ECTQG 2025

24th European Colloquium on Theoretical and Quantitative Geography

Novel Spatial Data and Indicators for Assessing the Reality of 15-Minute Cities

Assessing urban scenes for the 15-minute city through SAGAI (Streetscape Analysis with Generative AI)

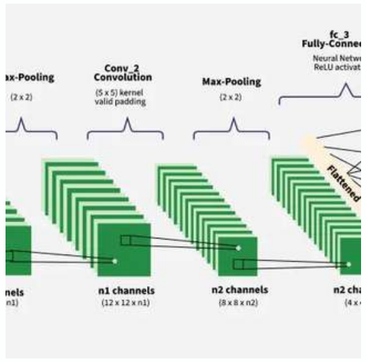
Perez, Joan - UrbanGeoAnalytics
Fusco, Giovanni - UMR 7300 ESPACE-CNRS

INTRODUCTION

The 15mC as a pedestrian experience

15mC has to cater to pedestrians, and pedestrians need appropriate streetscapes.
A **streetscape** refers to the **visual and functional characteristics of a street** and its immediate surroundings, as experienced by people in public space.
EMC2: Good Streetscapes → Essential factor of the pedestrian 15mC
Problem: streetscape **qualities** from the pedestrian point of view **hard to assess**

Steetscape example



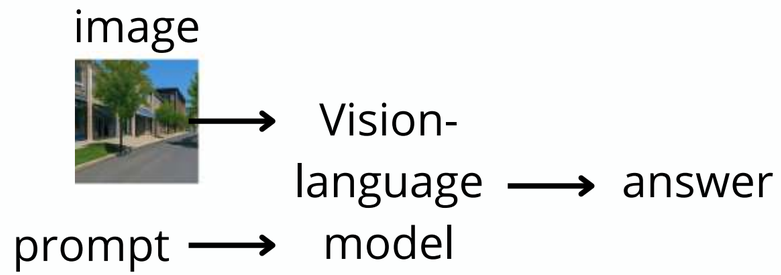
Field surveys & detection through CNNs

Traditional Approaches

Assessed through field surveys, audits, and **manual photo annotations** → costly, inconsistent, small-scale.
Later advances used Convolutional Neural Networks (CNNs) to **automatically classify and detect visual features**, but required large labeled datasets and were often **limited in scope** (task specific).

Towards Automation with GenAI

Vision–language models (coupling of vision encoder with LLMs) enable semantic interpretation of streetscapes directly from images through natural language.
Allow **scalable, reproducible, and low-cost detection** of streetscape elements and qualities.



Feature extraction using GenAI

VISION-LANGUAGE MODELS

Vision Tower : The Eye

- Takes an image and turns it into a an **embedding** (numerical vector)
- These embeddings are learned during **pretraining** on large datasets
- Examples: CLIP Vision Encoder, ViT (Vision Transformer), ResNet.

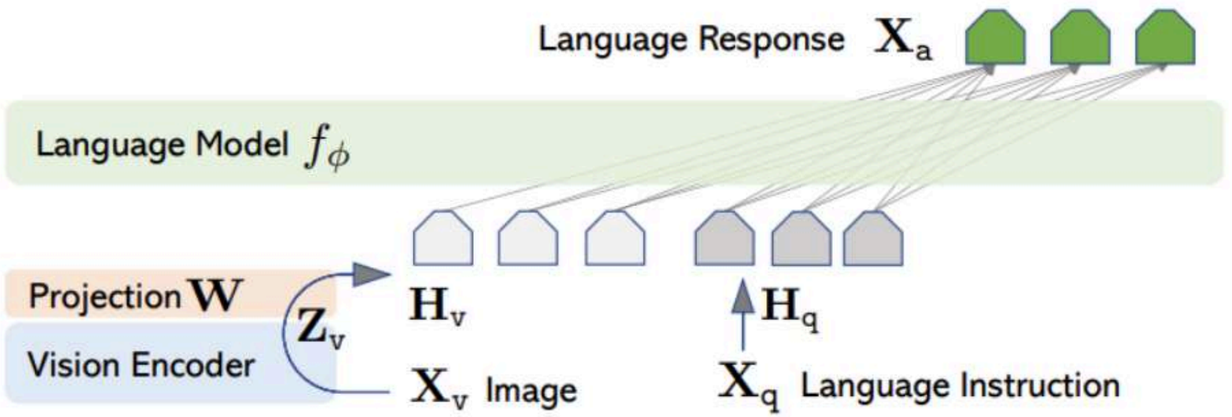
Large Language Model (LLM) : The Brain

- The LLM interprets inputs (prompts) in **natural language** and integrates contextual knowledge.
- Produces outputs (answers) in human-readable form or structured formats
- Examples: LLaMA, GPT-style transformers, Mistral, Falcon.

Vision-Language : The Combination

+

- Vision + Language: fusion enables semantic interpretation of images beyond object detection
- Supports flexible, task-specific queries without retraining (zero-shot capability)
- Examples: BLIP-2, LLaVA, GPT-4V, Kosmos-2

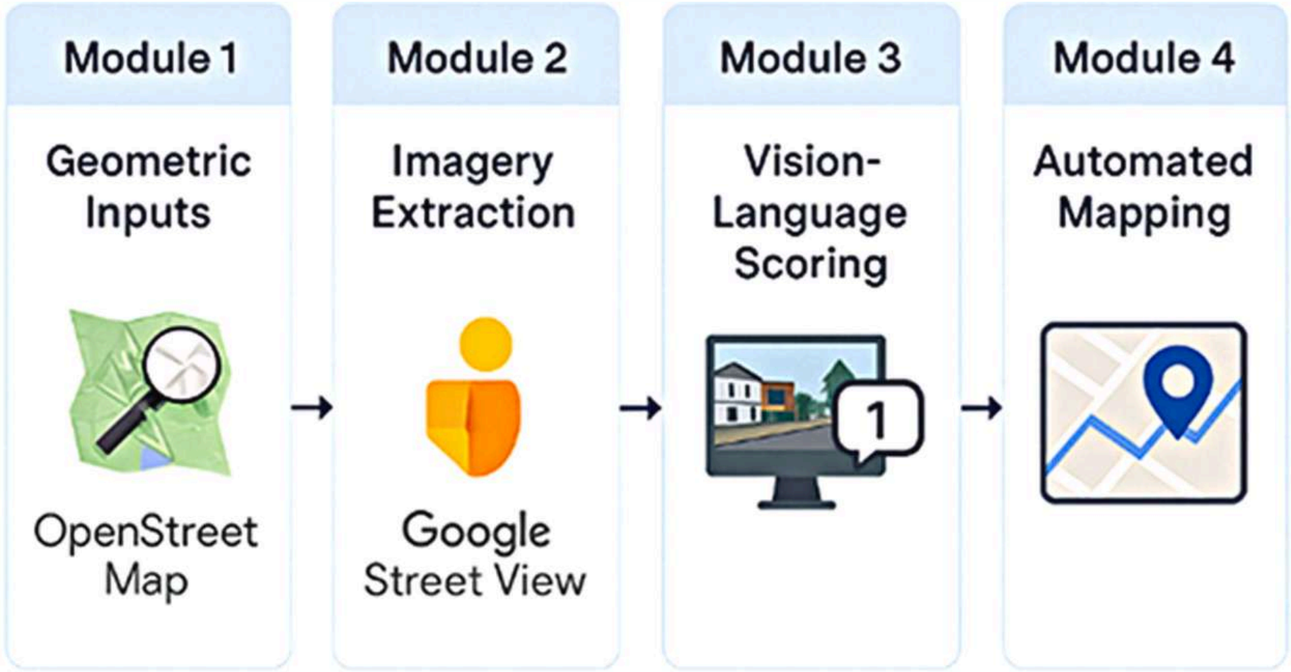


LLaVA algorithm (Liu et al., 2023))

SAGAI : STREETSCAPE ANALYSIS WITH GENERATIVE AI

The Starting Point

- If we can **automate the retrieval of street-level imagery** (e.g., from Google Street View) while retaining the spatial coordinates,
- Then we can use **vision–language models to extract key features of the streetscape** — sidewalks, storefronts, greenery, crossings — in a reproducible and scalable way.
- And finally, we can transform these outputs into **automated maps**, providing fine-grained urban indicators across entire neigborhoods or cities.



The four modules of SAGAI

This idea led to the design of SAGAI, structured into four modules:

- Geometric Inputs from OpenStreetMap,
- Imagery Extraction from Google Street View,
- Vision–Language Scoring of urban elements,
- Automated Mapping of results at scale.

MODULE 1: OSM POINTS GENERATOR

Purpose & Inputs

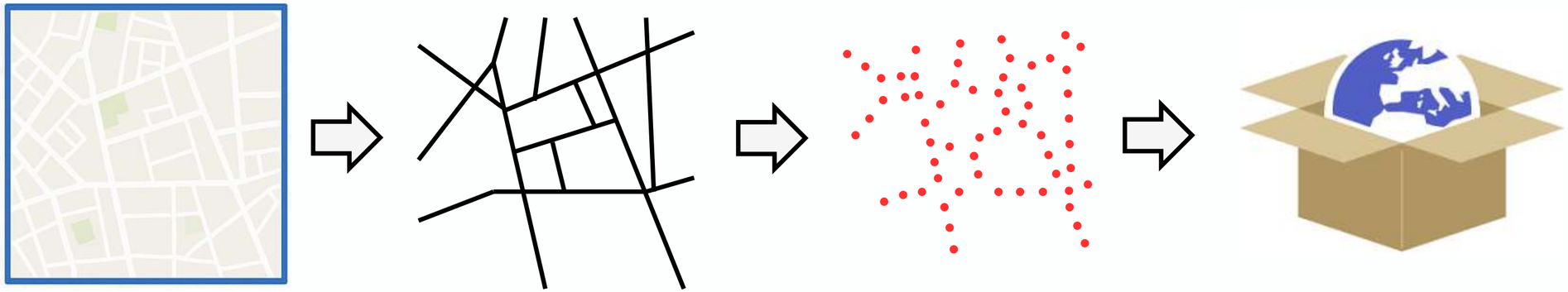
The system automatically **downloads the street geometries** from OpenStreetMap.
Only a bounding box (no heavy data preparation needed).

Parameters

- Spacing → distance between points.
- Offset → prevents points from sitting directly on intersections.
- Each point is linked to its street segment with a unique identifier, ensuring that later results can be aggregated at the street level.

Outputs

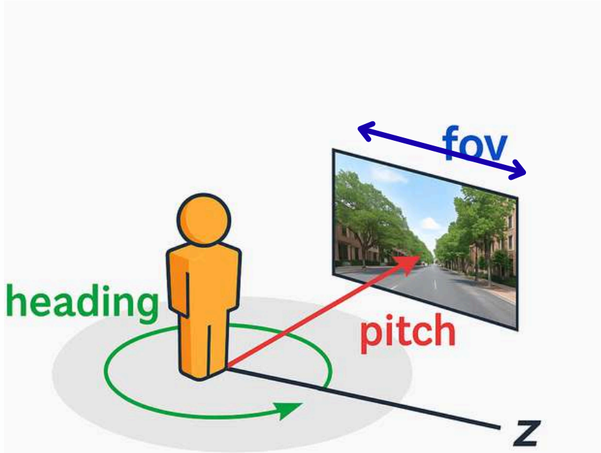
GeoPackage with both the cleaned street network and the generated points, ready to feed module 2.



MODULE 2: STREET-LEVEL IMAGES BATCH DOWNLOADER

Purpose & Inputs

Automatically **retrieves street-level images** from Google Street View.
Takes as input the points generated in Module 1.
Requires a Google Street View API key



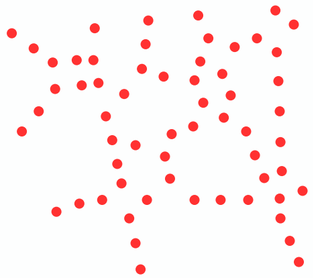
Headings, pitch and fov can be modified

Parameters

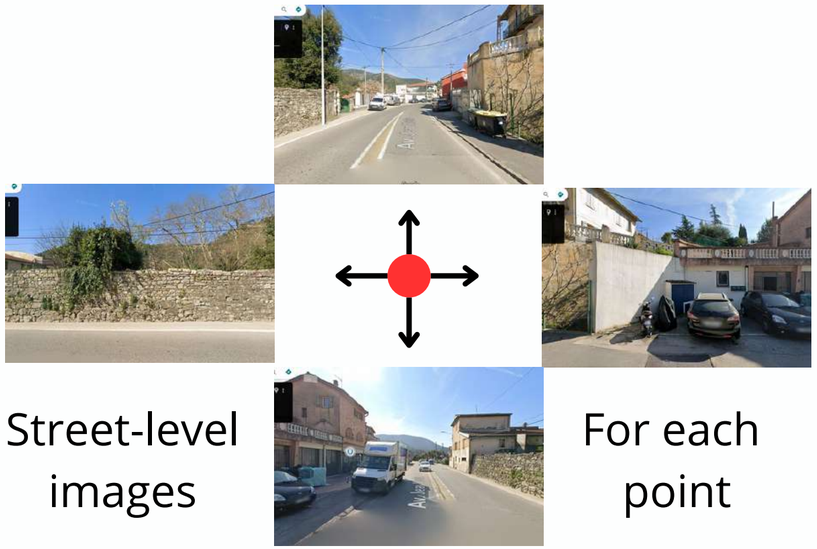
- Number of views → default is 4 per point (0°, 90°, 180°, 270°), but user can adjust.
- Camera settings → pitch and field of view can be modified.
- Filtering → invalid or placeholder images (“no imagery”) are flagged and removed.

Outputs

A structured dataset of georeferenced images linked to each input point, ready for vision–language scoring in Module 3

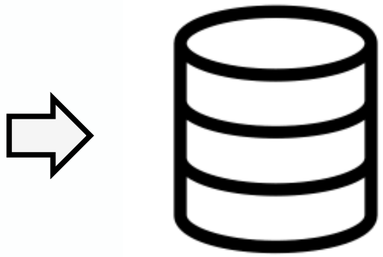


points



Street-level images

For each point

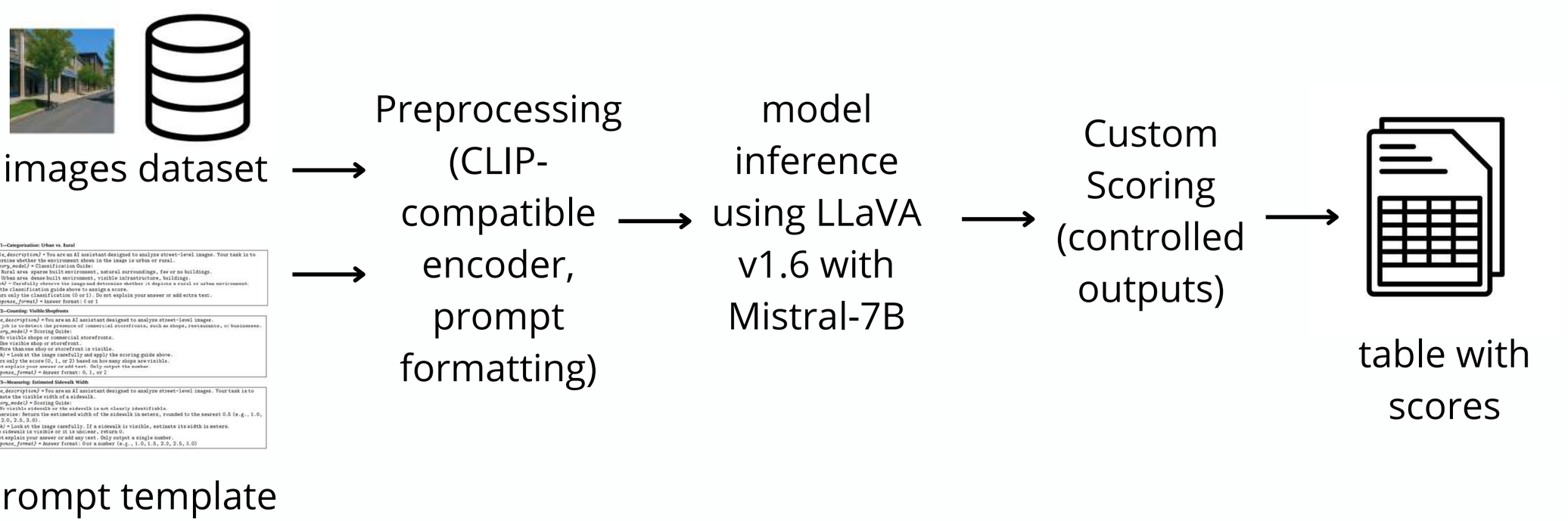


Geolocated images dataset

MODULE 3: SCENE ASSESSMENT WITH LLAVA 1/2

Purpose & Inputs

Core of the SAGAI workflow: performs **automated scoring of streetscape images**.
Input: images from Module 2, plus structured prompts defining the scoring task.
Model: **LLaVA v1.6 with Mistral-7B** backbone (loaded in 4-bit quantized format for lighweight inference).
Open-access implementation of the model and checkpoint via Hugging Face (default checkpoint: liuhaotian/llava-v1.6-mistral-7b).



Outputs

A table of numerical scores, each linked to its image and street segment. Preserves traceability: missing/unavailable images are flagged but still recorded

MODULE 3: SCENE ASSESSMENT WITH LLAVA 2/2

Parameters

- A **dedicated function** structures the interaction between image input and the LLaVA model (CLIP-compatible vision encoders, etc.)
- **Task selection** via identifiers: *T1 → Urban vs. rural classification / T2 → Shopfront counting / T3 → Sidewalk width estimation.*
- **Prompt templates** that can be modified including {role_description} , {theory_model}, {task} & {response_format}
- **Controlled generation**: low-temperature sampling + stopping criteria → enforce concise, numerical outputs.
- **Resuming mechanism**: continues from last processed image if run is interrupted.

Task T1—Categorization: Urban vs. Rural

{role_description} = You are an AI assistant designed to analyze street-level images. Your task is to determine whether the environment shown in the image is urban or rural.
{theory_model} = Classification Guide:
- 0: Rural area sparse built environment, natural surroundings, few or no buildings.
- 1: Urban area dense built environment, visible infrastructure, buildings.
{task} = Carefully observe the image and determine whether it depicts a rural or urban environment.
Use the classification guide above to assign a score.
Return only the classification (0 or 1). Do not explain your answer or add extra text.
{response_format} = Answer format: 0 or 1

Task T2—Counting: Visible Shopfronts

{role_description} = You are an AI assistant designed to analyze street-level images.
Your job is to detect the presence of commercial storefronts, such as shops, restaurants, or businesses.
{theory_model} = Scoring Guide:
- 0: No visible shops or commercial storefronts.
- 1: One visible shop or storefront.
- 2: More than one shop or storefront is visible.
{task} = Look at the image carefully and apply the scoring guide above.
Return only the score (0, 1, or 2) based on how many shops are visible.
Do not explain your answer or add text. Only output the number.
{response_format} = Answer format: 0, 1, or 2

Task T3—Measuring: Estimated Sidewalk Width

{role_description} = You are an AI assistant designed to analyze street-level images. Your task is to estimate the visible width of a sidewalk.
{theory_model} = Scoring Guide:
- 0: No visible sidewalk or the sidewalk is not clearly identifiable.
- Otherwise: Return the estimated width of the sidewalk in meters, rounded to the nearest 0.5 (e.g., 1.0, 1.5, 2.0, 2.5, 3.0).
{task} = Look at the image carefully. If a sidewalk is visible, estimate its width in meters.
If no sidewalk is visible or it is unclear, return 0.
Do not explain your answer or add any text. Only output a single number.
{response_format} = Answer format: 0 or a number (e.g., 1.0, 1.5, 2.0, 2.5, 3.0)

Prompt template for T1, T2 & T3

MODULE 4: GEOSPATIAL SCORING AGGREGATION & MAPPING

Purpose & Inputs

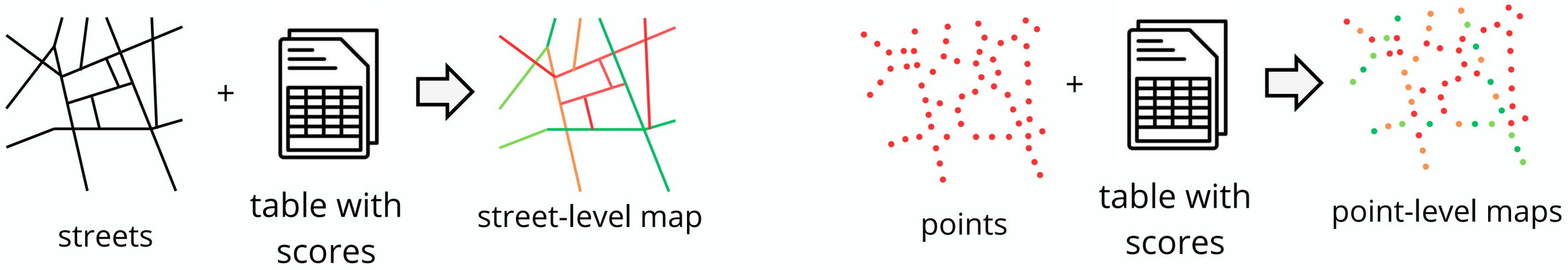
Aggregates the numerical scores from Module 3 with the spatial geometries from Module 1. Converts raw image-based assessments into **geospatial indicators with automated maps**.
Inputs: spatial geometries, task identifier (T1, T2, T3…), and aggregation mode (mean or sum).

Parameters

- Aggregation levels:
  - Point-level scores → retain values for each sampling point.
  - Street-level scores → aggregate by street segment using unique identifiers.
- Summary statistics: mean, sum, count, and standard deviation (stored for both points and streets).
- Cartographic rendering: two thematic maps automatically generated (point-level vs. street-level).
- Missing data handling: streets and points without valid imagery appear in gray.

Outputs

A GeoPackage with both point- and street-level indicators + thematic maps



OPEN-SOURCE AVAILABILITY & REPOSITORY ACCESS

Summary

- SAGAI v1.0: **publicly accessible** on a Git repository

https://github.com/perezjoan/SAGAI

- Designed for **zero-shot deployment on free-tier Google Colab** environments.
- Minimal configuration:
  - Bounding box coordinates.
  - Google Street View API key.
- No local installation or specialized hardware required.
- Released under an open license (Apache 2.0).

Repository content

- **Colab-ready Python notebooks** covering all four components of the pipeline and enabling **direct execution in the cloud**
- Predefined prompts for the three scoring tasks
- **Sample datasets**: inputs and outputs for Nice and Vienna case studies
- Detailed NOTICE files for each module: inputs, configuration, expected outputs.

Link to repo

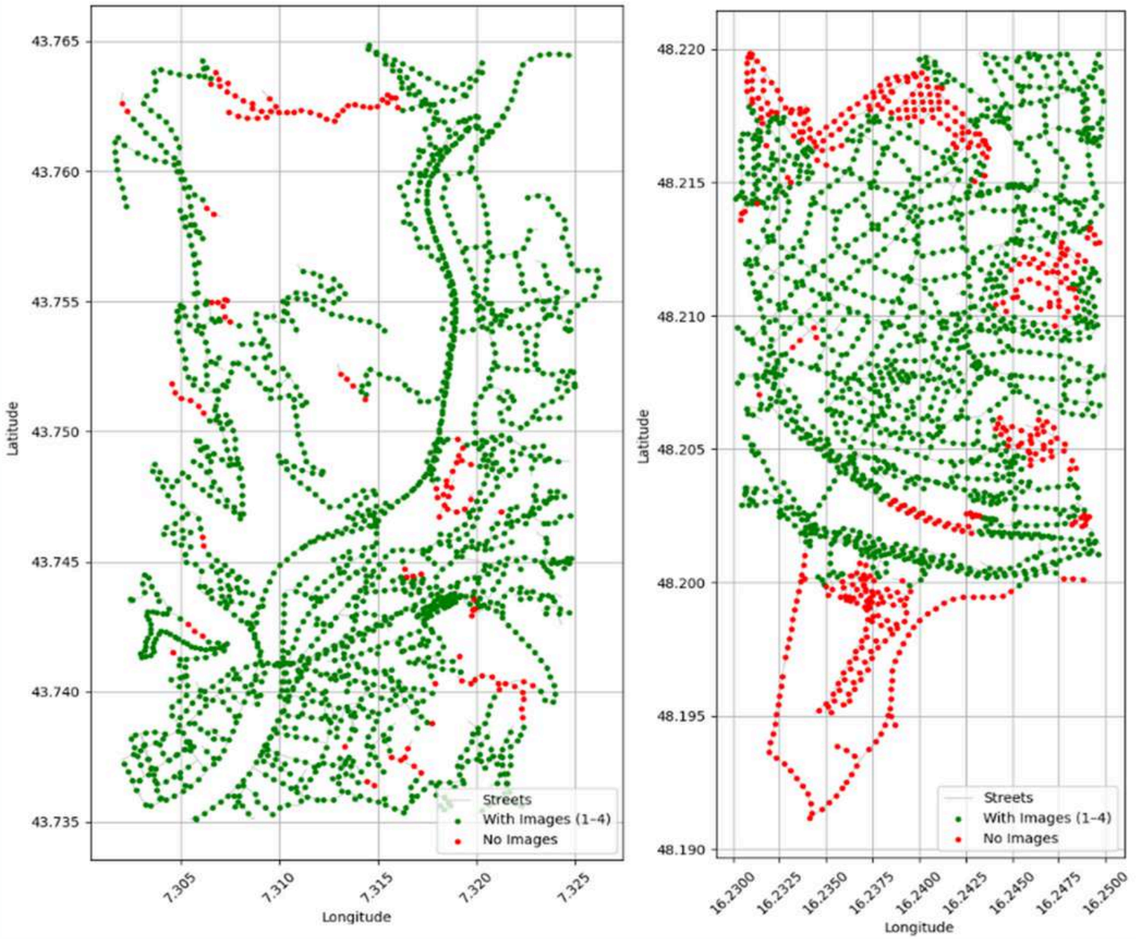
GitHub

EXPERIMENT : CASE STUDIES

Two suburban sectors with diverse morphological characteristics

Nice - Paillon Valley. Highly urbanized strip of flatland, strong topographic constraints.
Vienna - Penzig & Wolfersberg. Hilly suburb with rural patches.

Metric	Nice (Paillon Valley)	Vienna (Penzing & Wolfersberg)
Number of street segments	955	1228
Total street length	141.31 km	154.84 km
Total number of points	1898	1948
Bounding box surface	6.18 km2	4.96 km2
Points with 4 images	1775	1499
Points with no coverage	123	449

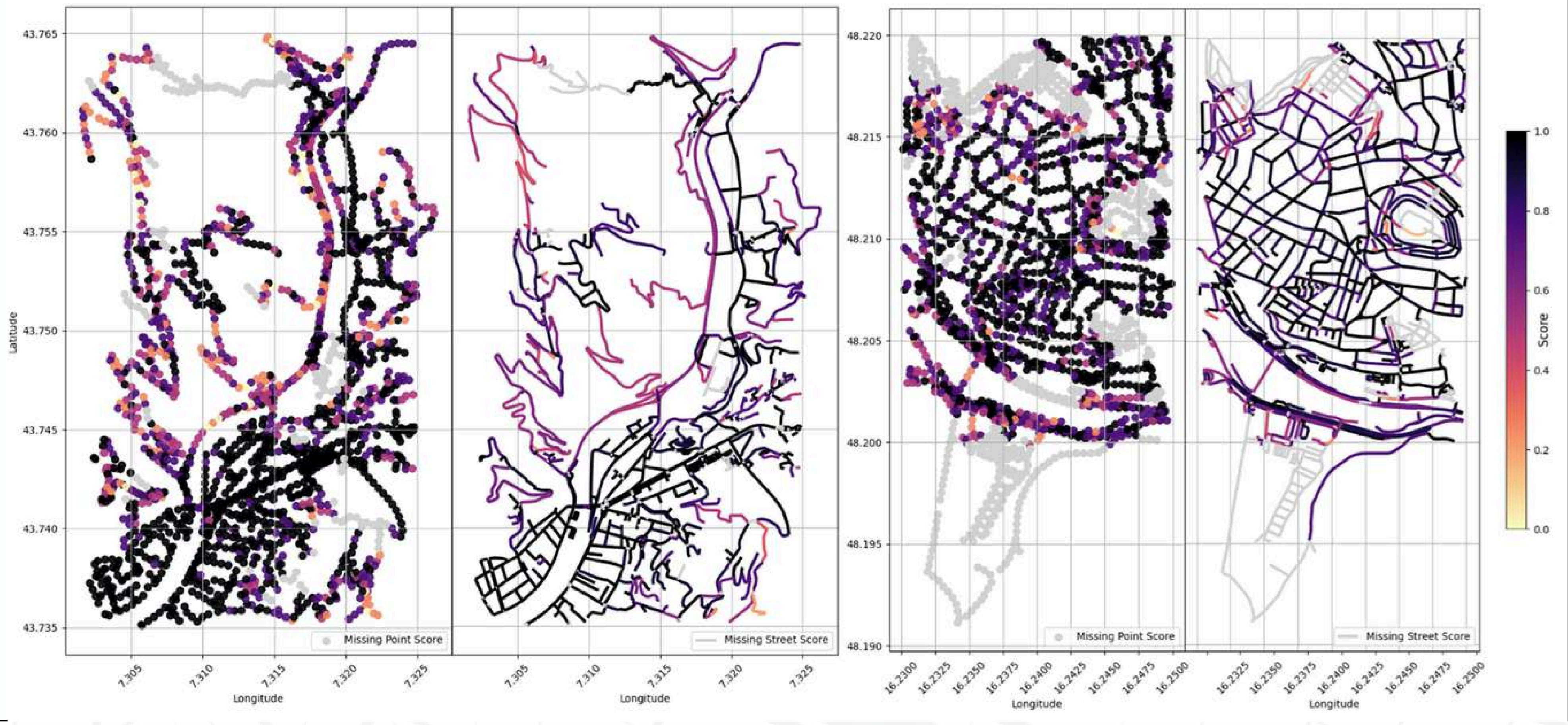


Good coverage by Google StreetView

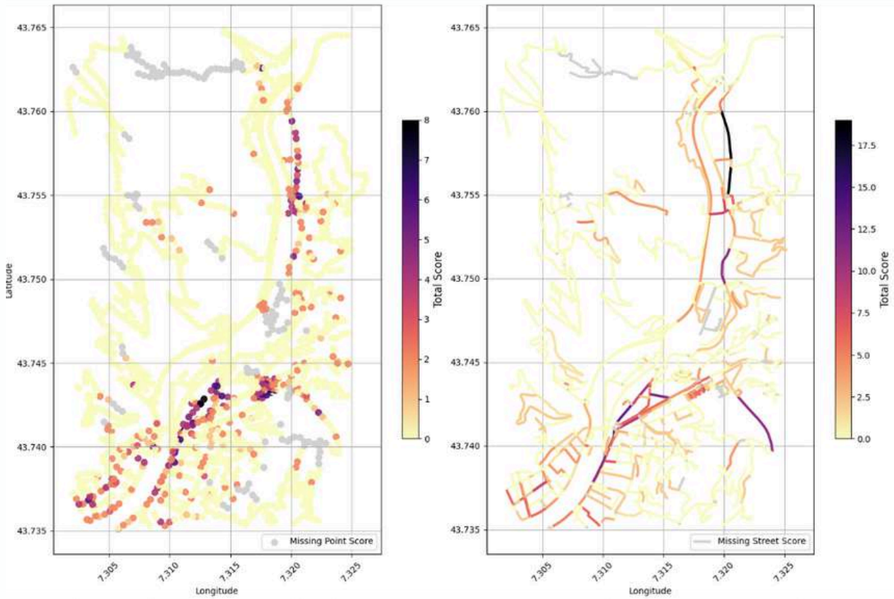
EXPERIMENT : CASE STUDIES

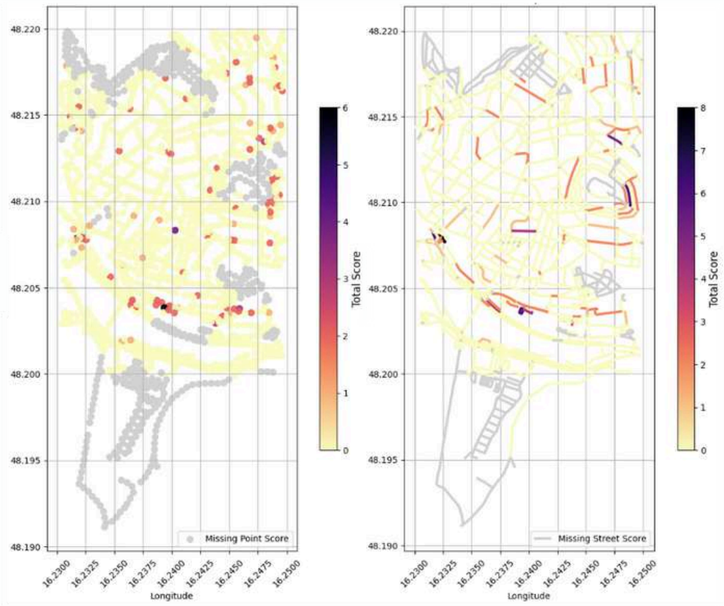
Comparison of model prediction with **manual annotations** (50 random frames for each task for each case study = 300 images, 2 annotators)

Task1 Urban-Rural characterization.
Good overall accuracy (91.7%)



EXPERIMENT : CASE STUDIES

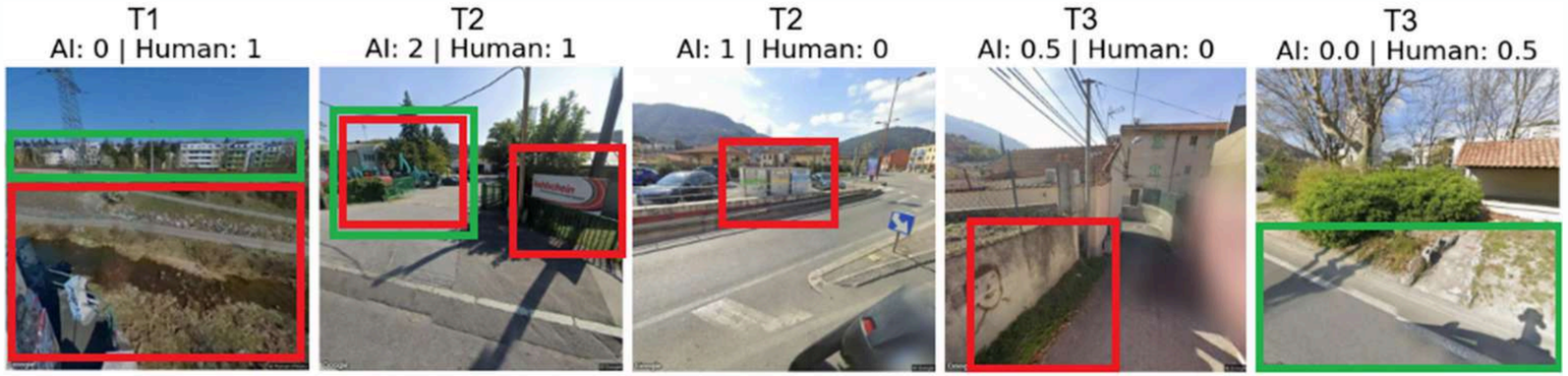




Task2 Shopfront counting. Good accuracy in Nice (73%), lower in Vienna (55%). Main commercial clusters identified.

Task3 Sidewalk depth measurement. More complex task. Lower accuracy (54%). A few misinterpretations.

Missclassified images. Background vs foreground, commercial signs which are not storefronts, colorful artefacts which are not storefronts, GAI cultural biases: grass stripes which are not sidewalks, pedestrian bands which are not considered sidewalks...



CONCLUSION

SAGAI could perform three complementary tasks (scence characterization, storefront counting, sidewalk depth measurement)**, with relatively good accuracy** (92%, 64%, 54%) in two different European suburbs.
Beyond accuracy in predicting human annotations, retreived spatial structures correspond to known charachteristics of case studies (tasks 1 & 2).

The **pipeline** is **fully automatized** and its performace is already acceptable (Step 3, 2h30 for 13k images) **with customizable prompts**.

What was the role of AI in the knowledge production process? (Messeri & Crocket 2023) … definetly **a "Surrogate"** of human analysts who could not manually evaluate the 13k images for each of the 3 tasks. **Humans keep control of the pipeline**, of prompt engineering to operationalize theory-driven concepts, and of downstream analyses.

The perspective is open for :

- **Automatize recognition of complex streetscape patterns** formalized through natural language (like Christopher Alexander's) or through quantitative metrics (like Ewing's)
- **Scale up** from analyzing urban districts **to** analyze **whole metropolitan regions**
- goal: a **metropolitan-wide assessment of fine-grained qualities of urban space**, to assess the **adequacy of 15mC to pedestrian needs**.

FUTURE WORK

Lightweight Improvements

Prompt engineering → enforce “NA” outputs or uncertainty tags.
Multi-pass voting → run several times per image, keep majority.
Rule-based validation → suppress unrealistic values, filter degraded images.

Model Development

Experiment with higher-precision quantization (8- or 16-bit)
Explore larger llm backbones (LLaMA-2 13B, Mixtral 8×7B) and/or vision towers (e.g, BLIP-2 for detection-based towers)
Test new LLaVA variants (e.g., LLaVA-Plus, LLaVA-NeXT).
Comparative benchmarks with proprietary VLMs (Gemini, GPT-4V) and CNN-based pipelines.

Learning Strategies

Current version = zero-shot only.
Add an optional **few-shot learning module** with small annotated datasets for local calibration.

Performance & Scaling

Current throughput ≈ 1,200 images/hour on Colab free tier: can be improved by parallel downloads & batch inference scorings
Optimize image retrieval → align capture with street orientation (forward/backward views).

Data Sources & Extensions

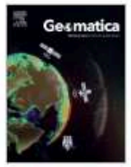
Leverage Street View metadata (camera coords, heading) for cleaner geolocation.
Explore temporal imagery (Mapillary, dynamic GSV APIs) for change detection.
Extend beyond street-level: high-resolution aerial imagery, micro-built environment analysis.

PUBLICATIONS

ELSEVIER

Geomatica

Volume 77, Issue 2, December 2025, 100063



Streetscape Analysis with Generative AI (SAGAI): Vision-language assessment and mapping of urban scenes

Joan Perez [a], Giovanni Fusco [b]

Show more

https://doi.org/10.1016/j.geomat.2025.100063 ↗

Get rights and content ↗

Link to paper

Thanks for your attention

jperez@urbangeoanalytics.com
giovanni.fusco@univ-cotedazur.fr