain changes are made. While exact score is hard to interpret, the relative increase or decrease is logical.

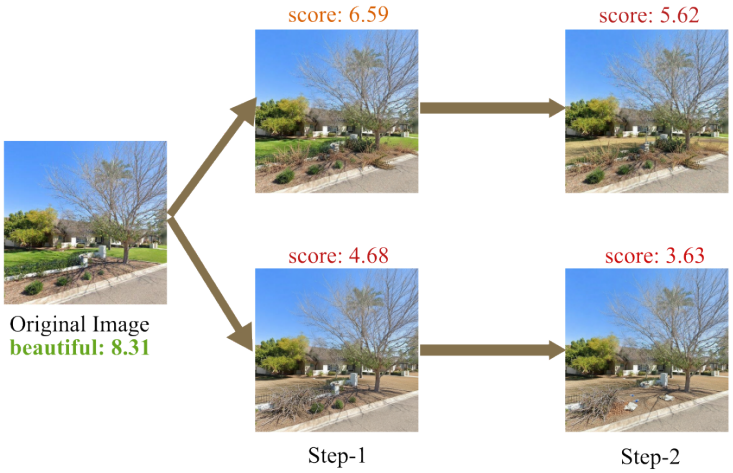


Figure 8: Counterfactual scene edits and corresponding predictor outputs. VIDA-Geo generates alternative versions of the same street-view image through two distinct editing directions. As vegetation is removed and the environment becomes less maintained, the black-box "beauty" predictor assigns progressively lower scores. The consistent and interpretable score changes suggest that the predictor captures meaningful scene attributes under counterfactual interventions, supporting its use as a component of a world model.

F Implementation and Experimental Setup Details

The pipeline integrates four models, FLUX-Fill Black Forest Labs (2024), LISAt Quenum et al. (2025), SAM Kirillov et al. (2023), and SAM3 Carion et al. (2026), which serve as local REST microservices. Region planning and all quality-control agents use **GPT-5.1** Singh et al. (2025). Semantically constrained regions rely on **Gemini 2.5 Flash** Comanici et al. (2025), also referred to as NanoBanana 2.5 (NB2.5), which edits via scene-level instruction rather than masked pixel replacement. All the cloud calls to GPT-5.1 and NB2.5 carry no local GPU requirements. The thresholds used in the quality check steps using the LLM judge were 0.7; the number of trials to mask and edit each region is limited to 3 trials per epoch. We set the steps to 30 and guidance to 20 for FLUX.

All experiments ran on a NVIDIA RTX 5000 Ada GPU with 32GB of memory. The simplest setup for the Agent is to use SAM/SAM3 and the black-box indicators which can occupy up to 10GB of memory. Experiments using the Street-view images, for 3 epochs, and NanoBanana as the generation model, take 6 minutes and 12 seconds on average. Experiments using satellite images, for 3 epochs using NanoBanana, take 13 minutes and 42 seconds on average.

G Prompts Used Within VIDA-Geo

This section reports the prompts and request formats for VIDA-Geo , that affect region selection, policy constraints, image generation, and evaluation. The complete domain-specific prompts are provided in the first link of the supplementary repository.

Text-Segmentation Request Format

LISAt and SAM3 do not receive a separate natural-language system prompt. Instead, they receive the planner-generated `seg_keyword`: a concrete one- or two-word noun describing the target region, such as `road`, `parking lot`, `street lamp`, or `tree canopy`. The SAM fallback receives only point coordinates and foreground/background labels.

Policy-Agent Decision Format

For each proposed region, the Policy Agent receives its identifier, label, description, edit type, and target metric. It assigns one of four edit classes:

- `free`: the proposed edit is physically plausible and may be applied directly;
- `constrained`: the region may be edited, but its physical or functional role must be preserved;
- `skip`: the proposed intervention is physically implausible or incompatible with the scene.

For constrained regions, the agent must provide an explicit `strategy_prompt_hint` describing what must remain intact. Every proposed region must appear in the structured output.

NanoBanana Edit Request Format

Masked editing. The image editor receives the original image and a binary mask, where white pixels denote the only region that may be changed.

Task: `{suggested_edit_prompt}`

- Modify only the white mask region.
- Preserve all pixels outside the white region.

Maskless fallback. If masked editing fails, the editor receives the original image and the same edit instruction, together with the following constraints:

- Keep all unrelated scene content unchanged.
- Blend the modification naturally with the existing scene.
- Do not add text, labels, watermarks, or borders.

For LLM-as-a-Judge evaluation, the LLM-Judge receives the original image first and the edited image second. All three criteria are scored jointly in one request.

LLM-as-a-Judge Prompt

You are a strict and consistent judge of an image edit. Be critical and discerning.
You receive two images: first the original/reference image and then the edited/output image. Judge only the edited image, using the original image to identify what changed. Score each of the following criteria on an integer scale from 1 to 10. Use the full range, reserve 9–10 for nearly flawless results, and assign low scores when substantial failures are present.

1. **realism**: Assess whether the edited world is physically, semantically, and contextually plausible in the specific scene. Objects should have plausible function, material, scale, geometry, attachment, and spatial relationships. Roads, sidewalks, doors, windows, wires, roofs, trees, parking areas, and buildings should fit the surrounding structure and land use. Penalize edits that insert context-breaking or excessive features, such as objects pasted into an implausible setting, repeated objects that would not actually exist there, or changes that conflict with the overall building, street, or neighborhood context.
2. **policy_preservation**: Assess whether the edited image preserves the original scene's core structure and function. Compare against the reference image. Penalize removal or severe distortion of roads, buildings, sidewalks, vehicles, street layout, vegetation layout, utility poles/wires, doors/windows, building envelopes, or essential infrastructure. Reward edits that keep the original scene recognizable and operational while changing only the intended local appearance.
3. **visual_quality**: Assess whether the edited image looks like a natural, authentic image in the same capture modality as the original. For street-view images, it should look like a real GSV/photo capture, not a cartoon, CGI render, painting, collage, or oddly formatted/generated image. Penalize synthetic style, over-smoothed AI texture, strange color processing, low sharpness, seams, warping, ghosting, garbled textures, melted objects, obvious compositing artifacts, or any visual presentation that makes the image itself feel unnatural. This criterion evaluates image naturalness and artifact quality, not whether the edited world is a good intervention.

Judge each criterion holistically. Return only a JSON object containing exactly the following integer fields and no additional text:

`{"realism":` s_r`, "policy_preservation":` s_p`, "visual_quality":` s_v`}`

where s_r, s_p, and s_v are integers from 1 to 10.

The reported LLM Judge Average is the arithmetic mean of the realism, policy-preservation, and visual-quality scores.

H Handling of Failure Cases

Inputs are not excluded when VIDA-Geo produces no accepted edit. This occurs when the original image already has a relatively optimal indicator score, or when no candidate edit satisfies the acceptance criteria; therefore, VIDA-Geo returns the original input image and flags it as "No Successful Edit." The input still remains part of the benchmark rather than being removed from the evaluation. A representation of the failure rate using the Street View images is as follows in table 5. This showcases an error rate of 1.7% for VIDA-Geo's pipeline.

I Analysis of the Best Output for Each Model

In order to make a fairer comparison between the outputs of the models, a smaller sample of outputs is selected to represent the best single output image for each model for each task. The results are shown in table 6, which summarizes the perceptual quality and policy-alignment results for every "Best" output for each input image. It is clear that VIDA-Geo still

Table 5: Edit success and failure counts by metric. Each metric contains 100 evaluated inputs.

Metric	Successful	No Improving Edit	Failure per 100
Safety	96	4	4%
Lively	100	0	0%
Beautiful	98	2	2%
Wealthy	99	1	1%
Boring	99	1	1%
Depressing	98	2	2%
Total	**590**	**10**	**1.7%**

outperforms LANCE and DIFFusion even when the best selected image is isolated and separately evaluated.

Table 6: Human-aligned quality, intervention strength, and distributional realism proxy for the ***Best*** edited satellite and Street View images across urban intervention tasks, evaluated by **GPT-5.2**. Tasks shown in green correspond to indicators where higher scores are desirable, while tasks shown in red correspond to indicators where lower scores are desirable. Bold judge scores indicate that the best-performing method is statistically separated from all alternatives by non-overlapping 95% confidence intervals.

Task	Method	Output Edit Avg. (%)	Delta Avg. (%)	Perceptual Quality		Policy Alignment		LLM Judge Avg. (%) ± 2.42
				FID-Proxy (↓) ± 1.97	Visual Quality (%) ± 2.48	Realism (%) ± 2.73	Policy Pres. (%) ± 3.45	
Greenery	**VIDA-Geo**	44.6	15.3	45.6	68.3	59.2	46.1	57.9
	LANCE	36.5	7.1	57.5	37.3	27.4	14.6	26.4
	DIFFusion	77.8	48.5	58.3	21.8	22.1	27.2	23.7
	NB2.5 (ZS)	56.3	26.9	47.7	65.0	57.3	51.0	57.8
Road Risk	**VIDA-Geo**	70.6	15.5	49.6	69.7	61.5	45.8	59.0
	LANCE	38.8	47.3	61.9	30.0	27.5	20.4	26.0
	DIFFusion	0.0	86.0	63.2	19.0	17.2	13.0	16.4
	NB2.5 (ZS)	63.7	22.4	50.7	65.5	59.1	51.3	58.6
Safety	**VIDA-Geo**	35.1	17.9	57.2	**66.0**	**70.5**	**66.6**	**67.7**
	LANCE	56.1	38.9	62.0	53.6	49.8	26.2	43.2
	DIFFusion	78.1	60.9	67.9	27.7	35.8	31.7	31.7
	URSimulator	50.2	33.0	57.4	57.6	49.7	40.8	49.4
	NB2.5 (ZS)	68.9	51.8	62.3	55.8	54.0	30.2	46.7
Lively	**VIDA-Geo**	49.2	35.6	62.6	58.6	**56.7**	**52.6**	**55.9**
	LANCE	40.2	26.6	63.8	56.8	47.8	23.6	42.7
	DIFFusion	66.0	52.4	65.0	32.5	39.1	32.3	34.6
	URSimulator	49.8	36.2	59.1	50.5	40.1	32.2	40.9
	NB2.5 (ZS)	79.4	65.7	67.9	49.6	48.1	36.5	44.7
Beautiful	**VIDA-Geo**	45.4	27.7	55.7	**63.9**	68.6	**63.1**	**65.2**
	LANCE	63.7	45.9	58.8	50.3	45.7	25.3	40.4
	DIFFusion	81.4	63.7	67.4	25.2	30.6	29.2	28.3
	URSimulator	48.1	30.3	55.2	55.5	51.5	40.8	49.3
	NB2.5 (ZS)	70.9	53.2	62.7	51.4	63.7	48.7	54.6
Wealthy	**VIDA-Geo**	48.0	30.1	59.0	**64.8**	**65.5**	**52.5**	**60.9**
	LANCE	54.7	36.8	61.5	55.6	50.0	28.0	44.5
	DIFFusion	77.9	60.0	71.4	24.7	32.5	29.0	28.7
	URSimulator	54.3	36.4	58.9	55.8	45.0	34.3	45.0
	NB2.5 (ZS)	87.2	69.2	63.6	60.7	34.3	13.0	36.0
Boring	**VIDA-Geo**	57.3	23.3	58.9	**65.6**	**70.6**	**64.3**	**66.8**
	LANCE	42.5	38.1	59.2	52.3	49.4	27.8	43.2
	DIFFusion	27.7	52.8	61.7	28.1	36.0	31.0	31.7
	URSimulator	44.7	35.9	60.9	51.6	41.2	32.2	41.7
	NB2.5 (ZS)	31.4	49.1	73.5	40.1	28.7	28.6	32.5
Depressing	**VIDA-Geo**	58.9	21.2	59.2	**64.1**	**69.7**	**64.4**	**66.0**
	LANCE	36.2	43.9	61.8	49.4	44.2	23.9	39.2
	DIFFusion	27.9	52.2	73.0	25.6	32.4	28.3	28.8
	URSimulator	54.6	25.5	59.6	53.9	46.5	37.1	45.8
	NB2.5 (ZS)	32.8	47.3	65.4	52.8	53.6	36.0	47.5

Table 7: Perceptual quality and policy alignment for edited satellite and Street View images. Green tasks correspond to indicators where higher scores are desirable, while red tasks correspond to indicators where lower scores are desirable. Visual Quality, Realism, and Policy Preservation are evaluated using **Qwen3-VL 32B**, and LLM Judge Avg. is their mean. All columns report average 95% confidence-interval half-widths. Bold values indicate the best-performing method for each task and metric.

Task	Method	**Visual Quality** (%) ± 0.67	**Realism** (%) ± 0.86	**Policy Pres.** (%) ± 1.15	**LLM Judge Avg.** (%) ± 0.76
Greenery	**VIDA-Geo**	**80.0**	68.7	69.1	72.6
	LANCE	41.5	37.3	23.8	34.2
	DIFFusion	28.1	25.2	30.0	27.7
	NB2.5 (ZS)	78.1	**70.2**	**77.7**	**75.3**
Road Risk	**VIDA-Geo**	**80.5**	**71.5**	65.6	**72.6**
	LANCE	45.1	43.4	26.9	38.5
	DIFFusion	25.3	23.3	29.1	25.9
	NB2.5 (ZS)	69.0	62.2	**66.6**	65.9
Safety	**VIDA-Geo**	**78.7**	**77.3**	**75.8**	**77.3**
	LANCE	62.5	60.8	40.5	54.6
	DIFFusion	30.8	33.8	33.5	32.7
	URSimulator	60.7	49.7	39.3	49.9
	NB2.5 (ZS)	65.6	65.3	45.8	58.9
Lively	**VIDA-Geo**	**71.8**	**66.7**	**65.8**	**68.1**
	LANCE	65.3	62.5	39.4	55.7
	DIFFusion	38.0	40.5	37.8	38.8
	URSimulator	54.7	43.8	34.9	44.4
	NB2.5 (ZS)	49.4	50.9	41.5	47.2
Beautiful	**VIDA-Geo**	**76.4**	**74.8**	**74.9**	**75.3**
	LANCE	58.0	55.6	35.8	49.8
	DIFFusion	27.5	28.8	29.1	28.5
	URSimulator	62.1	53.2	38.9	51.4
	NB2.5 (ZS)	50.3	61.0	56.0	55.8
Wealthy	**VIDA-Geo**	**74.9**	**70.8**	**66.1**	**70.6**
	LANCE	62.9	60.5	39.4	54.3
	DIFFusion	32.7	36.0	36.0	34.9
	URSimulator	56.7	45.3	35.2	45.7
	NB2.5 (ZS)	53.7	34.4	18.3	35.4
Boring	**VIDA-Geo**	**75.8**	**71.8**	**70.3**	**72.6**
	LANCE	63.4	59.6	38.0	53.7
	DIFFusion	38.0	41.0	37.0	38.6
	URSimulator	54.3	42.7	32.7	43.2
	NB2.5 (ZS)	36.2	32.3	29.1	32.5
Depressing	**VIDA-Geo**	**75.4**	**72.4**	**74.1**	**73.9**
	LANCE	61.2	57.2	35.7	51.4
	DIFFusion	32.3	36.2	35.1	34.5
	URSimulator	58.0	47.7	36.8	47.5
	NB2.5 (ZS)	55.4	56.7	43.8	52.0

J Analysis of Different LLM Judge

Table 7 illustrates the results of choosing a different LLM model, Qwen3 VL (qwen3-vl-32b-instruct), showcasing the consistent results for VIDA-Geo across the perceptual quality and policy alignment evaluation metrics. Similar to section I, the results for every "Best" output for each input image using the Qwen3 VL judge are summarized in table 8.

Table 8: Human-aligned quality and distributional realism proxy for the ***Best*** edited satellite and Street View images across urban intervention tasks, evaluated by **Qwen3-VL 32B**. Tasks shown in green correspond to indicators where higher scores are desirable, while tasks shown in red correspond to indicators where lower scores are desirable. FID-Proxy and judge-score columns report average 95% confidence-interval half-widths. Bold judge scores indicate that the best-performing method is statistically separated from all alternatives by non-overlapping 95% confidence intervals.

Task	Method	Perceptual Quality		Policy Alignment		LLM Judge
		FID-Proxy ↓ ± 1.97	**Visual Quality** (%) ± 2.77	**Realism** (%) ± 3.01	**Policy Pres.** (%) ± 3.19	**Avg.** (%) ± 2.50
Greenery	**VIDA-Geo**	45.6	**75.6**	59.0	51.5	62.0
	LANCE	57.5	40.3	36.4	22.4	33.0
	DIFFusion	58.3	22.3	22.3	30.7	25.1
	NB2.5 (ZS)	47.7	68.0	58.4	**70.3**	65.6
Road Risk	**VIDA-Geo**	49.6	**77.7**	66.1	54.2	66.0
	LANCE	61.9	38.0	40.5	27.4	35.3
	DIFFusion	63.2	19.2	12.3	15.9	15.8
	NB2.5 (ZS)	50.7	65.7	60.8	61.8	62.8
Safety	**VIDA-Geo**	57.2	**76.3**	**73.0**	**72.1**	**73.8**
	LANCE	62.0	57.2	52.9	33.4	47.8
	DIFFusion	67.9	30.6	33.5	30.9	31.7
	URSimulator	57.4	58.5	46.2	38.6	47.8
	NB2.5 (ZS)	62.3	56.5	58.8	35.8	50.4
Lively	**VIDA-Geo**	62.6	**68.7**	**63.3**	**52.4**	**61.5**
	LANCE	63.8	62.7	56.9	34.7	51.4
	DIFFusion	65.0	36.0	37.3	32.4	35.2
	URSimulator	59.1	51.4	39.4	31.1	40.6
	NB2.5 (ZS)	67.9	43.5	45.3	32.1	40.3
Beautiful	**VIDA-Geo**	55.7	**72.1**	**66.8**	**64.7**	**67.9**
	LANCE	58.8	52.9	51.1	32.2	45.4
	DIFFusion	67.4	27.5	29.3	29.8	28.9
	URSimulator	55.2	60.9	51.1	36.3	49.4
	NB2.5 (ZS)	62.7	45.7	56.6	49.4	50.6
Wealthy	**VIDA-Geo**	59.0	**70.9**	**63.4**	**54.9**	**63.1**
	LANCE	61.5	57.3	54.3	37.4	49.7
	DIFFusion	71.4	26.7	30.0	29.5	28.7
	URSimulator	58.9	55.4	43.6	33.3	44.1
	NB2.5 (ZS)	63.6	53.1	31.0	15.2	33.1
Boring	**VIDA-Geo**	58.9	**74.1**	**69.6**	**66.0**	**69.9**
	LANCE	59.2	55.4	55.0	36.3	48.9
	DIFFusion	61.7	31.2	34.4	30.9	32.2
	URSimulator	60.9	53.7	41.3	31.8	42.3
	NB2.5 (ZS)	73.5	32.5	26.1	25.7	28.1
Depressing	**VIDA-Geo**	59.2	**71.2**	**65.6**	**66.8**	**67.9**
	LANCE	61.8	53.4	52.5	31.9	45.9
	DIFFusion	73.0	28.2	32.3	31.5	30.7
	URSimulator	59.6	55.8	43.9	34.8	44.8
	NB2.5 (ZS)	65.4	47.2	50.2	35.0	44.1

K Benchmark Data Statistics

In this section, we provide additional statistics about the size and diversity of the dataset introduced in section 4.1. The selection process is an automated procedure that excludes cases where meaningful intervention would be infeasible, such as satellite images of an extremely dense urban area or street view images of surrounding cars on a road. The geographic distribution of the expanded benchmark is: North America (43%), Europe (18%), Asia (17%), South America (14%), Oceania (8%). The score distributions for the selected images are summarized in table 9 (in %).

Table 9: Summary statistics of indicator scores.

Indicator	Median Score	Score Range
Safety	13.0	3.0–49.0
Lively	11.0	2.0–43.0
Beautiful	15.0	2.0–47.0
Wealthy	14.0	2.3–48.0
Boring	83.0	54.0–98.0
Depressing	83.0	55.0–98.0
Greenery	29.0	5.0–64.0
Road Risk	90.0	51.3–99.7

For the satellite-image indicators, we measured the average semantic composition of the images: Urban: 57.6%, Vegetated: 41.1%, Road: 14.6%, Building: 18.9%. Across all Street View images, the average visual composition is: sky (39.9%), buildings (17.4%), vegetation (15.5%), road (14.1%), open ground (4.5%), and sidewalk or path (3.3%). These statistics provide a quantitative characterization of visual diversity in the benchmark.

L More Qualitative Examples

Table 10 shows a few more comparisons of VIDA-Geo vs. DIFFusion, specifically on aerial image indicators. While DIFFusion just increases the green saturation of the image for both increasing road safety and greenery, VIDA-Geo provides more realistic and policy-wise practical changes.

Table 11 shows an example of a single forward execution of a segmentation agent and generative agent. For the first row in greenery, the segmentation agent suggests a reason to improve. Then the planner suggests a prompt to edit said region with more trees. Finally, the generator agent calls FLUX model and fills the empty region with mixed species of trees. The black-box score increased by 7% with this regional intervention.

Table 10: Qualitative comparison across selected intervention examples. Scores beneath each generated image are reported in the order LLM-Judge, FID-proxy, and goal gain. Red task labels denote minimization objectives and green task labels denote maximization objectives.

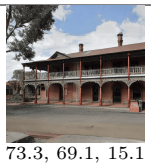
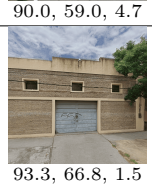
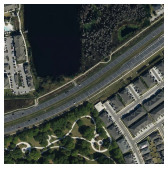
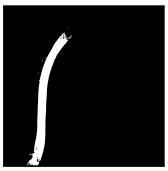
Task	Input	VIDA-Geo	DIFFusion	LANCE	NB2.5	URSimulator
		LLM-Judge ↑, FID ↓, Gain ↑				
Safety		 86.7, 52.5, 11.0	 86.7, 48.0, 33.8	 26.7, 58.1, 45.3	 83.3, 50.8, 17.2	 60.0, 53.0, 23.9
Safety		 83.3, 57.1, 31.6	 73.3, 62.8, 12.9	 73.3, 69.1, 15.1	 70.0, 73.0, 12.2	 63.3, 59.6, 25.3
Lively		 76.7, 50.7, 37.3	 90.0, 59.0, 4.7	 60.0, 60.3, 3.0	 66.7, 62.9, 15.5	 43.3, 56.2, 21.2
Lively		 80.0, 81.5, 13.1	 93.3, 66.8, 1.5	 50.0, 50.2, 8.8	 53.3, 87.6, 68.8	 40.0, 59.2, 62.4
Lively		 73.3, 73.9, 47.5	 86.7, 69.6, 14.3	 50.0, 74.1, 30.1	 63.3, 72.8, 66.3	 50.0, 62.7, 11.8
Wealthy		 73.3, 69.2, 45.3	 90.0, 69.1, 7.5	 43.3, 65.7, 40.9	 50.0, 77.3, 44.2	 50.0, 82.6, 71.9

Table 10 continued

Task	Input	VIDA-Geo	DIFFusion	LANCE	NB2.5	URSimulator
		LLM-Judge ↑, FID ↓, Gain ↑				
Depressing		 76.7, 67.7, 60.9	 83.3, 69.5, 35.9	 63.3, 65.0, 31.3	 70.0, 87.0, 48.9	 40.0, 62.8, 18.7
Depressing		 80.0, 61.7, 25.2	 63.3, 65.6, 28.7	 46.7, 71.3, 43.8	 76.7, 58.3, 13.6	 33.3, 70.5, 38.5
Boring		 80.0, 46.3, 26.9	 80.0, 47.1, 15.9	 46.7, 60.1, 32.9	 56.7, 60.5, 35.5	 63.3, 48.2, 7.5
Greenery		 90.0, 45.8, 4.8	 86.7, 43.0, 5.5	 46.7, 52.6, -12.5	 83.3, 40.3, 8.7	*N/A*
Road Risk		 80.0, 47.7, 16.7	 73.3, 46.6, 10.9	 36.7, 82.8, 2.4	 66.7, 51.2, 27.8	*N/A*
Road Risk		 90.0, 37.3, 42.9	 80.0, 35.0, 8.7	 23.3, 57.4, 34.1	 86.7, 38.8, 74.1	*N/A*

Table 11: Qualitative examples of counterfactual edits and indicator changes.

Task	Input	Seg. Mask	Gen. Prompt	Gen. Edit	Score
Greenery		 Seg: LISAt	*"Introduce a dense mixed-species tree canopy with distinct rounded crown shapes and natural shadows."*	 Gen: FLUX	+7.1%
		 Seg: SAM3	*"Convert the masked roadside area into a linear park with a dense tree canopy, walking paths, and seating areas."*	 Gen: NanoBanana	+5.5%
Road Risk		 Seg: SAM	*"Transform the masked bridge into a pedestrian and cycle bridge with barriers separating it from vehicle lanes."*	 Gen: NanoBanana	−9.0%

M Comparison with Real-World Interventions

To examine whether VIDA-Geo proposes interventions that resemble changes already implemented in practice, we qualitatively compare its outputs with real-world interventions. Figure 9 presents examples of real before-and-after changes alongside interventions generated by VIDA-Geo for similar urban settings. We observe that VIDA-Geo can generate several similar changes that target increases in beautifulness and safety. For example, improving the sidewalks, adding a bike-lane, incorporating a pedestrian-crossing line and improving the theater marquee.

N Ablation Studies Details

Ablation analysis was completed by disabling some of VIDA-Geo's sub-agents, using a subset of 10 images sampled uniformly from each task. The analysis particularly focused on the Policy Restrictions (restrictor) module, the Prompt Suggestion (suggestor) sub-agent, and the Generation QC and Segmentation QC sub-agents. Around 38.3% of the outputs of the models with disabled agents do not produce an image that improves upon the baseline score. Additionally, disabling the QC sub-agents causes a significant drop in scores from 77.7% to 67.1% for the Generation Edit QC and 64.2% for the Segmentation Mask QC in the average LLM-as-a-Judge score, as detailed in table 12.

O More Details on the User Interface for Human Evaluation

Here we provide more details on the instructions and the user interface (UI) for the human evaluation. Figure 10, shows the list of instructions given to the human evaluators. Since diffusion outputs look very unrealistic, to prevent them from biasing always select VIDA-Geo , we intentionally did not give any examples or explained what we mean exactly by indicator names such as "safety" or "wealthiness." Figure 11 shows an example of a UI where the users have to answer three questions about the pair of images: Which image do they

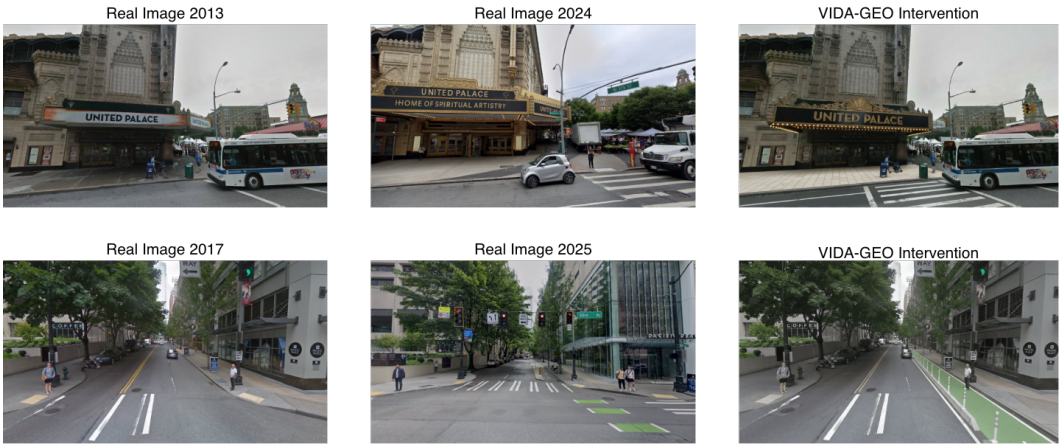


Figure 9: Qualitative comparison of VIDA-Geo interventions with real-world changes. From left to right: the original scene, the observed real-world intervention, and the intervention generated by VIDA-Geo. In the top example, VIDA-Geo similarly improves the pedestrian crossing, sidewalk, and theater marquee; in the bottom example, it introduces a bike lane similar to the one observed in the real-world intervention.

Table 12: Ablation study results. Each row disables one pipeline module (marked with ✗) while keeping all others active (✓). *Visual Quality*, *Realism*, and *Policy Preservation* are scored by an LLM judge; *Judge Avg.* is their mean. *FID-Proxy* measures domain typicality (lower is better).

Configuration	Edit QC	Policy	Mask QC	Suggestor	FID-Proxy ↓	Vis. Qual. (%)	Realism (%)	Policy Pres. (%)	Judge Avg. (%)
Full Pipeline (VIDA-Geo)	✓	✓	✓	✓	**54.13**	**71.9**	**75.6**	**85.5**	**77.7**
w/o Edit QC	✗	✓	✓	✓	57.05	65.0	68.8	67.5	67.1
w/o Policy Restrictor	✓	✗	✓	✓	56.09	68.0	64.6	62.5	65.0
w/o Mask QC	✓	✓	✗	✓	56.79	62.5	60.0	70.1	64.2
w/o Suggestor	✓	✓	✓	✗	55.87	60.0	65.5	80.0	68.5

believe has a higher change in the indicator score? Which image looks perceptually more real? And changes in which image align better with policy?

P Broader Impact of this Work

VIDA-Geo can support urban planning by making proposed interventions easier to visualize, compare, and prioritize before costly real-world deployment. It is the first of its kind to help in decision making to support engineers and urban planners in planning and modifying regions with multiple indicators in mind. At the same time, the system should not be treated as a replacement for community engagement, expert review, or detailed engineering studies. Because the generated edits and indicator scores depend on pretrained models, they may inherit geographic, socioeconomic, or cultural biases, and visually plausible interventions may still be legally, financially, or physically infeasible. There is also a risk that such tools could be used to optimize for superficial visual improvements rather than deeper structural needs.

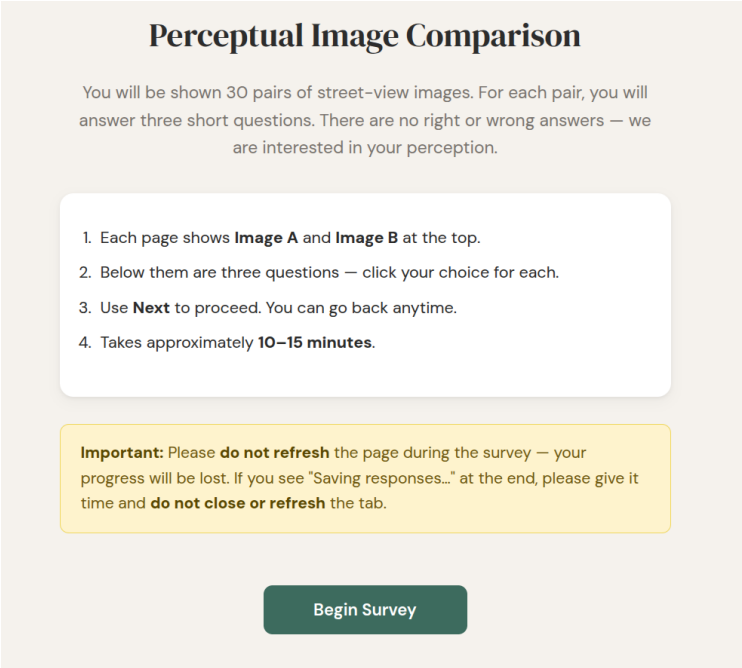


Figure 10: Instructions given to the human evaluators. The instructions were intentionally kept simple to prevent users from biasing themselves to select our method.

Q Licenses for Images

We do not redistribute any Google Street View, MapBox or third-party aerial imagery. The released benchmark contains only independently curated coordinates for each of our tasks. Users may obtain imagery from sources whose licenses permit their intended use.

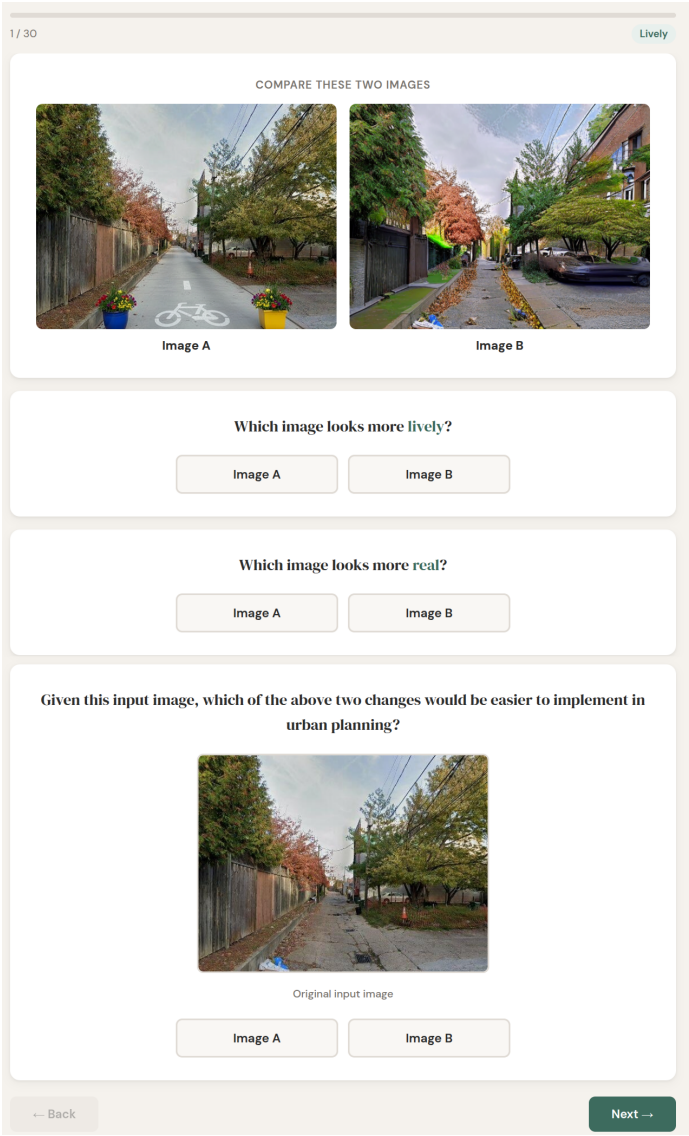


Figure 11: User interface with questions shown to the human annotators. They are randomly shown the best-performing DIFFusion intervention and the best VIDA-Geo intervention, and asked three questions about the indicator, perceptual quality, and policy alignment.