GEOAI 2026 · GHENT · 3–6 JUNE · ORAL PRESENTATION

Mapping the semantics of the *street.*

AUTHORS

Yunus Serhat Bıçakçı
Joseph Shingleton
Yu Wang
Ana Basiri

AFFILIATIONS

Marmara University, Istanbul, Türkiye
University of Glasgow, Glasgow, UK
Alan Turing Institute, London, UK

DATA & MODELS

Mapillary · Fatih, Istanbul
Qwen3-VL 4B Instruct
Qwen3-VL-Embedding 2B

WHY STREET-VIEW, WHY NOW

Streets are a rich data layer. *Mostly unread.*

CONVENTIONAL APPROACH

Coarse land-use classes.

Misses how streets *look, feel, and function.*

Expert street audits.

Rich, but do not scale beyond small study areas.

OUR PROPOSAL

VLM × multimodal embedding.

Semantically dense, scalable, no manual labels.

RQ · THE QUESTION WE SET

Does the joint distribution of VLM-derived images and descriptions yield *geographically coherent* structure when projected back into space?

IF YES

A scalable base for evidence-based urban analytics.

IF NO

VLM outputs are descriptive but not geographic.

A MAXIMALLY COMPLEX TEST AREA

Fatih, Istanbul: the city's *oldest core.*

FIGURE 01 *Schematic location of the Fatih study area. The historic peninsula bounded by the Golden Horn, the Bosphorus, and the Sea of Marmara.*

DISTRICT

Fatih, Istanbul

WHY THIS PLACE

A historically stratified, socio-spatially complex core.

OPERATING HYPOTHESIS

If a method works here , it has a real shot at generalising.

MAPILLARY + A VLM QUALITY FILTER

~130,000 frames in, *116,696 out.*

01 · COLLECT

Mapillary download

~130,000 frames

02 · FILTER

VLM quality check

−13,000 discarded

03 · ANALYSE

Retained set

116,696 usable frames

PIPELINE OVERVIEW

Three stages, all open-source.

Describe each frame with a VLM. Encode the image × description pair as a single multimodal vector. Cluster.

01 · DESCRIBE

GeoAI-oriented VLM description

Prompted for land-use character, street type, semantic tags, place-activity cues, streetscape quality, and a scene narrative, returned as a structured JSON object.

`Qwen3-VL 4B Instruct`

02 · EMBED

Joint image × text encoding

The frame and its JSON description are encoded together. Each pair becomes a vector $z_i \in \mathbb{R}^{2048}$.

`Qwen3-VL-Embedding 2B`

03 · CLUSTER

Unsupervised partition

MiniBatchKMeans over the 2048-d space. Selected k = 3 from an internal validity sweep over $k \in \{2, \ldots, 10\}$.

`MiniBatchKMeans · scikit-learn`

FIGURE 02 · PIPELINE SCHEMATIC

Joint encoding may amplify semantics; image-only / text-only / joint ablations are flagged as future work.

STAGE 01 · THE PROMPT

A constrained JSON schema, one frame at a time.

Rather than asking for a free-form caption, the VLM is asked to **act as a GeoAI analyst** , emitting structured fields on land-use, morphology, place character, heritage cues, and a short scene narrative.

LAND-USE

residential · commercial · institutional · historic_cultural · transportation

STREET TYPE

arterial · collector · local · pedestrian · alley · plaza

PLACE CHARACTER

dominant activity · temporal markers · human presence · visual complexity

HERITAGE CUES

period · mosque · minaret · cobblestone · turkish_script

```
{
"scene_narrative":        "<80–120 word urban analyst report>" ,
"land_use_character":     { "primary": "residential", "intensity": "medium"  },
"urban_morphology":       { "street_type": "local",    "enclosure_ratio":  "medium"  },
"streetscape_elements":   { "sidewalk_quality": "fair", "signage_density":  "medium"  },
"mobility_infrastructure":{ "modes_visible": ["pedestrian", "car" ] },
"place_character":        { "dominant_activity": "mixed_active"  },
"historic_cultural":      { "architectural_period": "ottoman" ,
"heritage_features":    ["cobblestone", "mosque" ] },
"environmental_quality":  { "greenery_coverage": "minimal", "cleanliness":  "fair"  },
"spatial_safety_cues":    { "lighting_adequacy": "good", "sightlines": "clear"  },
"geo_context":            { "neighborhood_type": "historic_core"  },
"image_quality":          { "usable_for_analysis": true, "issues": ["none" ] },
"semantic_tags":          ["historic", "narrow", "commercial", "residential" ]
}
```

WHY K = 3

Two metrics peak at *k = 3*; one favours *k = 8*.

We chose **k = 3** *to prioritise macro-scale interpretability, in line with the Silhouette and Calinski-Harabasz peaks. k = 8 is kept as a finer-grain reference.*

	K = 2	K = 3	K = 4	K = 5	K = 6	K = 7	K = 8	K = 9	K = 10
Silhouette ↑	0.091	**0.111**	0.083	0.089	0.070	0.081	0.078	0.073	0.065
Calinski-Harabasz ↑	416.3	**497.9**	427.7	370.6	344.3	293.7	310.2	268.3	269.0
Davies-Bouldin ↓	3.140	2.573	2.704	2.728	2.591	2.764	**2.498**	2.895	2.715

TABLE 01 · INTERNAL CLUSTERING VALIDATION

Higher is better for Silhouette and CH; lower is better for Davies-Bouldin.

2D PROJECTION OF THE 2,048-D SPACE

k = 3 *Selected* · coarse macro-typology

k = 8 Reference · finer DB-suggested partition

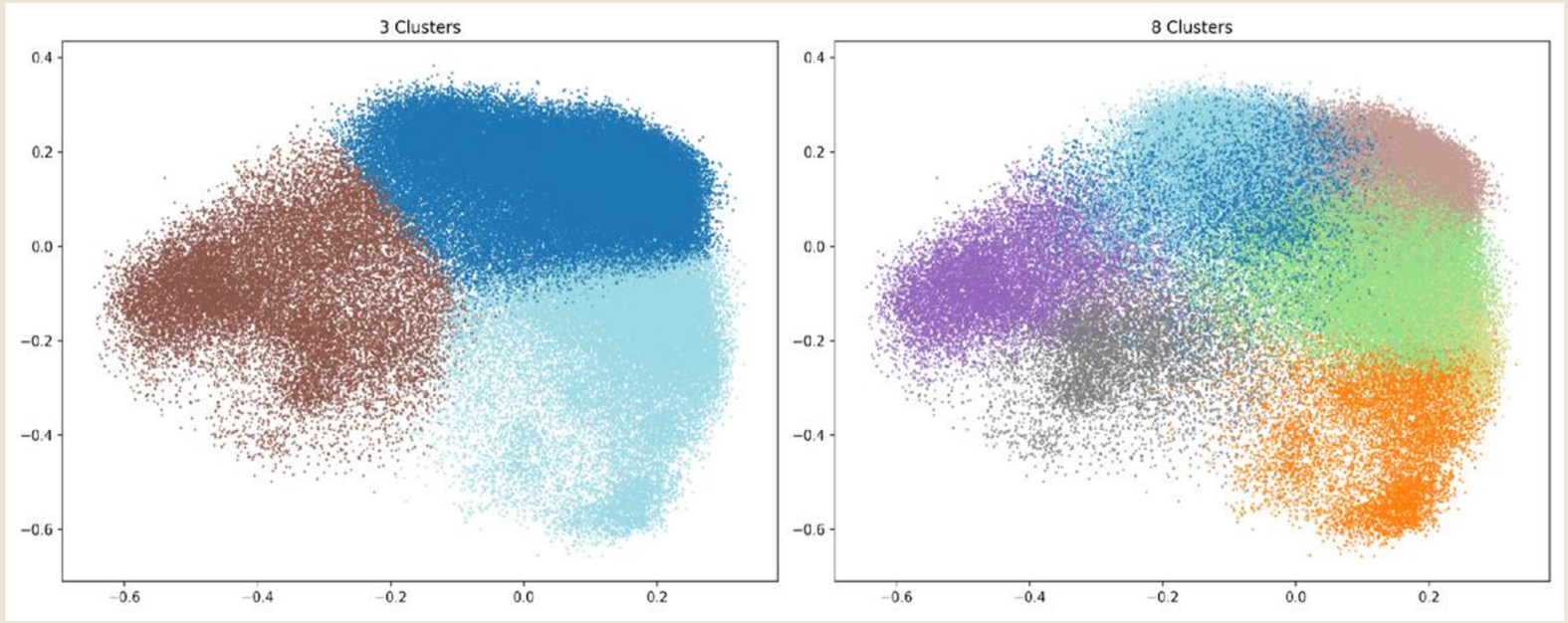


FIGURE 03 · 2D PROJECTION OF THE MULTIMODAL EMBEDDING SPACE

Fatih, by what it *looks like.*

CLUSTER 1

Ordinary local streets

Side streets and residential areas of the district.

Tags: *narrow, residential, commercial, mixed use.*

CLUSTER 2

Heritage & tourism

Historic and tourist environments, often pedestrian.

Tags: *tourism, historic, quiet, religious.*

CLUSTER 3

Traffic corridors

Wider roads, traffic-dominant environments.

Tags: arterial, moderate traffic, modern, commercial.

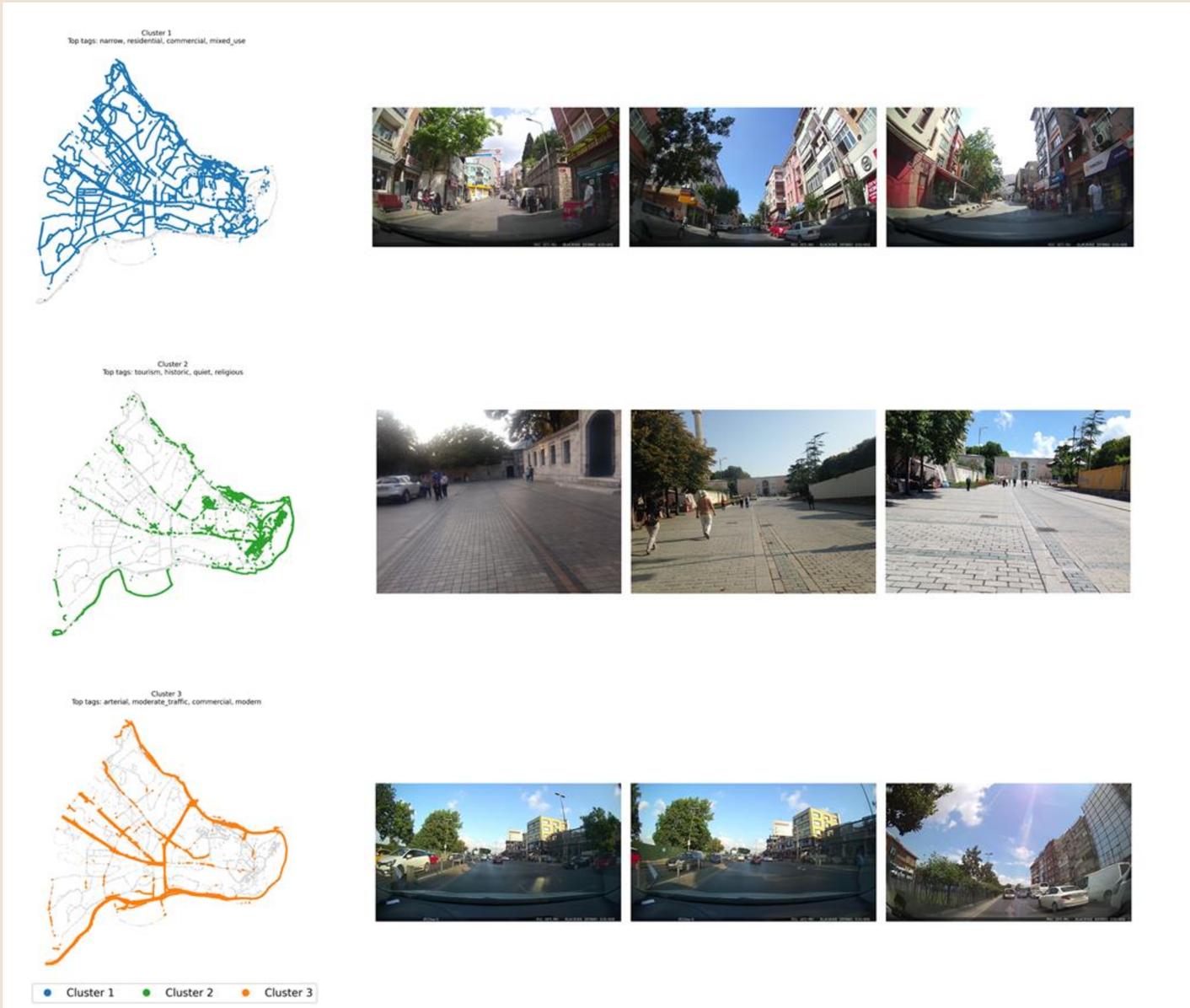


FIGURE 04 · REPRESENTATIVE EXAMPLES FOR EACH CLUSTER

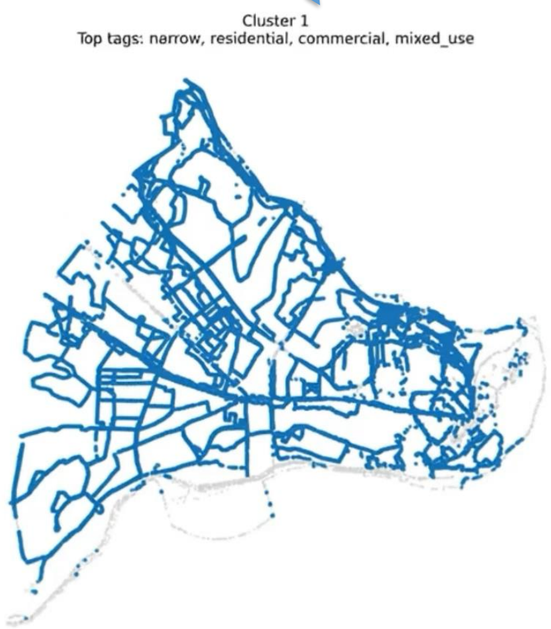
Fatih, by what it *looks like.*

CLUSTER 1

Ordinary local streets

Side streets and residential areas of the district.

Tags: *narrow, residential, commercial, mixed use.*



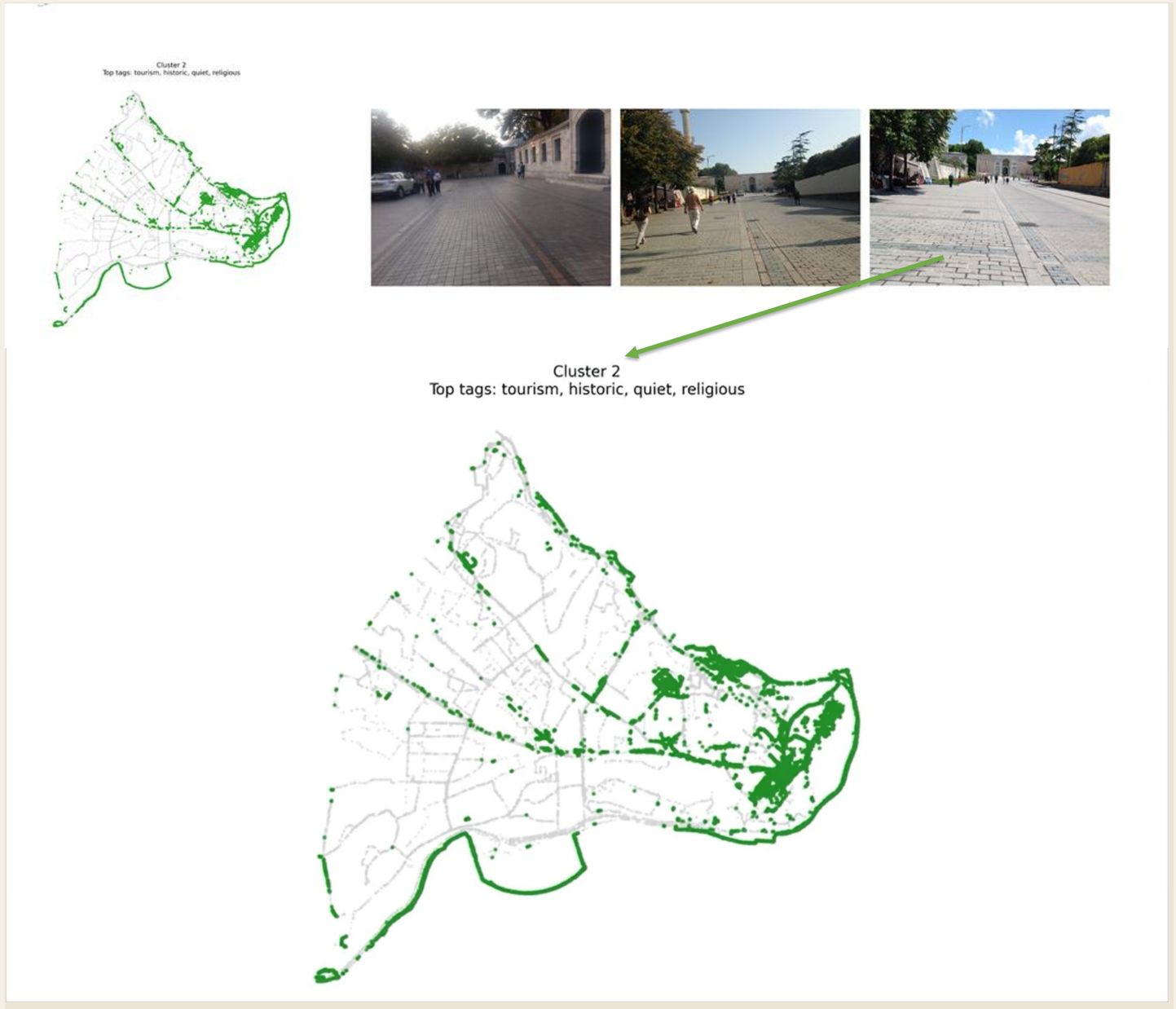
Fatih, by what it *looks like.*

CLUSTER 2

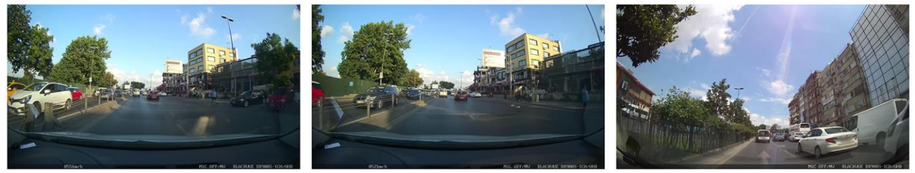
Heritage & tourism

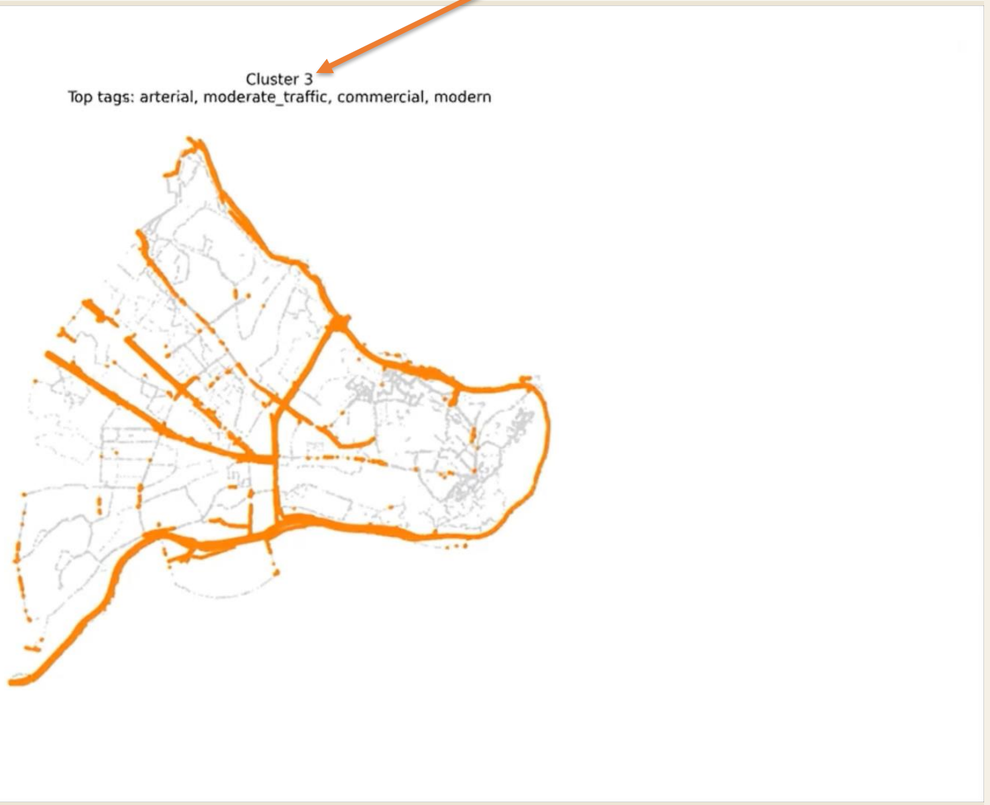
Historic and tourist environments, often pedestrian.

Tags: *tourism, historic, quiet, religious*.



Fatih, by what it *looks like.*





CLUSTER 3

Traffic corridors

Wider roads, traffic-dominant environments.

Tags: arterial, moderate traffic, modern, commercial.

TWO FINDINGS WORTH SAYING OUT LOUD

The clusters are *geographic*, not artefactual.

EXTERNAL VALIDATION · OSM OVERLAP

88.4 %

of the 29,968 frames in **Cluster 3 (traffic corridors)** *sit on streets tagged as highway in OpenStreetMap (26,484 of them).*

`n = 29,968 · OSM highway features`

INTERNAL VALIDATION · SEQUENCE SPANS

42.1 %

of capture sequences **span multiple clusters** *. The groupings reflect semantic transitions in the street, not camera or contributor artefacts.*

`capture sequences with ≥ 2 cluster IDs`

MICRO-SCALE SEPARABILITY · SOUTHERN COASTAL EDGE

Cluster 2 (pedestrian sidewalks by the sea) and Cluster 3 (the adjacent highway, only metres away) stay distinct. The VLM-derived representation can resolve micro-scale shifts in street function from visual perspective.

CONCLUSION IN ONE LINE

An interpretable semantic atlas, *without manual labelling.*

01	CONTRIBUTION	VLM-derived multimodal embeddings recover urban structure that aligns with established morphology at city-block resolution.	A scalable alternative to expert-driven audits for characterising urban environments.
02	LIMITATION	Cluster semantics are assigned *post hoc*.	Not yet validated against local expert knowledge. Stratified manual checks + independent ground-truth labels are needed.
03	LIMITATION	Possible cultural bias in VLM descriptors.	Concepts such as “place character” and “urban quality” are culturally situated; VLM training data may reproduce Western-centric interpretations.
04	LIMITATION	Mapillary coverage is uneven.	Spatial and temporal coverage may under-represent certain neighbourhoods or time periods.
05	FUTURE WORK	Image-only · text-only · joint embedding ablations, local expert review, and cross-city comparison.	Does the schema generalise outside Fatih? That is the natural next benchmark.

THANK YOU

Yunus Serhat Bıçakçı, PhD
yunus.serhat@marmara.edu.tr