The eye of the beholder: Neighbourhood embeddings and housing prices*

Laurent Gobillon† Clément Gorin‡ Stephan Heblich§
Amine Ouazad¶

Version 2026.03.27

Abstract

Urban neighborhoods differ along dozens of dimensions simultaneously—building stock, greenery, walkability, transport access—yet households cannot attend to all of them equally when making residential choices. This paper proposes a framework for measuring neighborhood amenities from street-level imagery and embedding them into a quantitative spatial model with limited attention. Using 3,825,241 Google Street View panoramas linked to 1,075,587 housing transactions in Paris over 2010-2024, we construct visual embeddings summarizing street appearance. In hedonic regressions, these embeddings explain a substantial share of housing price variation beyond standard controls, with spatial spillovers decaying beyond 250 meters. We decompose the image embeddings into interpretable semantic objects using a vision-language model and sparse coding. Amenities are then embedded into a quantitative spatial model where households allocate limited cognitive attention across dimensions, determining which amenities influence equilibrium prices and residential sorting. Counterfactuals allow us to evaluate the impact of changes in visual characteristics.

Keywords: Urban amenities, housing prices, street-view imagery, representation learning, quantitative spatial models.

*We thank seminar, conference, and workshop participants for comments and suggestions. The usual disclaimer applies.

†Paris School of Economics-CNRS, CEPR and IZA. Email: laurent.gobillon@psemail.eu
‡Université Paris 1 Panthéon-Sorbonne, Email: clement.gorin@univ-paris1.fr
§University of Toronto, CESifo and NBER. Email: stephan.heblich@utoronto.ca
¶HEC Montreal, Department of Economics. Email: amine.ouazad@hec.ca

1 Introduction

Much of what distinguishes neighborhoods is visible but hard to measure. Households care about features such as greenery, streetscape, building design, and access to public space, but these characteristics do not enter standard datasets in a systematic way. They are high-dimensional, vary sharply over short distances, and tend to resist simple summary measures. Because households value these characteristics and we cannot directly observe them, empirical work typically absorbs them into neighborhood effects.[1] While composite neighborhood effects help rationalize observed prices and location choices, they remain a black box that comes at a cost: counterfactuals that alter the built environment directly require the amenity content of each location to be observable and not merely inferred as a residual.

This paper opens the black box by replacing composite neighborhood effects with directly measured visual characteristics. The starting point is that much of what distinguishes neighborhoods is visually encoded in the built environment, and that recent advances in representation learning make it possible to extract this information systematically. We map street-level imagery into embeddings that preserve semantic content, decompose them into interpretable object scores, and bring these measures into a quantitative spatial model. Hedonic regressions confirm that our measures capture meaningful variation in neighborhood quality: object scores explain housing price variation with expected signs and magnitudes, and the implied implicit prices align with independent geographic data. Having established this, we embed the measures into a quantitative spatial model where we decompose the residual amenity term into systematic visual variation and a truly idiosyncratic component. This decomposition enables internally consistent counterfactuals: because the systematic part is directly observed and alterable, we can predict the equilibrium consequences of specific changes to the built environment.

We implement this approach using a new dataset of 3,825,241 *Google Street View* panoramas linked to 1,075,587 housing transactions in the Paris *petite couronne*[2] be-

[1] In reduced-form estimations, neighborhood fixed effects control for unobserved location-level confounders, recovering within-neighborhood variation in prices or outcomes. In quantitative spatial models, residual amenity terms rationalize observed location choices that prices and wages net of commuting costs alone cannot explain.

[2] This includes Paris and the 3 surrounding *départements*, namely Hauts-de-Seine, Seine-Saint-Denis, and Val-de-Marne, totalling 143 *communes*.

tween 2010 and 2024. We map these high-dimensional images into more compact embeddings using the *Contrastive Language–Image Pretraining* (CLIP) model of Radford et al. (2021), which project images and text descriptions into a shared latent space in which visually similar environments and conceptually related objects are geometrically close. For each panorama, we compute embeddings for four directional views and average them to a single vector, decompose it into interpretable object scores using a sparse coding LASSO approach (Tibshirani, 1996), and aggregate both to the hexagon–year level using Uber's *H3* spatial indexing system at 50-meter resolution.

When included in hedonic regressions, these embeddings explain a substantial share of housing price variation and complement standard unit-level controls. Conditional on year effects alone, the adjusted R^2 is 0.0193, adding unit attributes it increases to 0.0974 while it is 0.2604 with embeddings; including both leads to a statistic of 0.2767. Controlling additionally for neighborhood fixed effects, the within-area R^2 with both attributes and embeddings reaches 0.3152, indicating that embeddings capture variation in neighborhood quality that is not absorbed by broad spatial aggregates. Adding embeddings averaged over concentric rings around each transaction further increases explanatory power by up to 4 percentage points. Incremental gains decay smoothly and become negligible beyond 300 meters. This spatial decay is consistent with CLIP embeddings capturing local neighborhood characteristics rather than broad city-wide trends.

Because embedding dimensions are not directly interpretable, we express each image embedding as a sparse linear combination over a semantically rich dictionary spanning a wide range of urban elements, including vegetation, architectural styles, transport infrastructure, pedestrian amenities, and commercial establishments. Each dictionary element is defined using CLIP text embeddings derived from natural language descriptors, leveraging CLIP's shared image-text embedding space. We then estimate this decomposition using a sparse coding LASSO approach, which selects a compact subset of objects to characterize each image. The resulting object scores recover recognizable urban patterns: vegetation scores are concentrated along tree-lined boulevards and near major parks as mapped by *OpenStreetMap*; Haussmannian scores are densest in the historical *arrondissements*; and cobblestone streets scores peak in *Montmartre*. Hedonic regressions that use these object scores as regressors yield coefficients with expected signs and magnitudes, providing further validation.

The high dimensionality of neighborhood characteristics raises a further conceptual issue: households are unlikely to process all available information when making residential decisions. Even if many aspects of the built environment are potentially relevant, attending to all of them simultaneously would be cognitively costly. We capture this idea by embedding the measured amenities into a quantitative urban model with limited attention, following the insight of Gabaix (2014) that decision-makers simplify complex environments by selectively focusing on a subset of salient attributes. Households allocate attention weights across amenity dimensions, trading off the utility gain from finer environmental assessment against a cognitive cost. Attention operates by distorting the valuation of attributes rather than by introducing noisy choices: households behave as if they evaluate neighborhoods using a simplified representation that places positive weight on only a limited set of characteristics, remaining fully rational conditional on this perceived utility. Attention is chosen ex ante, before idiosyncratic preference shocks are realized, and thus reflects a forward-looking decision about how to process neighborhood information. This interpretation is consistent with residential choice being an infrequent, deliberate decision.[3] This formulation distinguishes our framework from standard rational inattention models, in which limited attention typically manifests as stochastic choice.

In the model, the amenities that receive positive attention are exactly those capitalized into equilibrium prices and that drive residential sorting. This provides a structural link between the empirical selection of features in hedonic regressions—which may be understood as measuring revealed attention—and the equilibrium consequences of amenity changes. Because amenities are explicitly measured rather than absorbed into fixed effects, the framework supports counterfactuals that alter exogenous neighborhood characteristics directly. As an illustration, we trace the effect of replacing Haussmannian architectural style with modern design across Paris by modifying image embeddings along the corresponding direction in CLIP space, then feeding the modified embeddings through the equilibrium model to compute the implied changes in prices, wages, and population shares.

Our paper relates to several strands of the literature. The first is representation learning (Bengio et al., 2013; LeCun et al., 2015), which provides the conceptual foundations

[3]This formulation follows Gabaix (2014) rather than information-theoretic models of rational inattention (Sims, 2003; Matějka and McKay, 2015), in which limited attention manifests as residual choice uncertainty arising from costly signal acquisition.

for learning compact, informative representations from high-dimensional unstructured data. A growing empirical literature applies image and text embeddings to neighborhood outcomes: Naik et al. (2017) pioneer the use of street view images to measure perceived safety and socioeconomic conditions; Cafarella et al. (2023), Yazdani and Raissi (2023), Wheaton and Xu (2024), and Bajari et al. (2025) use visual or textual embeddings to study variation in housing prices, neighborhood change, or urban amenities. Relative to this work, our contribution is to move beyond reduced-form estimates by decomposing embeddings into interpretable components, validating them against external data, and integrating them into a quantitative urban model.

A second strand combines unstructured data with economic models of market demand and asset prices. Compiani et al. (2023), Bach et al. (2025), Gabaix et al. (2025), and Magnolfi et al. (2025) show how text or image data can inform preferences, beliefs, or valuations. We share the objective of grounding economic primitives in rich data, but focus on residential choice and explicitly address high dimensionality through selective attention.

Third, our paper contributes to the literature on amenities in quantitative spatial models. Ahlfeldt et al. (2015) provide a canonical framework in which amenities enter utility as a residual term inferred from prices and quantities. More recent work, including Ang et al. (2024) and Almagro et al. (2025), seeks to discipline the role of amenities and improve counterfactual analysis. Our contribution takes this one step further: rather than inferring amenities indirectly, we measure them directly using visual data and embed them into a spatial equilibrium model with limited attention, allowing amenities to affect utility and equilibrium outcomes through observable, interpretable channels while preserving the tractability of the Ahlfeldt et al. (2015) framework.

The remainder of the paper is organized as follows. Section 2 describes the data. Section 3 introduces the representation learning procedure and the CLIP model. Section 4 presents hedonic regression evidence, describes the decomposition of embeddings into semantic objects, and validates the object measures. Section 5 presents the quantitative spatial model with limited attention. Section 7 then illustrates the properties of the model using equilibrium simulations. Section 8 describes the counterfactual exercises and Section 9 concludes.

2 Data

2.1 Spatial grid

Our study area covers Paris and the three surrounding *départements* of the *petite couronne*, totaling 143 French communes, including the 20 *arrondissements* of Paris. We discretized this area using the Uber H3 hierarchical spatial indexing system, which tessellates the earth's surface into hexagons at a fixed resolution. Our study area contains 397,146 such hexagons of approximately 50 m width, corresponding to resolution 11 in the Uber H3 system. This resolution is fine enough to capture local variation in neighborhood appearance, i.e. the level at which street-level amenities such as tree cover, building style, and streetscape features vary, while coarse enough to allow meaningful aggregation of multiple panoramas and transactions within each cell.

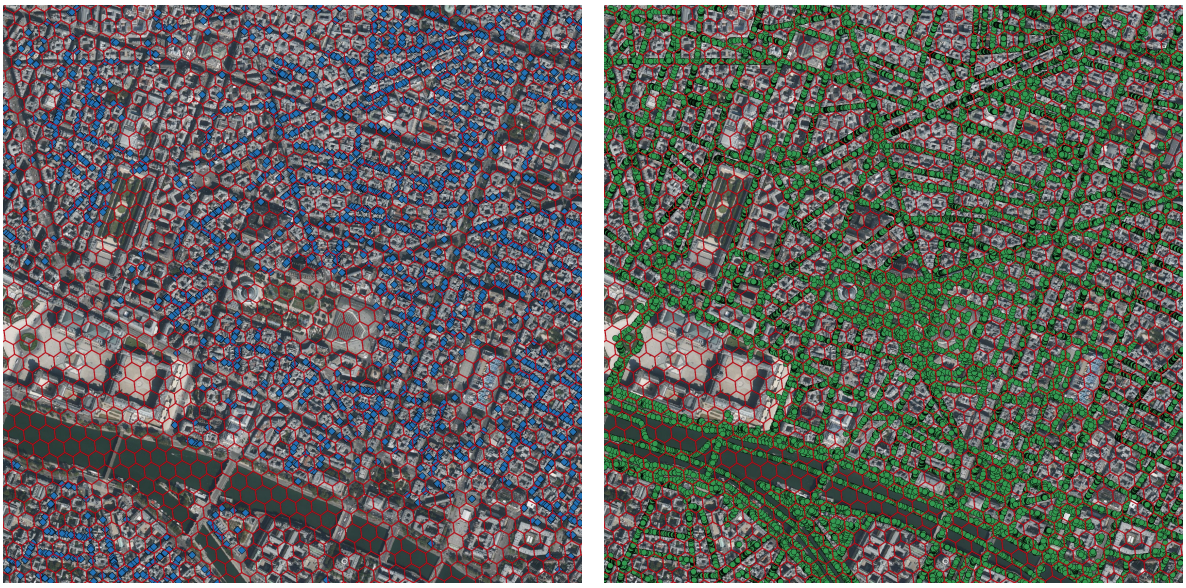
2.2 Housing transactions

Transaction-level data on residential property sales in the Paris urban area come from the *Demandes de Valeurs Foncères* (DV3F) dataset, provided by the French Ministry of Public Finance (DGFiP). DV3F combines transaction records from land registry services (*Services de la Publicité Foncière*) with detailed information on premises from the cadastral property files (*Fichiers Fonciers*), allowing transactions to be allocated to individual parcels.

The sample covers all transactions between January 1, 2010 and December 31, 2024. We restrict our dataset to standard sales of a single flat or house, excluding auctions, social and expropriation transfers, transactions involving multiple or geographically distant properties, extreme dwellings, premises with professional use or dependencies, and any transaction for which premises cannot be identified in the cadastral files. Prices are measured as the total transaction price and standardized to log price per square meter. The DV3F provides a rich set of property characteristics, including built area, land area, floor, number of rooms, bathrooms, bedrooms, kitchens, garages, terraces, conservatories, and annexes, as well as year of construction. The full variable list is given in Appendix A.

After applying these restrictions, the dataset contains 1,075,587 transactions, of which 601,220 (56%) can be matched to hexagons where panorama data are also available for that year. This constitutes our main estimation sample.[4] The spatial and temporal distribution of transactions is shown in Figure 13 in Appendix B. Transaction counts are broadly stable over the sample period, with a mild dip in 2020 reflecting the COVID-19 lockdowns.

Figure 1: Housing transactions and Google Street Google Street View panoramas in central Paris.



(a) Housing transactions (b) Panoramas

Notes: Map overlay showing transactions (blue diamonds) and panoramas (green dots) matched to the H3 grid (red hexagons) over 2010–2024. Transaction locations snapped to building entrance coordinates. Panoramas are retrieved within 50 meters of hexagon centroids and transaction locations.

2.3 Street-level imagery

To measure neighborhood visual characteristics, we collect street-level panoramas from the *Google Street View* API via a third-party service. Query locations are generated from hexagon centroids and from individual transaction addresses, corresponding to building entrance locations as defined in the French Topographic Database (BD TOPO),

[4]The discrepancy between the full transaction count and the matched sample reflects the coverage of the Street View data, which covers 65,699 hexagons, or 17% of the full grid.

produced by the French National Geographic Institute (IGN). For each query point, we retrieve the nearest available panoramas within a 50-meter radius and record associated metadata, including the exact geographic coordinates and capture date. The data span the period from 2010 to 2024.[5]

In total, we collect 3,825,241 panoramas covering 195,756 hexagons, or 49% of the grid. Figure 1b shows the spatial distribution of panorama locations, which largely follows the road network in central Paris. Coverage is densest in the inner *arrondissements* and along major thoroughfares. In addition, Google also captures images in public spaces not accessible by car, such as squares, parks, riverbanks, and other pedestrian areas. The temporal distribution of panoramas across collection years reflects Google's re-collection cycles: captures are not uniformly distributed over time but cluster around years in which Street View vehicles traversed the city.

Each panorama is a $360° \times 180°$ equirectangular image that is projected into four standard perspective views – left, front, right, and back – each corresponding to a 90° horizontal field of view. An example is shown in Figure 2. Each view is encoded as a separate image embedding (see Section 3), and the four directional embeddings are then averaged to produce a single $\mathbb{R}^{1024}$-valued vector per panorama. Panoramas are then aggregated to the hexagon-year level by averaging all embeddings within each hexagon and year. This two-step aggregation—first across directions within a panorama, then across panoramas within a hexagon-year—leverages the additive compositionality of the CLIP embedding space, which we document in Section 3.

2.4 Validation data

To validate the semantic object measures derived in Section 4.4, we use data from *OpenStreetMap* (OSM). These data provide geospatial information on the Paris public transport network (metro, tram, and bus lines and stops), parks, gardens, forests, and other urban features at a fine spatial scale. For each hexagon, we compute the total length of public transport infrastructure and the total area of vegetated land, which we use as external benchmarks for the corresponding object scores. The construction and validation of the OSM-based measures is described in detail in Appendix E.

[5]Since Street View data are not available annually, we choose a longer time horizon to get the best possible overlap with the transaction data.

Figure 2: Example Street View panorama.



Notes: Street view panorama split and its four directional views. The full panorama (top) is split into left, front, right, and back views (bottom row), each of which is encoded separately using CLIP.

3 Neighborhood Embeddings from Imagery

3.1 Representation Learning and CLIP

Representation learning (Bengio et al., 2013) aims to map high-dimensional data, such as text or images, into compact numerical representations that capture the underlying factors of variation, typically in the form of vectors called embeddings. In neural networks, these representations are learned by optimizing a predictive objective that encourages the model to transform and organize input information in ways that are useful for the task at hand. In images, these representations often capture the identity and visual properties of objects, while in text they tend to capture semantic content as well as the relationships between words. The resulting embeddings form a latent geometric structure, the embedding space, in which each observation corresponds to a point. In this space, observations with similar characteristics tend to lie close to one another, while systematic directions correspond to meaningful variations in the

underlying data.

We construct our visual measures using the *Contrastive Language-Image Pretraining* objective (CLIP, Radford et al., 2021). The CLIP framework relies on two separate neural networks, one that computes image embeddings and another that computes text embeddings, mapping both into a common latent space. The training objective encourages the representation of each image to be close to that of its corresponding text description, while pushing apart representations of non-matching image-text pairs, a mechanism known as contrastive learning. As a result, the learned space organizes images and text descriptions geometrically, so that visually or semantically related observations lie close to one another. Formally, given N imagetext pairs, the CLIP loss averages row- and column-wise cross-entropy over the matrix of cosine similarities:

$$\mathcal{L} = -\frac{1}{2N} \sum_{i=1}^{N} \left[\log \frac{e^{s_{ii}/\tau}}{\sum_{j=1}^{N} e^{s_{ij}/\tau}} + \log \frac{e^{s_{ii}/\tau}}{\sum_{j=1}^{N} e^{s_{ji}/\tau}} \right] \tag{1}$$

where s_{ij} is the cosine similarity between image embedding z_i^{img} and text embedding z_j^{txt}, defined as the inner product of the two vectors after normalizing each to unit length, and τ is a learned temperature parameter, controlling how sharply the model distinguishes between matching and non-matching image-text pairs. Geometrically, cosine similarity equals one when two embeddings point in exactly the same direction, zero when they are orthogonal, and minus one when they point in opposite directions; it therefore measures directional agreement independently of how large the vectors are. Minimizing $\mathcal{L}$ pushes matched image-text pairs to point in the same direction while pushing unmatched pairs apart: the two log terms penalize, respectively, the case where image i is closer to the wrong text descriptions than to its own, and the case where text i is closer to the wrong images than to its own.

Regarding the model architecture, we use the open-source *OpenCLIP* implementation (Cherti et al., 2023), which relies on a vision transformer (Dosovitskiy et al., 2021) and a text transformer (Vaswani et al., 2017) trained on the LAION-2B dataset of two billion image-text pairs (Schuhmann et al., 2022). The image transformer divides each image into a grid of fixed-size patches, treated as a sequence of tokens analogous to words in a sentence. Similarly, the text transformer tokenises the input string into a sequence of subword units. Both sequences are passed through successive transformer layers, which

progressively aggregate local information into global representations. For images, early layers detect elementary visual features such as edges and colour gradients; deeper layers combine these into higher-level concepts such as building facades, vegetation, or street furniture. For text, early layers capture lexical meaning while deeper layers resolve syntactic structure and semantic context. Each encoder produces a single 1,024-dimensional vector summarizing the content of its input. The two encoders are trained jointly so that semantically related image-text pairs are mapped to nearby points in this shared space, enabling direct comparison between visual and textual representations. Additional details on the model architecture are provided in Appendix C.

Representations from neural networks are distributed, meaning that any given concept is encoded across multiple dimensions, and each dimension contributes to many concepts. In the geometry of the embedding space, concepts correspond to directions rather than individual axes. Semantically related features therefore induce correlation across dimensions, producing multicollinearity that inflates variance and destabilises coefficient estimates in downstream estimation exercises. We apply Principal Component Analysis (PCA) to obtain an orthogonal rotation of the embedding space, retaining all components ordered by decreasing explained variance. This eliminates collinearity among regressors and stabilises estimation. The corresponding eigenspectrum is reported in Appendix 17.

3.2 Geometric Properties of the Embedding Space

The CLIP training objective encourages three properties of the embedding space are particularly important for our application.

Semantic image representations. The joint imagetext training objective encourages the embedding to capture semantic content such as object identities and visual attributes. These objects and attributes are encoded as directions in the embedding space, spanning multiple dimensions, rather than a single interpretable dimension. Figure 3 illustrates this by computing, for each image patch token, its cosine similarity to text embeddings of descriptions such as “a cobblestone road”, “lined with parked cars”, or “with vegetation overhead”. The similarity scores highlight the image regions that correspond to each concept, confirming that the embeddings localize semantic content

spatially within the image.

Figure 3: Image-text alignment.



Notes: Thanks to the transformer architecture, the *OpenCLIP* model generates intermediate embeddings for each 14×14 pixel patch. The colors represent the cosine similarity between patch embeddings (after applying the final projection) and the text embeddings for "cobblestone road" (left), "parked cars" (centre), and "vegetation" (right). Similarity values are clipped at zero and bilinearly interpolated to match the image dimensions. Blue indicates orthogonality, while red indicates a positive similarity.

Consistent distances. The contrastive training objective ensures that distances between embeddings, measured by cosine similarity, are semantically meaningful and comparable across images: images with similar visual semantic characteristics map to nearby points in the embedding space. This property motivates our practice of averaging embeddings within hexagons and across directions, since the mean of two embeddings in the latent space corresponds to a semantic midpoint. As an illustration, Figure 4 shows the result of a linear interpolation between two images in the embedding space, visualized using the pre-trained *UnCLIP* decoder (Ramesh et al., 2022), a diffusion model mapping a *CLIP* embedding back to the image space.

Additive semantic compositionality. The CLIP training objective is based on the dot product between image and text embeddings. Since the dot product is linear, independently varying visual concepts are efficiently encoded as additive directions in embedding space interference between unrelated concepts is minimised when their representations are orthogonal. This encourages the embedding space to decompose approximately as a sum of contributions from individual visual concepts (Bhalla et al., 2024), which is the geometric basis for the sparse decomposition in Section 4.4. This property is well-established empirically, though it remains approximate, due to

Figure 4: Embedding interpolation.



$\alpha = 0.25$ $\alpha = 0.50$ $\alpha = 0.75$

Notes: Images decoded from weighted combinations $\alpha z_i^{img} + (1-\alpha) z_j^{img}$ for $\alpha \in \{0.25, 0.50, 0.75\}$, with endpoints at left and right.

the presence of systematic co-occurrences (e.g. "car" and "wheel") and hierarchical relationships (e.g. "car" systematically entails "vehicle"), since they are never observed separately during training. These systematically co-occurring concepts are grouped together in the semantic dictionary used for the decomposition.

3.3 From Images to Concepts

A image is a three-dimensional matrix of pixels of dimension d_i, in which pixels are strongly correlated and organized into spatial patterns. Our goal is to convert each image x_{rt}^{img} observed at location r and time t into a set of urban concepts —such as *Haussmannian building*, *paved street*, *green space*— that can be related to housing prices and used in the spatial equilibrium model. This requires solving two problems: First, raw images are high dimensional, and the relevant information is encoded in complex spatial patterns rather than in individual pixels. Second, the universe of possible concept labels is combinatorially vast.

We address the first problem using the OpenCLIP model. The image encoder E^{img} maps each view to a vector $z_{rt}^{img} \in \mathbb{R}^{1024}$, compressing high-dimensional pixel information into a lower-dimensional representation aligned with natural language semantics. These embeddings are averaged at the hexagon-year level and used as our primary measure of neighborhood appearance, entering directly into the hedonic regressions in Section 4.

We address the second problem —interpretability— by constructing a *semantic dictio-*

nary: a finite set of M urban concepts described in natural language. Each concept w_m is encoded by the text encoder as $z_m^{txt} = E^{txt}(w_m) \in \mathbb{R}^{1024}$, and the M text embeddings are collected into the dictionary matrix. We then represent each image embedding as a sparse, non-negative linear combination of these dictionary elements, yielding an interpretable sparse coding of street-view images. The full set of urban concepts is available in table 22.

$$O = \left[z_1^{txt} \ \cdots \ z_M^{txt} \right] \in \mathbb{R}^{1024 \times M}. \tag{2}$$

Because image and text embeddings share the same latent space, and that the embedding space is approximately additive, we decompose image embeddings into concept embeddings. The dictionary thus provides a coordinate system in which image embeddings can be expressed in terms of interpretable urban concepts.

For each image embedding z_{rt}^{img}, we seek a sparse, non-negative representation in terms of this dictionary:

$$\hat{\beta}_{rt} = \arg\min_{\beta} \ \frac{1}{2} \| z_{rt}^{img} - O\beta \|_2^2 + \lambda \|\beta\|_1 \quad \text{s.t. } \beta \geq 0 \tag{3}$$

where $\lambda > 0$ controls the overall degree of regularization. Non-negativity ensures that the decomposition is interpretable as a mixture: a positive weight indicates that a concept is present in the image, while zero indicates absence. Because semantic objects are often correlated, we also estimate a sparse group LASSO (Simon et al., 2013), using object categories to define groups. This approach trades off a group-level penalty, which encourages selection of entire semantic categories, and an individual-level penalty, which selects specific concepts within groups. We find that the solution converges toward a simple LASSO, reflecting the dominance of individual sparsity in the data. We select the values of λ via grid search to minimize reconstruction error, measured by the cosine similarity between observed and reconstructed embeddings, as detailed in Appendix E.

The resulting object scores $\hat{\beta}_{rt}$ map each image to a low-dimensional, interpretable summary of its visual content. Aggregated to the hexagon–year level, they serve two roles in what follows. In Section 4, they enter hedonic regressions alongside the raw embeddings, providing an interpretable decomposition of the price effects of visual neighbor-

hood characteristics. In Section 5, the model shows that the hedonic coefficients on these object scores recover the attention weights that households optimally assign to each amenity dimension in equilibrium. This is to say that the hedonic regressions are not meant to provide estimates of causal amenity values but descriptive moments that the structural model interprets.

4 Visual Embeddings and Housing Prices

4.1 Specification

The hedonic regressions in this section serve as a validation of our visual measures and as the source of the reduced-form moments that the structural model of Section 5 interprets. We regress log price per square meter on embedding vectors together with unit-level controls. The baseline specification is

$$\log p_{ihlt} = \alpha + X'_{it}\beta + Z'_{ht}\gamma + \delta_l + \tau_t + \varepsilon_{ihlt} \tag{4}$$

where i indexes the transaction, h the hexagon, l the neighborhood, and t the year. X_{it} is a vector of unit attributes; Z_{ht} is the hexagon–year embedding vector; δ_l and τ_t are neighborhood and year fixed effects respectively; and ε_{ihlt} is an idiosyncratic error term. Neighborhood fixed effects are defined using Uber H3 hexagons at resolution 7, corresponding to a resolution of about 2 km. Standard errors are clustered at the hexagon level in all specifications.

4.2 Explanatory Power of Embeddings

Table 5 reports adjusted R^2 values across eight specifications. Columns (1)–(4) progressively add unit attributes X_{it} and embedding measures Z_{ht} to a baseline with year fixed effects only. Columns (5)-(8) replicate this progression conditional on neighborhood fixed effects.

Figure 5: Hedonic regressions: adjusted R^2.

	log(Price sqm.)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Attributes	–	✓	–	✓	–	✓	–	✓
Embeddings	–	–	✓	✓	–	–	✓	✓
Year FE	✓	✓	✓	✓	✓	✓	✓	✓
Neighborhood FE	–	–	–	–	✓	✓	✓	✓
Significants embed.	–	–	449/802	429/802	–	–	282/802	277/802
Adjusted R^2	0.0193	0.0974	0.2600	0.2770	0.2790	0.2990	0.2980	0.3150
Observations	601,220	601,220	601,220	601,220	601,220	601,220	601,220	601,220

Notes: Standard errors clustered at the hexagon level.

The embeddings alone (column 3) achieve an adjusted R^2 of 0.2604, substantially more than unit attributes alone (column 2, 0.0974). Their joint inclusion (column 4) reaches 0.2767, indicating that the two sets of controls are complementary rather than substitutes: unit attributes capture property-specific characteristics, while embeddings capture the surrounding visual environment. The incremental R^2 gain from adding embeddings to unit attributes greatly exceeds the gain from adding unit attributes to the baseline, confirming that the visual neighborhood is quantitatively more important for prices than the standard set of property characteristics, conditional on size and year.

The specification that includes neighborhood fixed effects including both attributes and embeddings (column 8, 0.3152) achieves a higher adjusted R^2 than specifications including either unit attributes or embeddings alone (columns 6 and 7, 0.2990 and 0.2975, respectively). The additional explanatory power from embeddings within neighborhoods is comparable to that from unit attributes. Besides, 274 out of 1,024 remain significant at the 10% level in the joint specification. This indicates that embeddings capture fine-grained spatial variation that is not fully absorbed by neighborhood fixed effects. Full regression coefficients are reported in Appendix 18.

4.3 Spatial Decay of Neighborhood Effects

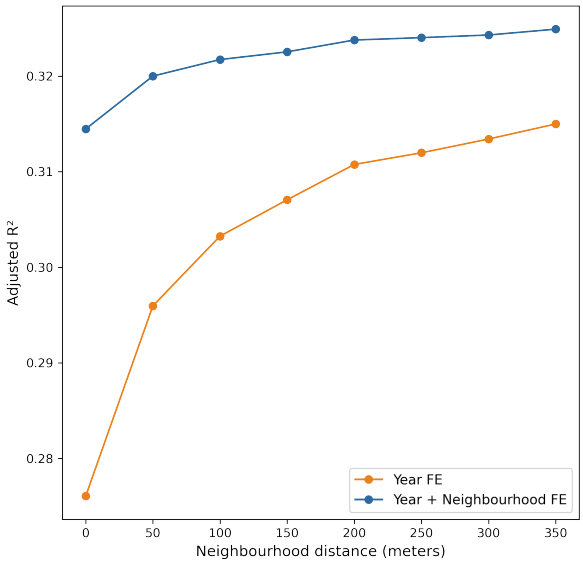
Neighborhood visual characteristics may affect housing prices not only at the transaction location but also in its vicinity. We examine this by augmenting specification 4 with average embeddings computed over successive concentric rings around each hexagon:

$$\bar{Z}_{ht}^{(k)} = \frac{1}{|\mathcal{N}_k(h)|} \sum_{j \in \mathcal{N}_k(h)} Z_{jt} \tag{5}$$

where $\mathcal{N}_k(h)$ denotes the set of hexagons at ring distance k from hexagon h, with each ring adding approximately 50 meters of radius. We add rings $k = 1, \ldots, 7$ separately to specifications (4) and report the incremental adjusted R^2 gain from each ring.

Figure 6 shows that explanatory power declines with distance: embeddings from nearby rings yield the largest gains in adjusted R^2, while gains diminish rapidly for more distant rings and become negligible beyond ring 6 (approximately 300 meters). The maximum gain from neighborhood embeddings is approximately 3 percentage points. We match this smooth and convex spatial decay profile at the scale of a few city blocks with an exponential spatial decay specification in the structural model of Section 5.

Figure 6: Spatial decay of neighborhood embedding effects on log price.



Notes: Each point reports the adjusted R^2 from a hedonic regression including embeddings of neighbouring hexagons within a given distance, defined in successive rings of approximately 50 meters. The two series differ in their fixed effects specification: year fixed effects only (orange) and year plus neighbourhood fixed effects (blue). The sample consists of 258,181 hexagons that have neighbours at all distances considered. Standard errors are clustered at the hexagonyear level.

4.4 Decomposing Embeddings into Semantic Objects

Semantic dictionary and decomposition. The 1,024-dimensional CLIP embeddings provide rich predictive power but resist direct interpretation. We address this by decomposing each image embedding as a sparse linear combination of a semantic dictionary of urban objects, using the sparse group LASSO estimator introduced in Section 3.3 (equation (3)). The dictionary covers categories such as vegetation, building types (Haussmannian, historical, modern), public transport infrastructure, walkability features, and nightlife establishments, among others. The resulting object scores $\hat{\beta}_{rt}$ measure the intensity of each urban concept in the visual environment of a given location and year.

Figure 7 illustrates the decomposition for three images with contrasting visual content. Notre-Dame Cathedral receives high scores for religious buildings, statues, and trees, all of which are visually prominent. A busy street in a central neighbourhood scores highly on pedestrians, Haussmannian architecture, cobblestone streets, and café. A street scene with visible signs of urban precarity scores highly on makeshift shelters and degraded infrastructure. The decomposition is interpretable and semantically faithful across visually distinct environments.

Validation. To validate the semantic object measures, we examine whether their spatial distribution across Paris hexagons aligns with independent geographic data. Appendix E presents full heatmaps, comparisons with OpenStreetMap, and a two-dimensional UMAP (Uniform Manifold Approximation and Projection, McInnes et al., 2018) projection of the embedding space, where distances represent visual similarity between hexagons. Vegetation scores are elevated along tree-lined boulevards and near major parks, and additionally recover street-level tree cover along the Seine embankments that is absent from the administrative green space data — a form of measurement that standard geographic sources do not provide. Haussmannian building scores are densest in the inner *arrondissements* and along the Haussmann-era boulevards; walkability scores peak in Montmartre and the Marais. In the UMAP projection, hexagons within the same urban environment cluster together, while visually similar neighborhoods across administrative boundaries also lie close in embedding space, confirming that the representations encode genuine urban characteristics rather than spatial fixed effects.

Figure 7: Sparse group LASSO decomposition: selected examples.



Notes: The top panel shows the Street View image; the bottom panel shows the object scores $\hat{\beta}_i$. Additional examples are shown in Appendix E.

4.5 Hedonic Evidence on Object Scores

Having established that the object scores recover interpretable urban characteristics, we use them as regressors in the hedonic specification to assess the price premia associated with distinct amenity dimensions. Table 1 reports results with and without neighborhood fixed effects for the 20 *arrondissements* of Paris.

Most coefficients have the expected sign and are precisely estimated. A one-standard-deviation increase in the Haussmannian score is associated with a 7.4 percentage point price premium without neighborhood fixed effects (column 1), falling to 1.6 points within *neighborhood* (column 2), indicating that Haussmannian character is correlated with other characteristics which neighborhood fixed effects account for. Walkability, retail, and outlet scores all carry positive and significant premia. Dirtiness is signif-

Table 1: Hedonic regressions on semantic object scores

	log(Price sqm.)			log(Price sqm.)	
	(9)	(10)		(9)	(10)
Building			*Public services*		
Commercial building	0.0227***	-0.0045	Administrative building	0.0178***	0.0031
Derelict building	-0.0285***	-0.0116***	Cemetery	-0.0053	0.0049*
Haussmannian building	0.0739***	0.0158***	Education building	-0.0015	0.0017
Heritage site monument	-0.0026	-0.0018	Police / fire station	-0.0048*	-0.0008
High-rise building	-0.0458***	-0.0273***	Post office	-0.0022	-0.0001
Historical building	0.0145***	0.0087***	Public transport	-0.0058*	0.0019
Industrial building	-0.0661***	-0.0273***	*Retail*		
Modern building	0.0086**	0.0018	Bakery	0.0065***	-0.0013
Religious building	0.0022	0.0018	Clothing shop	0.0199***	0.0012
Residential building	0.0174***	0.0059***	Cultural retail	0.0151***	0.0029***
Cycling			Florist shop	0.0032***	0.0011*
Bicycle	0.0162***	0.0022	Grocery store	-0.0009	-0.0010
Cycling lane	0.0611***	0.0127***	Health retail	0.0021*	0.0005
Electric scooter	0.0128***	0.0035*	Luxury retail	0.0252***	0.0073***
Greenery			Market stall	0.0030	0.0001
Other greenery	-0.0199***	-0.0030	Shopping mall	-0.0034	-0.0057*
Trees	-0.0336**	-0.0078	Speciality retail (food)	0.0081***	0.0024**
Hospitality			Speciality retail (other)	-0.0165***	-0.0030
Bar / cafe	0.0199***	0.0039***	*Services*		
Fast food	-0.0240***	-0.0054**	Beauty services	0.0005	-0.0001
Hotel	0.0027*	-0.0003	Financial services	0.0004	0.0024**
Restaurant	0.0021**	-0.0001	Health services	0.0062***	0.0011
Infrastructure			Laundry services	-0.0014	0.0005
Bridge	-0.0146	-0.0114	Travel agency	0.0112***	0.0005
Exterior staircase	-0.0105***	0.0017	*Street furniture*		
Retaining wall	-0.0070	-0.0010	Bench	0.0034	-0.0001
Underpass	-0.0129	-0.0080*	Fence	-0.0277***	-0.0081**
Item			Panel poster	0.0032	-0.0003
Statue	0.0025	0.0015	Public toilet	0.0112***	0.0026**
Water fountain	0.0110***	0.0030*	Streetlight	-0.0104**	0.0003
Outlet			Traffic signage	-0.0311***	-0.0011
Cinema / theatre / concert hall	0.0018	-0.0002	Trash can	-0.0012	-0.0003
Museum / gallery / exhibition	0.0050*	0.0005	*Traffic*		
Sports facility	-0.0156***	-0.0057*	Road	-0.0455***	-0.0060
Pedestrian			Vehicle	-0.0066*	-0.0001
Cobblestone street	0.0133***	0.0039***	*View*		
Pavement	-0.0093***	-0.0022	Construction	-0.0053	-0.0010
Pedestrian	0.0073***	0.0026	Countryside / field crops	-0.1380***	-0.0270***
Pedestrian crossing	-0.0026	0.0008	Water (river / canal)	0.0127*	0.0183***
...			*Other*		
			Graffiti	0.0166***	0.0012
			Homeless	0.0021	-0.0026
			Trash	-0.0094*	-0.0038
Attributes	✓	✓		✓	✓
Year FE	✓	✓		✓	✓
Neighborhood FE	–	✓		–	✓
Adjusted R^2	0.157	0.302		0.157	0.302
Within R^2	–	0.0650		–	0.0650
Observations	601,220	601,220		601,220	601,220

Notes: The dependent variable is the logarithm of price per square meter. Standard errors are clustered at the hexagon level. For readability, we report only coefficients that are statistically significant in at least one of the two specifications.

icantly negative without fixed effects but becomes insignificant within *neighborhood*, suggesting it proxies for broader quality gradients rather than capturing fine-grained variation. In the structural model of Section 5, these coefficients recover the optimal

attention weights that households assign to each amenity dimension in equilibrium: the hedonic regression is the reduced-form counterpart of the equilibrium pricing equation derived there.

4.6 Prediction of policy-induced change (in progress).

As a further validation of the predictive content of our representations, we examine whether written descriptions of planned urban interventions can be used to predict the direction of change in observed neighborhood embeddings. Paris implemented a series of neighborhood greening initiatives over our sample period, and for each treated area, planning documents describe the intended intervention in natural language, for instance the installation of street trees, planted facades, urban gardens, and so on. We encode these verbal descriptions using the CLIP text encoder and project the resulting text embedding onto the semantic dictionary to obtain a predicted change in object intensity scores: a greening proposal should predict an increase in the greenness concept score $\hat{\beta}_{rt,\text{green}}$ and leave unrelated scores such as Haussmannian architecture or nightlife unchanged. We then compare these predicted changes to the actual changes in hexagon-level object scores between the Street View vintage before and after the intervention. Alignment between predicted and actual intensity changes—measured by their correlation across treated hexagons, relative to a placebo distribution from untreated hexagons over the same period—validates our ability to translate written policy proposals into predicted visual changes. A positive and significant alignment would confirm that the representations support out-of-sample prediction of the visual consequences of planned interventions, which is precisely what the counterfactual exercises of Section 8 require.

5 Limited Attention in a Quantitative Urban Model

The model embeds the measured amenities of Section 4.4 into a quantitative spatial equilibrium framework following Ahlfeldt et al. (2015), augmented with a limited attention mechanism following Gabaix (2014). The central insight is that households face a dimensionality problem: Paris has 397,146 hexagonal cells, each characterized by a large number of potentially relevant visual concepts like vegetation, building types, walkability features, public transport infrastructure, nightlife, and many others captured by our

semantic dictionary. Attending to all concepts at all locations is cognitively infeasible. Households therefore select a sparse subset of the feature space on which to base their location choices, a decision that is itself endogenous to the distribution of features and their true impact on welfare. The same households whose attention choices generate the hedonic coefficients in Section 4 are the agents whose equilibrium location choices determine prices, wages, and population shares. Making this connection explicit allows us to use observed prices and sorting patterns to identify the attention-weighted amenity values $m_k^* \gamma_k$ as the product of optimal attention weights and intrinsic utility effects and to perform counterfactuals that are internally consistent with the spatial equilibrium.

Note that separately identifying m_k^* and γ_k requires additional variation beyond equilibrium prices and quantities. One approach is to use experimental variation that directly manipulates the salience of specific features in the decision environment. By varying which attributes are made more visually or cognitively prominent, one can distinguish between features that households do not value and those that are simply not attended to in equilibrium. We are working on this addition.

The model delivers three key objects. First, an equilibrium pricing equation that links housing prices to amenity features weighted by attention, thus providing the structural foundation for the hedonic regressions of Section 4 and showing that the LASSO coefficients recovered there are proportional to optimal attention weights. Second, location choice probabilities under rational inattention that determine equilibrium population shares across space. Third, a mapping from exogenous changes in visual features to equilibrium changes in prices, wages, and sorting patterns, enabling the counterfactual exercises of Section 8.

5.1 Setup

There are J locations indexed by $i = 1, \ldots, J$ for residence and $j = 1, \ldots, J$ for workplace. A mass N of ex ante identical individuals choose a residence-workplace pair (i, j). Conditional on choosing (i, j), individuals consume a composite good c and housing ℓ to obtain utility

$$U_{i,j}(c, \ell) = \frac{1}{\beta^\beta (1-\beta)^{1-\beta}} z_{i,j} \frac{B_i}{d_{i,j}} c^\beta \ell^{1-\beta}, \tag{6}$$

where $d_{i,j} > 0$ denotes iceberg commuting costs between residence i and workplace j, $\beta \in (0,1)$ is the Cobb–Douglas consumption share, $B_i = \widetilde{B}(\mathbf{b}_i)$ is a composite amenity index, and $z_{i,j}$ is an idiosyncratic taste draw. Residential amenities depend on a K-dimensional vector $\mathbf{b}_i = (b_{i,1}, \ldots, b_{i,K})'$, where each component $b_{i,k} \in [\underline{b}_k, \overline{b}_k]$ measures the level of amenity k at location i. In our empirical application, $\mathbf{b}_i$ corresponds to the vector of semantic object scores $\hat{\boldsymbol{\beta}}_i$ recovered by the sparse group LASSO decomposition of Section 4.4, so that each component $b_{i,k}$ measures the presence of urban concept k, i.e. vegetation, Haussmannian architecture, walkability, etc., at location i, and $K = d_o$ is the number of concepts in the semantic dictionary.

Idiosyncratic taste draws $z_{i,j}$ are independently distributed Fréchet:

$$F_{i,j}(z) = \exp\left(-T_i E_j \, z^{-\varepsilon}\right), \qquad \varepsilon > 1, \tag{7}$$

where $T_i > 0$ and $E_j > 0$ are exogenous residential and workplace fundamentals that shift the mass of agents selecting each location in equilibrium, and ε is the Fréchet shape parameter governing the dispersion of idiosyncratic tastes. The larger ε, the more homogeneous are preferences and the more responsive are location shares to systematic utility differences; as $\varepsilon \to \infty$, all agents concentrate at the highest-utility location, whereas as $\varepsilon \to 1$ location choices become insensitive to systematic utility. We calibrate ε from the gravity equation for commuting flows, using the estimate of Ahlfeldt et al. (2015) of approximately 6 as a benchmark.

The composite good is freely traded and its price is normalized to one. The budget constraint is

$$c + Q_i \ell = w_j, \tag{8}$$

where w_j is the wage at workplace j and Q_i is the residential price of floor space at location i.

5.2 Limited Attention to Amenities

Households face a high-dimensional problem. Each location is characterized by a large number of potentially relevant urban concepts (e.g. vegetation, building types, walkability, public transport, nightlife, etc.) captured by our semantic dictionary, and there are tens of thousands of locations to evaluate. A fully attentive household would eval-

uate all features at all locations before choosing where to live, which is cognitively infeasible. We model this by allowing households to choose, prior to observing their idiosyncratic taste draws $z_{i,j}$, a vector of attention weights $\mathbf{m} = (m_1, \ldots, m_K)' \in [0,1]^K$. Under attention weights m, households evaluate locations using a perceived amenity composite $B(\mathbf{b}_i, \mathbf{m})$, while the true amenity index is given by $\widetilde{B}(\mathbf{b}_i)$. We impose that full attention recovers the true amenity index, $B(\mathbf{b}_i, \mathbf{1}_K) = \widetilde{B}(\mathbf{b}_i)$. Perceived indirect utility is

$$V_{i,j}(\mathbf{m}) = z_{i,j}\, v_{i,j}(\mathbf{m}), \qquad v_{i,j}(\mathbf{m}) = \frac{B(\mathbf{b}_i, \mathbf{m})\, w_j}{d_{i,j} Q_i^{1-\beta}}. \tag{9}$$

When $m_k = 0$, feature k does not enter perceived utility; when $m_k = 1$, it enters fully. We impose the following properties on the aggregate amenity function $B(\mathbf{b}, \mathbf{m})$.

First, attention operates as a selection mechanism: if the household does not attend to a subset of amenities, those amenities do not affect perceived utility. Formally, if $m_k = 0$ for some k, then feature k does not enter $B(\mathbf{b}, \mathbf{m})$.

Second, the aggregate amenity is increasing in feature levels: for any k,

$$\frac{\partial B(\mathbf{b}, \mathbf{m})}{\partial b_k} > 0.$$

Third, the aggregate amenity is increasing in attention weights: for any k, $\partial B(\mathbf{b}, \mathbf{m})/\partial m_k \geq 0$, with strict inequality when the amenity is present above its baseline level. Paying more attention to a feature increases its contribution to perceived utility.

Fourth, attention and feature levels are complementary:

$$\frac{\partial^2 B(\mathbf{b}, \mathbf{m})}{\partial m_k \partial b_k} \geq 0,$$

so that the marginal value of attention is higher in locations where the feature is more strongly present.

Finally, we impose mild regularity conditions ensuring that $B(\mathbf{b}, \mathbf{m})$ and its derivatives are bounded. These conditions guarantee that the marginal value of attention can be uniformly bounded, which is sufficient for the existence of corner solutions in the attention problem.

Appendix G.1 shows that these properties are satisfied by the Cobb–Douglas aggregator

$B(\mathbf{b}, \mathbf{m}) = b_1^{m_1\gamma_1} \cdots b_K^{m_K\gamma_K}$ and by the exponential aggregator $B(\mathbf{b}, \mathbf{m}) = \exp(\sum_k m_k\gamma_k b_k)$, both of which arise naturally from log-utility specifications.

Timing and Choices The model has four stages. In the first stage, households choose attention weights $\mathbf{m}$ to maximize expected utility net of attention costs, before observing idiosyncratic draws. In the second stage, taste draws $z_{i,j}$ are realized. In the third stage, households choose the residence-workplace pair (i, j) that maximizes perceived utility $V_{i,j}(\mathbf{m})$. In the fourth stage, households obtain their true utility at the chosen location.

The timing of attention choice *before* location choice is natural for residential decisions, which are infrequent and deliberate: households commit to a strategy for how to evaluate neighborhoods before undertaking the search that precedes any given move. This contrasts with settings where attention is allocated *after* observing a subset of options; here attention is a fixed decision rule rather than a case-specific response.

Location Choice Probabilities Under Attention Given attention weights $\mathbf{m}$ and Fréchet taste draws, the probability that a household chooses residence-workplace pair (i, j) is

$$\pi_{i,j}(\mathbf{m}) = \frac{T_i E_j \, v_{i,j}(\mathbf{m})^{\varepsilon}}{\sum_{r,s=1}^{J} T_r E_s \, v_{r,s}(\mathbf{m})^{\varepsilon}}. \tag{10}$$

Under perfect attention ($\mathbf{m} = \mathbf{1}_K$), this reduces to the standard Ahlfeldt et al. (2015) expression, with $\pi_{i,j}(\mathbf{1}_K) \equiv \pi_{i,j}$. Under inattention, households distort the systematic utility of residential locations through $B(\mathbf{b}_i, \mathbf{m})$, which in turn changes the unconditional probability of selecting residence-workplace pair (i, j).

A key property is that conditional commuting probabilities, i.e. the probability of choosing workplace j given residence i, do *not* depend on attention:

$$\pi_{i,j|i} = \frac{E_j (w_j / d_{i,j})^{\varepsilon}}{\sum_{s=1}^{J} E_s (w_s / d_{i,s})^{\varepsilon}}. \tag{11}$$

The intuition is that amenity terms are multiplicative in utility and cancel from the ratio when conditioning on the residence. Attention therefore affects the spatial distribution of residents across locations, but not the workplace choices of residents conditional on where they live. As a result, expected wages conditional on residence are invariant to attention:

$$E(w_j \mid i) = \sum_{j=1}^{J} \pi_{i,j|i}\, w_j. \tag{12}$$

This decoupling simplifies both the equilibrium characterization and the estimation of commuting costs. In particular, it implies that the Fréchet shape parameter ε can be identified from conditional commuting flows without being confounded by attention.

Expected Utility Under Inattention Since attention is chosen before taste draws are realized, households evaluate their attention strategy by computing expected *true* utility—evaluated under perfect attention—over the distribution of location choices that their attention strategy induces. Formally,

$$V(\mathbf{m}) = E_{\mathbf{z}|v(\mathbf{m})}\left[V_{i(\mathbf{z},\mathbf{m}),\, j(\mathbf{z},\mathbf{m})}(\mathbf{1}_K)\right], \tag{13}$$

where $v(\mathbf{m}) = (v_{1,1}(\mathbf{m}), \ldots, v_{J,J}(\mathbf{m}))'$ collects systematic utilities under attention $\mathbf{m}$, and the expectation is over realized choices induced by perceived utilities but evaluated at true utilities. Appendix G.2 derives the closed-form expression

$$V(\mathbf{m}) = \widetilde{V}(\mathbf{1}_K) \sum_{i,j=1}^{J} \pi_{i,j}^{1/\varepsilon} \left[\pi_{i,j}(\mathbf{m})\right]^{1-1/\varepsilon}, \tag{14}$$

where $\widetilde{V}(\mathbf{1}_K) = \Gamma\big(\frac{\varepsilon-1}{\varepsilon}\big)\big(\sum_{r,s} T_r E_s v_{r,s}^{\varepsilon}\big)^{1/\varepsilon}$ is expected utility under perfect attention, as in Ahlfeldt et al. (2015).

This representation in equation 14 makes the welfare cost of inattention transparent: $V(\mathbf{m})$ is a weighted geometric mean of the true choice probabilities $\pi_{i,j}$ and the attention-distorted probabilities $\pi_{i,j}(\mathbf{m})$. The exponents $1/\varepsilon$ and $1 - 1/\varepsilon$ sum to one,

and by the inequality of power means, $V(\mathbf{m}) \leq \widetilde{V}(\mathbf{1}_K)$ with equality at $\mathbf{m} = \mathbf{1}_K$. Perfect attention maximizes expected utility; any deviation reduces it by inducing mistakes in location choices. The welfare cost of inattention is therefore measured by the gap $\widetilde{V}(\mathbf{1}_K) - V(\mathbf{m}^*)$, where $\mathbf{m}^*$ denotes the optimal attention vector defined in the next paragraph.

Appendix G.4 establishes the equivalent probability representation: the function $V^P(\theta) = \widetilde{V}(\mathbf{1}_K) \sum_{i,j} \pi_{i,j}^{1/\varepsilon} \theta_{i,j}^{1-1/\varepsilon}$ is maximized at $\theta_{i,j} = \pi_{i,j}$. The attention problem is thus isomorphic to choosing a location probability vector to be as close as possible to the true optimum, in a sense made precise by this objective.

Optimal Attention and the Kink Attention is costly. Let attention costs be linear: $\lambda \sum_k m_k$, so that each unit of attention paid to any feature costs $\lambda > 0$. The household chooses attention to maximize net expected utility:

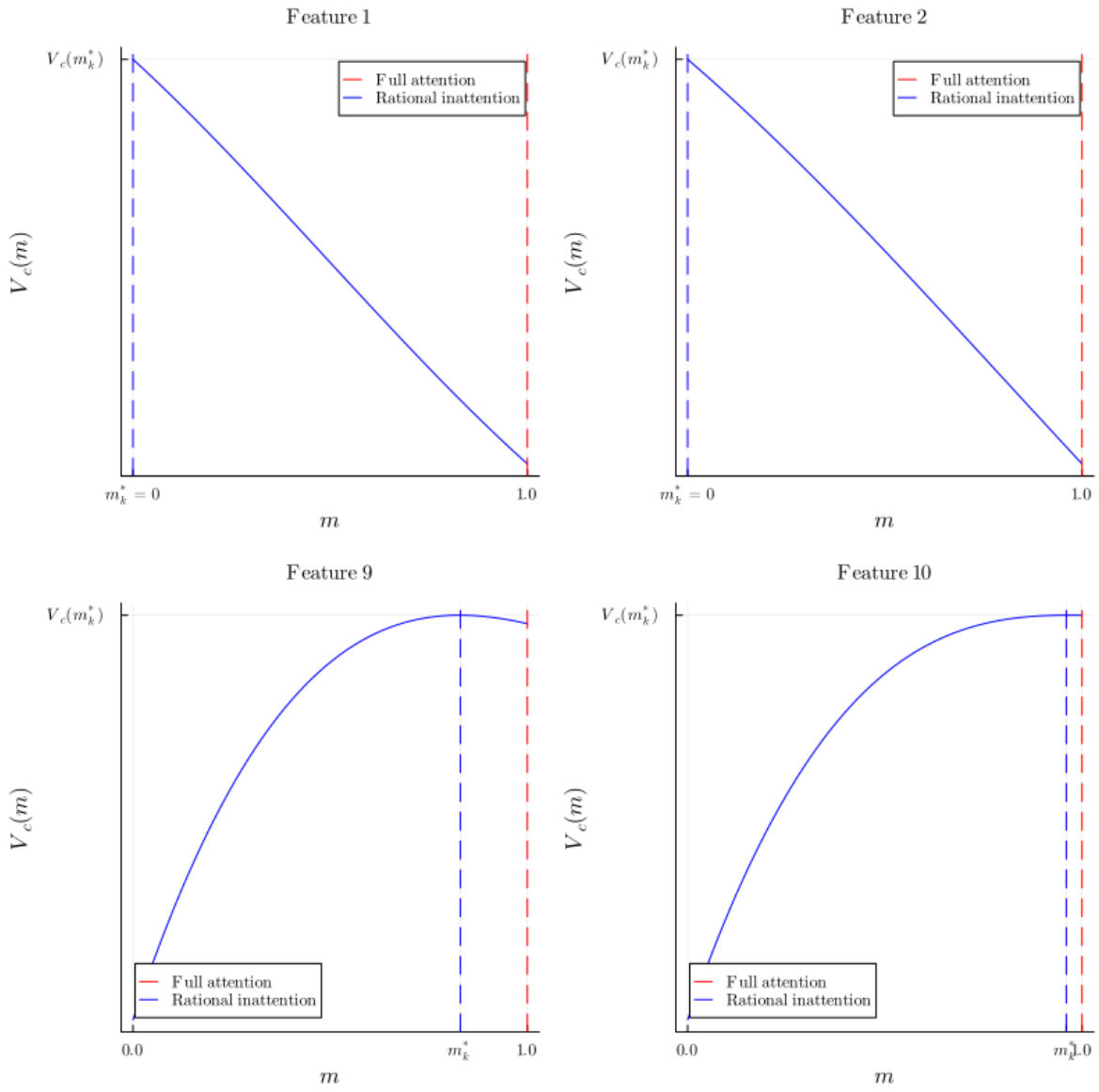
$$V_c(\mathbf{m}) = V(\mathbf{m}) - \lambda \sum_{k=1}^{K} m_k. \tag{15}$$

Since $V_c(\mathbf{m})$ is continuous on the compact set $[0,1]^K$, the Weierstrass theorem guarantees the existence of a maximum. Appendix G.3 shows that the marginal value of attending to feature k is

$$\frac{\partial V(\mathbf{m})}{\partial m_k} = (\varepsilon - 1)\widetilde{V}(\mathbf{1}_K) \sum_{i,j=1}^{J} \left[\frac{\partial \log B(\mathbf{b}_i, \mathbf{m})}{\partial m_k} - \sum_{r=1}^{J} \pi_r(\mathbf{m}) \frac{\partial \log B(\mathbf{b}_r, \mathbf{m})}{\partial m_k} \right] \pi_{i,j}^{1/\varepsilon} \pi_{i,j}(\mathbf{m})^{1-1/\varepsilon}, \tag{16}$$

where $\pi_r(\mathbf{m}) = \sum_j \pi_{r,j}(\mathbf{m})$ is the residential population share of location r. The expression in brackets is the deviation of the log marginal amenity $\partial \log B(\mathbf{b}_i, \mathbf{m})/\partial m_k$ at location i from its average across locations, weighted by residential population shares $\pi_r(\mathbf{m})$. The full expression sums this deviation against a weight $\pi_{i,j}^{1/\varepsilon} \pi_{i,j}(\mathbf{m})^{1-1/\varepsilon}$ that is large when location (i,j) is likely under both the true and distorted distributions. The marginal value of attending to feature k is therefore high when that feature varies substantially across locations the household is likely to choose, i.e. when attending to it would materially change which location is selected.

Figure 8: Optimal attention weights from simulations.



Notes: Each panel plots the objective function optimized by the rationally inattentive household:

$$V_c(\mathbf{m}) = V(\mathbf{m}) - \lambda \sum_k m_k$$

as a function of the candidate attention weight for a given feature, with true importance $m_k = 1.0$. This example created using a city with $K = 10$ features of increasing importance $\beta_1 < \beta_2 < \cdots < \beta_K$.
This optimization generates attention weights that (i) exhibit kinks at zero, (ii) increase in the true importance of the amenity in experienced utility, and (iii) decrease in attention cost.

The optimality condition for an interior solution is $\partial V(\mathbf{m}^*)/\partial m_k = \lambda, \forall k = 1, 2, \ldots, K$. For features with low spatial variance or low intrinsic importance, the marginal values of

attention are small and can be bounded uniformly by a value close to zero. This implies that $\partial V(\mathbf{m})/\partial m_k < \lambda$ for all $m_k \in [0,1]$, so that the optimal solution is the corner $m_k^* = 0$. This generates the kink at zero that characterizes LASSO-type solutions, as confirmed by simulation reported in Figure 8. The simulation uses 10 features ranked by true importance ($\beta_k = k/10$ for $k = 1, \ldots, 10$), 100 locations, and 10,000 households. Features with lower true importance β_k receive lower attention weights m_k^*, and features below a threshold receive zero attention. The attention weight is decreasing in the attention cost λ.

Uniqueness and the Role of ε A natural question is whether the attention problem admits a unique interior solution. Differentiating equation 16 a second time yields $\partial^2 V(\mathbf{m})/\partial m_k^2$ as shown in Appendix G.3. Evaluated at $\mathbf{m} = \mathbf{1}_K$, the first-order condition simplifies: the cross-sectional mean of $\partial \log B/\partial m_k$ cancels, and the marginal value of attention at the full-attention point equals zero, i.e. $\partial V(\mathbf{1}_K)/\partial m_k = 0$. This confirms that the interior of $[0,1]^K$ at $\mathbf{m} = \mathbf{1}_K$ is a critical point of $V(\mathbf{m})$, not of $V_c(\mathbf{m})$. The sign of $\partial^2 V/\partial m_k^2$ determines the local curvature. For large values of ε, consistent with the estimate of $\varepsilon \approx 6$ from Ahlfeldt et al. (2015), the second derivative is negative, so $V(\mathbf{m})$ is locally concave in m_k and the net utility $V_c(\mathbf{m})$ has a unique interior maximizer, provided the attention cost λ places the first-order condition in the interior. We proceed under the assumption that ε is large enough to guarantee this, which we verify numerically in the quantitative analysis.

5.3 The Attention–LASSO Correspondence

Under the exponential aggregator $B(\mathbf{b}_i, \mathbf{m}) = \exp(\sum_k m_k \gamma_k b_{i,k})$, so that $\log B(\mathbf{b}_i, \mathbf{m}) = \sum_k m_k \gamma_k b_{i,k}$, the marginal value of attending to feature k simplifies to

$$\frac{\partial V(\mathbf{m})}{\partial m_k} \propto (\varepsilon - 1)\, \widetilde{V}(\mathbf{1}_K) \cdot \gamma_k\, \widehat{\mathrm{Cov}}(b_{i,k},\, \log Q_i)\,, \tag{17}$$

where $\widehat{\mathrm{Cov}}$ denotes the cross-sectional covariance between the semantic object intensity and the log housing price, weighted by the distorted location probabilities $\pi_{i,j}(m)$. The interior first-order condition $\partial V/\partial m_k = \lambda$ therefore equates the covariance of feature k with housing prices to a common threshold λ: features whose spatial pattern

covaries strongly with prices receive positive attention, and those below the threshold receive zero attention. This is precisely the structure of the LASSO first-order condition (Tibshirani, 1996). In the unconstrained hedonic regression of $\log Q_i$ on $b_{i,k}$, the OLS estimator $\hat{\beta}_k^0$ is proportional to $\widehat{\text{Cov}}(b_{i,k}, \log Q_i)$.

The attention FOC $\partial V/\partial m_k = \lambda$ is thus isomorphic to the LASSO condition $|\hat{\beta}_k^0| \geq \lambda$: features whose hedonic coefficient exceeds the threshold λ receive positive attention weight $m_k^* > 0$, while those below it are set to zero. In this sense, λ corresponds simultaneously to the cognitive cost of attention and to the LASSO regularization penalty, and the hedonic regression with LASSO selection directly recovers the product $m_k^* \gamma_k$ up to the scaling $(\varepsilon - 1)\widetilde{V}(\mathbf{1}_K)$.

Two differences from the standard LASSO deserve emphasis. First, our LASSO is nonlinear: the covariance in equation (17) is weighted by the distorted probabilities $\pi_{i,j}(\mathbf{m})$, which themselves depend on the attention vector $\mathbf{m}$. In the standard LASSO, the covariance is computed once from the data; here, it is re-evaluated at each candidate attention weight because location shares change as attention changes. Second, the threshold λ is the cognitive cost of attention rather than a statistical regularization parameter, giving it an economic interpretation: it is identified from the share of features that are empirically selected relative to the cross-sectional variance of those features' hedonic prices.

Note that there is also a second LASSO in our empirical procedure: the *Sparse Group LASSO* used to decompose CLIP embeddings into semantic objects in Section 4.4. These two LASSOs play different roles. The first (the attention LASSO in this section) microfounds the selection of features in the hedonic regression as a consequence of rational inattention. The second (the embedding LASSO, Section 4.4) is a dimensionality reduction tool that maps raw CLIP embeddings into a sparse vector of semantic object intensities, measuring the extent to which each urban concept appears in the images of a given location. The two connect directly: these intensity scores become the feature levels $b_{i,k}$ that enter the attention model, so that the amenities households potentially attend to are the same objects recovered from the visual data.

5.4 Equilibrium

Production Production in workplace j is Cobb–Douglas, $Y_j = A_j H_{Mj}^{\alpha} L_{Mj}^{1-\alpha}$, where A_j is productivity, H_{Mj} is employment, and L_{Mj} is commercial floor space. Profit maximization and zero profits imply a commercial floor price $q_j = (1-\alpha)(\alpha/w_j)^{\alpha/(1-\alpha)} A_j^{1/(1-\alpha)}$. This block is the same as in Ahlfeldt et al. (2015); we abstract from firm inattention, which would be relevant for commercial real estate prices but is outside our current scope.

Market Clearing Let $H_{Ri}(\mathbf{m}^*) = \pi_i(\mathbf{m}^*)N$ denote the residential population in location i under optimal attention. Labor market clearing in workplace j then requires

$$H_{Mj} = \sum_i \pi_{i,j|i}\, H_{Ri}(\mathbf{m}^*).$$

The key feature is that attention affects where people live, through $\pi_i(\mathbf{m}^*)$, but not how they commute once resident, through $\pi_{i,j|i}$. This recursive structure enables identification of attention separately from commuting costs.

Let θ_i denote the share of floor space used commercially in location i. Residential land market clearing implies that total residential floor space $(1-\theta_i)L_i$ equals aggregate housing demand. This yields the residential floor price

$$Q_i = \frac{(1-\beta)\, E(w_j \mid i)\, \pi_i(\mathbf{m}^*)\, N}{(1-\theta_i)L_i}. \tag{18}$$

Substituting the residential population share $\pi_i(\mathbf{m}^*) = \Psi_i B(\mathbf{b}_i, \mathbf{m}^*)^{\varepsilon} / Q_i^{\varepsilon(1-\beta)}$ into equation 18 and solving yields the equilibrium pricing relationship:

$$\log Q_i = \frac{1}{1+\varepsilon\,(1-\beta)} \log\left[\Psi_i \frac{(1-\beta)N}{(1-\theta_i)L_i} E(w_j \mid i)\right] + \frac{\varepsilon}{1+\varepsilon\,(1-\beta)} \log B(\mathbf{b}_i, \mathbf{m}^*), \tag{19}$$

5.5 From Theory to Empirics

The model can be taken to the data in a sequence of steps, following the approach of Ahlfeldt et al. (2015). The equilibrium pricing equation (19) provides the main bridge from the model to the data. Its first term collects location fundamentals that vary at a broader spatial scale—including commuting access Ψ_i, residential fundamentals T_i, floor space supply, and expected wages—and can be absorbed by a location fixed effect δ_i in the empirical implementation. The second term captures the capitalization of attention-weighted amenities into housing prices.

Under the exponential aggregator $B(\mathbf{b}_i, \mathbf{m}) = \exp(\sum_k m_k \gamma_k b_{i,k})$, so that $\log B(\mathbf{b}_i, \mathbf{m}) = \sum_k m_k \gamma_k b_{i,k}$, equation (19) becomes

$$\log Q_i = \delta_i + \frac{\varepsilon}{1 + \varepsilon\,(1-\beta)} \sum_{k=1}^{K} m_k^* \gamma_k b_{i,k} + \varepsilon_i, \tag{20}$$

where the feature levels $b_{i,k}$ are the semantic object intensity scores recovered by the sparse group LASSO decomposition of Section 4.4. The coefficients $\tilde{\beta}_k^* \equiv m_k^* \gamma_k \varepsilon / \left(1 + \varepsilon\,(1-\beta)\right)$ are proportional to optimal attention-weighted amenity values. The hedonic regression therefore identifies the product $m_k^* \gamma_k$ up to a constant.

This mapping highlights a small set of sufficient objects that can be taken to the data. Three objects are sufficient for the baseline quantitative analysis. First, the Fréchet shape parameter ε, which governs the elasticity of commuting and residential location shares to systematic utility differences. Second, the optimal attention weights $\mathbf{m}^* = (m_1^*, \ldots, m_K^*)'$, which determine which amenity features are capitalized into prices and which are ignored. Third, the spatial decay parameter ρ, which governs how quickly neighborhood influences attenuate with distance (estimated in the hedonic regressions of Section 4).

The products $m_k^* \gamma_k$ are identified by the hedonic regression coefficients in equation 20: up to the aggregate of parameters $\left(1 + \varepsilon\,(1-\beta)\right)/\varepsilon$, the coefficient on $b_{i,k}$ recovers a value proportional to $m_k^* \gamma_k$. Separately identifying the attention weight m_k^* from the intrinsic utility weight γ_k requires additional variation beyond equilibrium prices, which we intend to derive from an experiment. The LASSO penalty λ in the hedonic regression can be interpreted as reflecting the cognitive cost of attention; its magnitude governs

how many features receive positive attention weight in equilibrium.

The following lemma summarizes the identification content of the model.

Lemma 1 (Identification of Attention-Weighted Amenity Values)**.** *Suppose that the amenity aggregator is exponential,* $B(\mathbf{b}_i, \mathbf{m}) = \exp(\sum_k m_k \gamma_k b_{i,k})$*, and that a location fixed effect absorbs* $\Psi_i T_i E(w_j \mid i)$*. Given observed housing prices* $\{Q_i\}$ *and feature vectors* $\{b_i\}$*, the hedonic regression* (20) *identifies the products* $m_k^* \gamma_k = \hat{\beta}_k \left[1 + \varepsilon \left(1 - \beta\right)\right] / \varepsilon$*, where* $\hat{\beta}_k$ *is the estimated coefficient on* $b_{i,k}$*. Features with* $\hat{\beta}_k = 0$ *receive zero attention-weighted amenity value:* $m_k^* \gamma_k = 0$*. Separate identification of* m_k^* *and* γ_k *requires knowledge of one of the two, which an experimental design is intended to provide in the future.*

A complete equilibrium definition and existence argument follow from the standard conditions in Ahlfeldt et al. (2015), modified by the replacement of the full-attention amenity term with $B(b_i, \mathbf{m}^*)$. For a given attention vector $\mathbf{m}^*$, the equilibrium in prices, wages, and population shares is uniquely determined, following the log-linear fixed point argument of Ahlfeldt et al. (2015).

Uniqueness of $\mathbf{m}^*$ itself depends on the global concavity of $V(\mathbf{m})$ in $\mathbf{m}$, which in turn depends on the sign of the second derivative derived in Appendix G.3. We treat uniqueness of the attention vector as a quantitative property of the calibrated model and verify it numerically for the parameter values used in the analysis.

6 Quantitative Analysis

In this section, we take the model of Section 5 to the data. Following Ahlfeldt et al. (2015) and Heblich et al. (2020), we proceed in a sequence of steps, each relying on a minimal set of additional assumptions. In Subsections 6.1 through 6.3, we implement this methodology and estimate the model using our Paris data. In Subsections 6.1 through 6.3, we estimate the key objects of the model using our Paris data. Subsection 6.4 then describes how we conduct counterfactual exercises.

6.1 Step 1: Commuting Probabilities and Fréchet Shape Parameter

In the first step, we estimate the Fréchet shape parameter ε from conditional commuting probabilities. The model implies a gravity equation for log bilateral commuting probability conditional on residence:

$$\log \pi_{i,j|i} = \mathcal{W}_j - \varepsilon \log d_{i,j} + e_{i,j}, \tag{21}$$

where $\mathcal{W}_j$ and $\mathcal{R}_i$ are workplace and residence fixed effects. Note that the unconditional residence-workplace probability $\pi_{i,j}(\mathbf{m}^*)$ also satisfies a gravity equation with origin and destination fixed effects, as shown in the model. However, the residential component of this equation depends on attention through $B(\mathbf{b}_i, \mathbf{m}^*)$ and housing prices, and is therefore not directly suitable for identification. Conditioning on residence eliminates this dependence, yielding equation (21).

We estimate this equation by OLS and by two-stage least squares, instrumenting bilateral commuting times with straight-line distances, following the identification strategy of Ahlfeldt et al. (2015). The composite estimate $-\varepsilon\phi$ identifies the product of the Fréchet shape parameter ε and the elasticity of commuting costs to travel time ϕ. We calibrate ε using the benchmark estimate from Ahlfeldt et al. (2015) of approximately 6 and report robustness to values in the range $[3, 9]$.

6.2 Step 2: Attention Weights and Hedonic Identification

In the second step, we estimate the attention weights $\mathbf{m}^*$ from the hedonic regression 20. This step directly connects the econometric results of Section 4 to the structural model: the LASSO-selected coefficients in the hedonic regression are proportional to optimal attention weights by Lemma 1. The penalty parameter λ is selected by cross-validation, providing a data-driven estimate of the cognitive cost of attention.

In practice, the attention-weighted amenity values $m_k^*\gamma_k$ are estimated at the hexagon–year level using the semantic object intensity scores $\hat{\beta}_{i,k}$ recovered by the sparse group LASSO decomposition of Section 4.4, so that each $b_{i,k}$ measures the intensity of urban concept k in the visual environment of location i.

6.3 Step 3: Baseline Equilibrium

In an third step, the model can be used to recover equilibrium wages and location fundamentals from observed data. Given the observed residential population shares $\{\pi_i\}$, conditional commuting probabilities $\{\pi_{i,j|i}\}$, floor prices $\{Q_i\}$, and model parameters $\{\beta, \varepsilon\}$, the land-market clearing condition 18 provides a system of equations in workplace wages $\{w_j\}$. Under standard regularity conditions, this system can be solved for equilibrium wages. Given wages, we recover expected wages conditional on residence from equation (12), and infer residential and workplace fundamentals $\{T_i, E_j\}$ as residuals.

6.4 Step 4: Counterfactual Exercises

The model provides a framework for evaluating counterfactual changes in neighborhood characteristics. Given a counterfactual amenity vector $\{b_{i,k,\tau}\}$ constructed using the CLIP space interpolation described in Section 8, equation (22) describes how changes in amenities translate into changes in equilibrium prices, wages, and residential sorting. This allows us to evaluate questions such as how housing prices and the residential distribution of Paris would respond to changes in the visual environment.

Given the estimated objects from the previous steps, we evaluate the impact of changes in neighborhood visual characteristics using a comparative statics approach analogous to the "exact hat algebra" of Dekle et al. (2007). Let t denote the baseline equilibrium and $\tau \neq t$ a counterfactual equilibrium. For any endogenous variable x, define its proportional change as $\hat{x} = x_\tau / x_t$. This notation allows us to express counterfactual changes in equilibrium outcomes without solving explicitly for all underlying fundamentals. Log-differentiating the equilibrium pricing relationship (19) yields the following expression in changes:

$$\log \hat{Q}_i = \frac{1}{1 + \varepsilon(1-\beta)} \left[\log \hat{\Psi}_i + \log \widehat{E(w_j \mid i)} \right] + \frac{\varepsilon}{1 + \varepsilon(1-\beta)} \log \widehat{B(\mathbf{b}_i, \mathbf{m}^*)}. \tag{22}$$

We consider two counterfactual exercises. In the first, attention weights are held constant at their estimated baseline values $\mathbf{m}^*$ and only the exogenous visual features $\hat{b}_{i,k}$ change. This is a valid first-order approximation for small changes in the feature dis-

tribution, and captures the short-run response when household search heuristics adjust slowly. In this case $\widehat{B(\mathbf{b}_i, \mathbf{m}^*)} = \exp(\sum_k m_k^* \gamma_k (b_{i,k,\tau} - b_{i,k,t}))$ can be constructed from observed changes in features and estimated attention-weighted amenity coefficients. Given known model parameters $\{\beta, \varepsilon, \rho\}$ and baseline equilibrium vectors $\{\mathbf{Q}, \boldsymbol{\pi}, \boldsymbol{w}\}$, the system (22) can be solved for changes in equilibrium wages $\hat{w}_j$, residential population shares $\hat{\pi}_i$, and floor prices $\hat{Q}_i$.

In the second exercise, households re-optimize attention weights at the counterfactual feature distribution $b_{i,k,\tau}$, capturing the long-run response when search heuristics fully adjust. This requires iterating between the attention optimization and the spatial equilibrium until both are mutually consistent at the counterfactual.

A key advantage of this approach is that we need not specify the full vector of location fundamentals (T_i, E_j, A_j); these are absorbed by the baseline equilibrium values and cancel from the comparative statics. The sufficient statistics for the counterfactual are the baseline commuting probabilities $\pi_{i,j|i}$, the baseline residential and workplace population shares, wages, and floor prices, the estimated attention weights $\mathbf{m}^*$, and the changes in exogenous features $\hat{b}_{i,k}$.

7 Equilibrium Simulations

This section illustrates the properties of the model using equilibrium simulations. We present (i) a definition of the equilibrium consistent with the model, (ii) the contraction algorithm used to estimate it, and (iii) simulation results for a stylized city with spatial heterogeneity in amenities and productivities. The simulations also allow attention weights to adjust endogenously.

Definition 1. *Given primitives* $\{\mathbf{T}, \mathbf{E}, \mathbf{A}, \mathbf{b}, \boldsymbol{\gamma}, \boldsymbol{\tau}, \mathbf{L}, \boldsymbol{\xi}\}$ *and parameters* $(\kappa, \varepsilon, \beta, \alpha, \lambda)$*, an equilibrium with rational inattention consists of attention weights* $\mathbf{m}$*, location choice probabilities* $\boldsymbol{\pi} = [\pi_{ij}]$*, residential and workplace shares* $(\boldsymbol{\pi}_R, \boldsymbol{\pi}_M)$*, wages* $\mathbf{w}$*, prices for residential and commercial floor space* $(\mathbf{Q}, \mathbf{q})$*, and land-use shares* $\boldsymbol{\theta}$ *such that:*

1. *Attention weights* $\mathbf{m}$ *maximize net expected utility* $V_c(\mathbf{m})$*.*
2. *Households choose residence-workplace pairs* (i, j) *to maximize perceived utility.*

3. Firms choose labor and floor space to maximize profits.

4. Labor and land markets clear.

Equilibrium Iteration with Endogenous Attention Weights We compute the equilibrium using an iterative algorithm in which attention weights adjust, affecting floor surface prices through the residential demand for floor space, and in turn influencing commercial demand for floor space and wages. This mechanisms shapes the equilibrium hedonic gradient, as illustrated in the results below.

The algorithm generates a sequences of attention weights $\mathbf{m}_t$, location choices $\boldsymbol{\pi}_t$, prices $\mathbf{Q}_t, \mathbf{q}_t$, wages $\mathbf{w}_t$, and land-use shares $\boldsymbol{\theta}_t$. At each iteration, attention weights are updated by maximizing expected utility net of cognitive attention costs, given current location choices, prices, wages and shares. Location choices, prices (through floor space demand and shares), and wages are then updated using the firm's first order conditions. We verify that both the land and labor market clear. In the numerical analysis, the algorithm uses a learning rate and converges at log-linear speed, consistent with a solution exhibiting multiplicity.

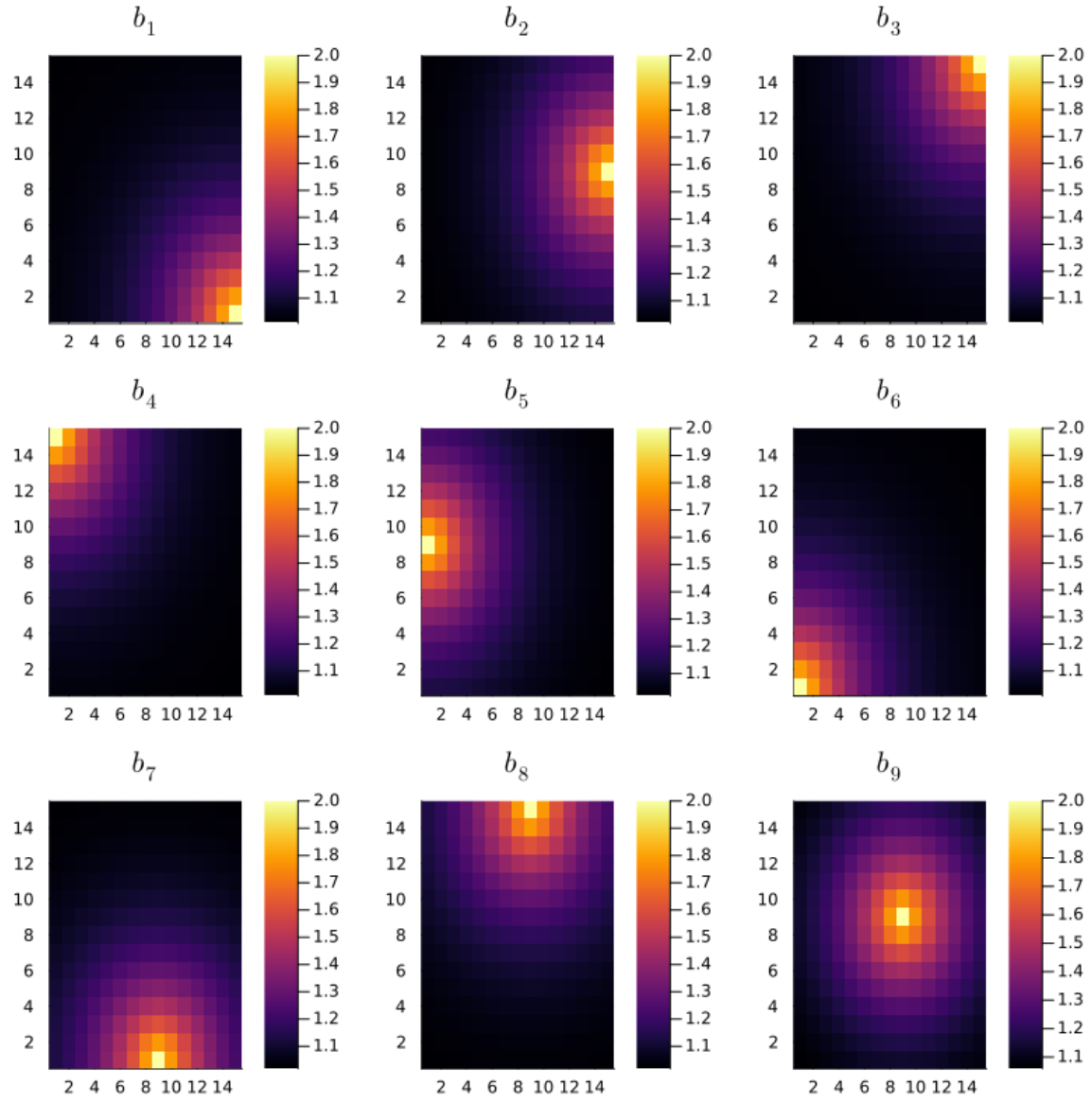
7.1 Equilibrium Results

Choice of Parameters We estimate the equilibrium with rational inattention in a gridded city, following the quantitative exercises of Redding and Rossi-Hansberg (2017). The parameters β, α, κ, and ε are set to consensus values in the literature: the preference for numeraire consumption $\beta = 1 - 0.25$, the marginal productivity of labor $\alpha = 1 - 0.20$, the gravity parameter $\kappa\varepsilon = 0.09$ and the dispersion of Fréchet unobservable utility $\varepsilon = 6.62$. The city consists of 15×15 locations that are heterogeneous in amenity characteristics $\mathbf{b} = [b_{ik}]$, with amenity dimensions ranked in increasing importance:

$$\gamma_1 < \gamma_2 < \cdots < \gamma_K$$

These multidimensional features b_{ik} are depicted in Figure 9. They show $K = 9$ features that are more important for utility (γ_K high) at the center and in the western part of the city than in the eastern part (where γ_1, γ_2, γ_3 are relatively small). The resulting

Figure 9: General Equilibrium Exercise – Exogenous Amenity Features $\mathbf{b} = [b_{ik}]$ in Space.



Notes: Residents choose attention weights $\mathbf{m} = [m_k]$ that determine their perceived bundle B_i of amenities. In this exercise, we use a Cobb–Douglas aggregator, $\mathbf{B}(\mathbf{b}, \mathbf{m}) = \Pi_k b_{ik}^{m_k \gamma_k}$. The amenity components are specified as $b_{ik} = b_k e^{-\varpi_k d_i^k}$, where d_i^k denotes distance feature-specific centers. Features are ordered from least relevant in the East, to the West, and finally the most relevant feature at the center of the city.

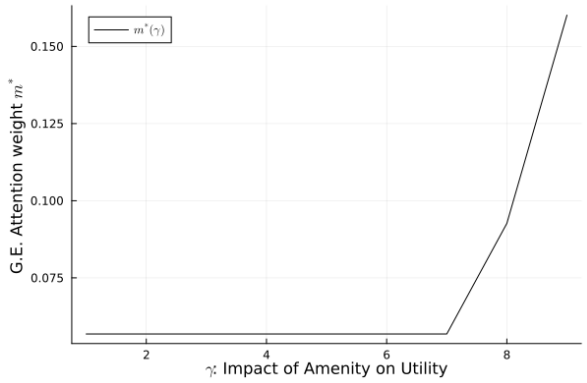
equilibrium suggests that endogenous attention amplifies the role of high-importance features in shaping the spatial distribution of population.

Results Figure 10 shows that the optimal attention weights estimated endogenously naturally follow the kinked shape of an *activation function* as in Gabaix (2014). In our setting, this activation function is determined in general equilibrium, as it depends on the equilibrium distribution of prices and attention weights.

$$\mathbf{m}^*(\boldsymbol{\gamma}; \boldsymbol{\pi}, \mathbf{w}, \mathbf{Q}, \mathbf{q}, \theta)$$

Equilibrium outcomes are plotted in Figure 11. Panel (a) presents the spatial distribution of population (residential location choices π_R), residential floor surface prices Q, and the share of floor surface $1 - \theta$ used for residential purposes. The equilibrium display a strong monocentricity, as agents put more weight on the central feature b_K than on the peripheral features b_1, b_2, b_{K-1} of this simulation.

Figure 10: General Equilibrium Exercise – Activation function as in Gabaix (2014)

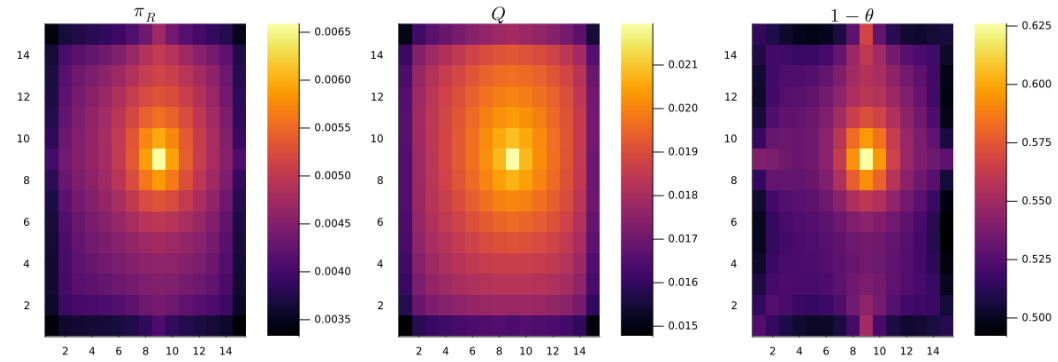


Notes: This graphs presents the optimal attention weights $m^*(\gamma)$ as a function of the impact γ_k of the feature on utility. These optimal attention weights are estimated at each step of the equilibrium iteration as the maximum of $V_c(\mathbf{m})$ given the spatial distribution of prices and wages.

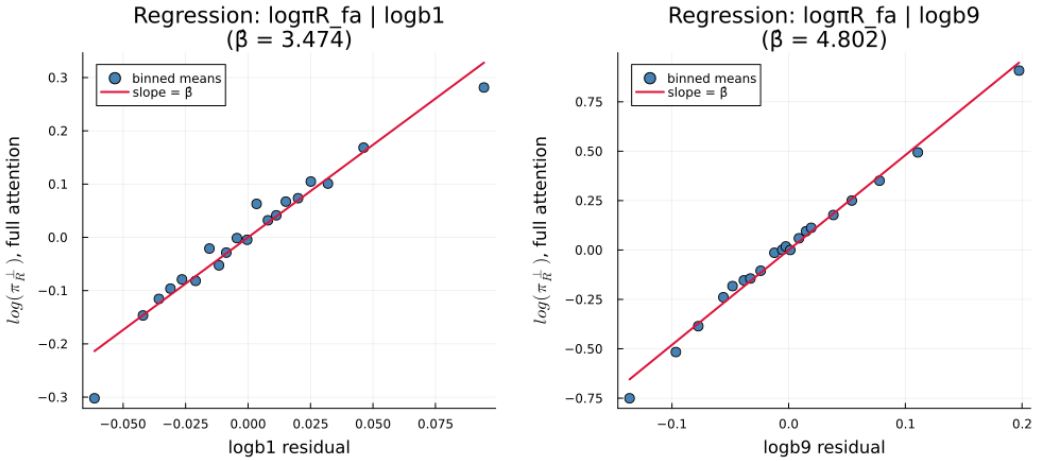
The attention weights affect equilibrium location choices. Panel (b) recomputes the equilibrium forcing full attention $\mathbf{m} = \mathbf{1}_K$, and runs regressions of equilibrium location choices on features 1 (the least important, peripheral) and 9 (central feature, most important in the simulation). Panel (c) estimates these regressions in the model with endogenous rational inattention m^*. The comparison of panels (b) and (c) shows that in the model with rational inattention, location choices are weakly related to feature 1 relative to feature 9. In the model with full attention, location choices are driven by all features in equilibrium.

Figure 11: General Equilibrium Exercise – Endogenous Equilibrium Outcomes.

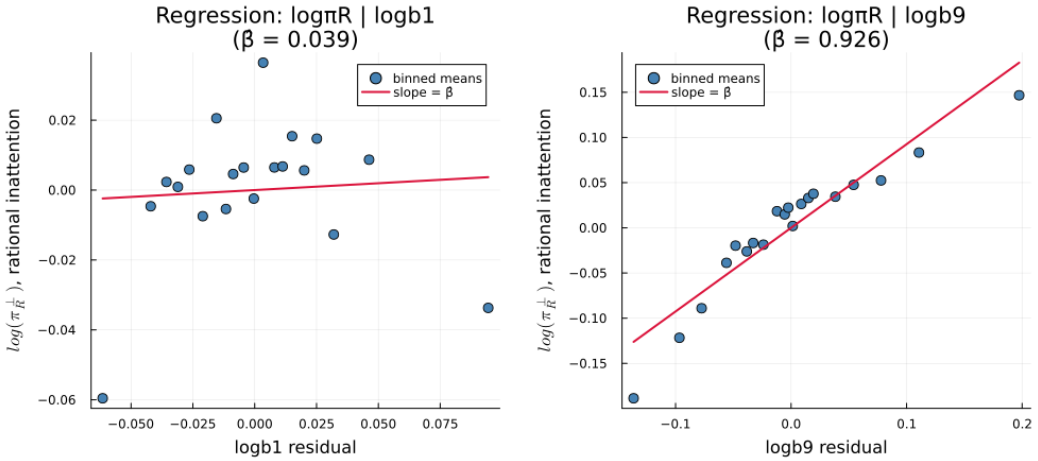
(a) Spatial Equilibrium, Rational Inattention



(b) Location Choices, Full Attention Equilibrium



(c) Location Choices, Rational Inattention Equilibrium



Notes: Panel (a) presents the spatial distribution of general equilibrium outcomes π_R, Q, $1-\theta$ in the model with rational inattention. Panel (b) presents the regressions of location choices in the equilibrium with full attention and panel (c) the same regressions for the equilibrium with rational inattention. In the model with full attention, equilibrium location choices are more steeply related to the features for both dimensions $k = 1$ and $k = 9$.

8 Counterfactuals

TBA

9 Conclusions

This paper opens the black box of neighborhood effects by replacing residual amenity terms with directly measured visual characteristics of the built environment. Using street-level imagery and representation learning, we construct high-dimensional embeddings of neighborhoods, decompose them into interpretable object scores, and show that these measures capture economically meaningful variation in local environments. Hedonic regressions confirm that visual features explain a substantial share of housing price variation and align with independent geographic information.

We embed these measures into a quantitative spatial model with limited attention. Households face a high-dimensional decision problem and optimally allocate attention across amenity dimensions, evaluating locations using a simplified representation of the environment. In equilibrium, only the features that receive attention are capitalized into housing prices and drive residential sorting. This provides a structural interpretation of hedonic coefficients: estimated amenity prices reflect both intrinsic utility weights and endogenous attention.

The model delivers a tractable mapping from changes in visual features to equilibrium outcomes. Holding attention fixed, this framework allows us to translate observed changes in neighborhood characteristics into predicted changes in prices and sorting patterns. This enables counterfactual analysis of the built environment based on directly observed, policy-relevant attributes, rather than composite residual amenity terms.

Our results highlight two broader insights. First, high-dimensional neighborhood characteristics can be measured in a systematic and interpretable way using modern representation learning tools. Second, limited attention plays a central role in shaping how these characteristics are valued: only a subset of features is incorporated into decision-making, and this selection governs both estimated amenity prices and equilibrium outcomes.

By combining visual data with a structural model of attention and spatial equilibrium, this paper provides a new approach to measuring, interpreting, and evaluating urban amenities.

References

Ahlfeldt, Gabriel M., Stephen J. Redding, Daniel M. Sturm, and Nikolaus Wolf, "The Economics of Density: Evidence from the Berlin Wall," *Econometrica*, 2015, *83* (6), 2127–2189.

Almagro, Milena, Manuel Domínguez, and Tomohiro Iino, "Location, Productivity, and Amenities," *Econometrica*, 2025.

Ang, Amanda, Daniel Angel, and Andrii Parkhomenko, "Amenities in Quantitative Spatial Models," Working Paper 2024.

Bach, Philipp, Victor Chernozhukov, Sven Klaassen, Martin Spindler, Jan Teichert-Kluge, and Suhas Vijaykumar, "Adventures in Demand Analysis Using Artificial Intelligence," Working Paper 2025.

Bajari, P., Z. Cen, V. Chernozhukov, M. Manukonda, S. Vijaykumar, J. Wang, R. Huerta, J. Li, L. Leng, G. Monokroussos, and S. Wang, "Hedonic prices and quality adjusted price indices powered by AI," *Journal of Econometrics*, 2025, *251*, 106052.

Bengio, Yoshua, Aaron Courville, and Pascal Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, *35* (8), 1798–1828.

Bhalla, Usha, Alex Oesterling, Suraj Srinivas, Flavio P. Calmon, and Himabindu Lakkaraju, "Interpreting CLIP with Sparse Linear Concept Embeddings," in "Advances in Neural Information Processing Systems" 2024.

Cafarella, Michael, Gabriel Ehrlich, Tian Gao, John C. Haltiwanger, Matthew D. Shapiro, and Laura Zhao, "Using Machine Learning to Construct Hedonic Price Indices," NBER Working Paper 2023.

Cherti, Mehdi, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev, "Reproducible Scaling Laws for Contrastive Language-Image Learning," in "Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)" June 2023, pp. 2818–2829.

Compiani, Giovanni, Ilya Morozov, and Stephan Seiler, "Demand Estimation with Text and Image Data," Working Paper 2023.

Dekle, Robert, Jonathan Eaton, and Samuel Kortum, "Unbalanced Trade," *American Economic Review*, 2007, *97* (2), 351–355.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An Image is Worth 16Œ16 Words: Transformers for Image Recognition at Scale," in "International Conference on Learning Representations" 2021.

Gabaix, Xavier, "A Sparsity-Based Model of Bounded Rationality," *Quarterly Journal of Economics*, 2014.

_ , Ralph S. J. Koijen, Robert Richmond, and Motohiro Yogo, "Asset Embeddings," Working Paper 2025.

Heblich, Stephan, Stephen J. Redding, and Daniel M. Sturm, "The Making of the Modern Metropolis: Evidence from London," *The Quarterly Journal of Economics*, 2020, *135* (4), 2059–2133.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," *Nature*, 2015, (521), 436–444.

Magnolfi, Lorenzo, Jonathon McClure, and Alan Sorensen, "Triplet Embeddings for Demand Estimation," *American Economic Journal: Microeconomics*, 2025, *17* (1), 282–307.

Matějka, Filip and Alisdair McKay, "Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model," *American Economic Review*, January 2015, *105* (1), 27298.

McInnes, Leland, John Healy, Nathaniel Saul, and Lukas Großberger, "UMAP: Uniform Manifold Approximation and Projection," *Journal of Open Source Software*, 2018, *3* (29), 861.

Naik, Nikhil, Scott Duke Kominers, Ramesh Raskar, Edward L. Glaeser, and César A. Hidalgo, "Computer vision uncovers predictors of physical urban change," *Proceedings of the National Academy of Sciences*, 2017, *114* (29), 7571–7576.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark et al., "Learning transferable visual models from natural language supervision," in "International conference on machine learning" PMLR 2021, pp. 8748–8763.

Ramesh, Aditya, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen, "Hierarchical Text-Conditional Image Generation with CLIP Latents," *arXiv preprint*, 2022.

Redding, Stephen J and Esteban Rossi-Hansberg, "Quantitative spatial economics," *Annual Review of Economics*, 2017, *9*, 21–58.

Schuhmann, Christoph, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev, "LAION-5B: an open large-scale dataset for training next generation image-text models," in "Proceedings of the 36th International Conference on Neural Information Processing Systems" NIPS '22 Curran Associates Inc. Red Hook, NY, USA 2022.

Simon, Noah, Jerome Friedman, Trevor Hastie, and Robert Tibshirani, "A Sparse-Group Lasso," *Journal of Computational and Graphical Statistics*, 2013, *22* (2), 231–245.

Sims, Christopher A., "Implications of Rational Inattention," *Journal of Monetary Economics*, 2003, *50* (3), 665–690.

Tibshirani, Robert, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, 1996, *58* (1), 267–288.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in "Advances in Neural Information Processing Systems," Vol. 30 2017.

Wheaton, William C. and Cunjia Xu, "Using AI to Improve Price Transparency in Real Estate Valuation," Research Paper, MIT Center for Real Estate 2024.

Yazdani, Mahdieh and Maziar Raissi, "Real Estate Property Valuation using Self-Supervised Vision Transformers," Working Paper 2023.

Appendix

A Data: Variable List and Summary Statistics

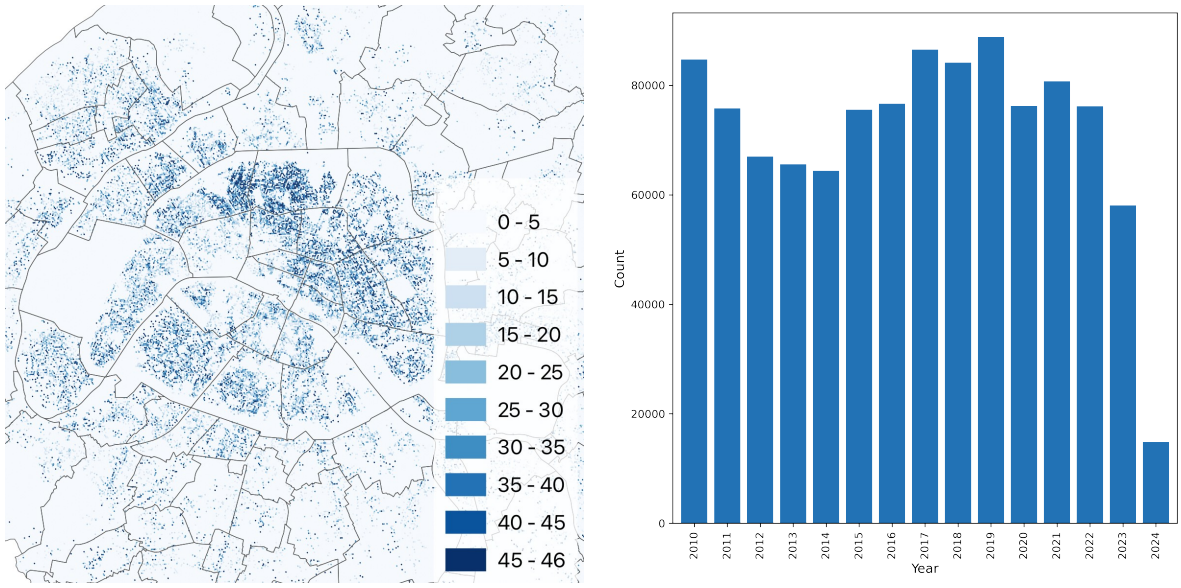
Table 12 lists the variables available from DV3F and used in the construction of the unit attribute vector X_{it}.

Figure 12: DV3F variable list

Variable	**Description**
tid	Transaction identifier
hid	Hexagon identifier
year	Year of transaction
hid7	Neighbourhood identifier
price	Transaction price
apartment	Indicator for apartment property type
age	Building age at time of transaction
age_2	Building age squared
type_prop	Property type
n_lot	Number of lots
n_unit	Number of housing units
year_built	Year of construction
unit_area	Unit area (m^2)
land_area	Land area (m^2)
annex_area	Annex area (m^2)
n_rooms	Number of main rooms
n_floor	Number of floors
n_levels	Number of building levels
n_diningroom	Number of dining rooms
n_bedroom	Number of bedrooms
n_bathroom	Number of bathrooms
n_kitchen	Number of kitchens
n_annex	Number of annexes
n_mainroom	Number of principal rooms
n_garage	Number of garages
n_conservatory	Number of conservatories
n_terrace	Number of terraces
n_pool	Number of swimming pools
n_other	Number of other rooms

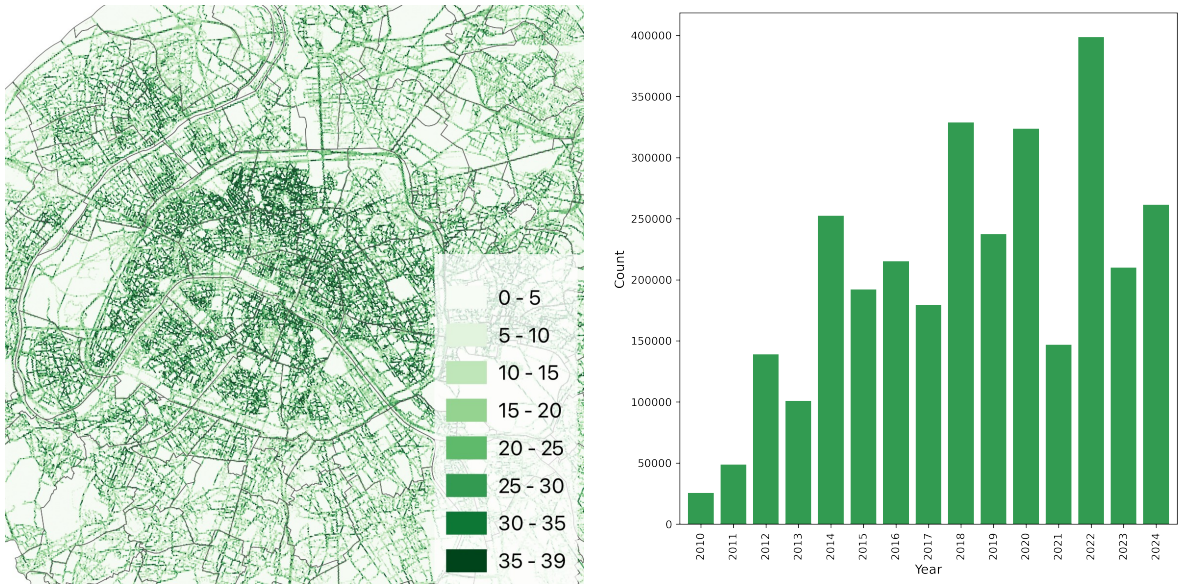
B Data: Spatial and Temporal Distributions

Figure 13: Distribution of transactions.



Notes: Distribution of transactions by hexagon (left) and by year (right).

Figure 14: Distribution of Street View panoramas.



Notes: Distribution of Street View panoramas by hexagon (left) and by year (right).

C CLIP Architecture Details

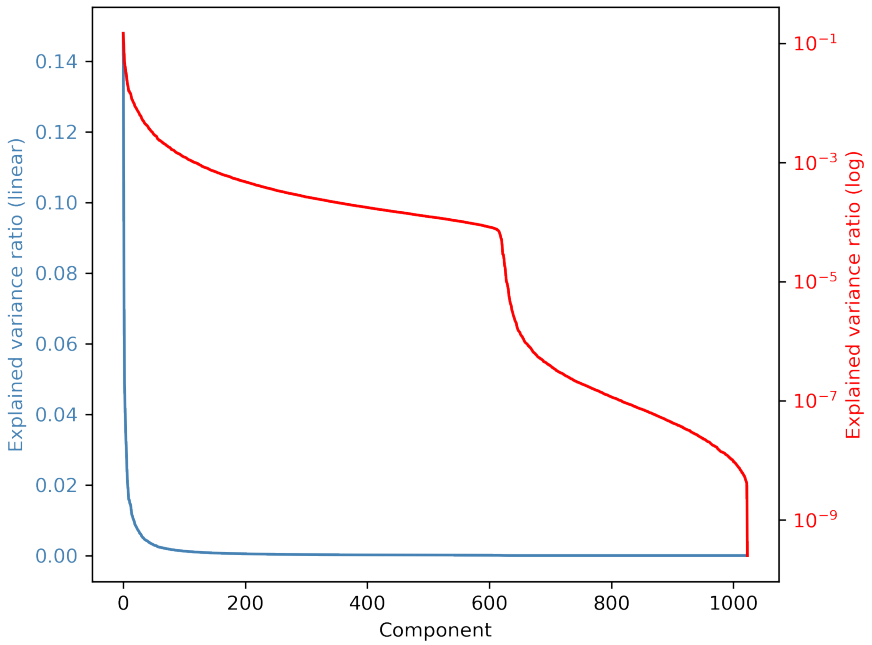
Figure 15: OpenCLIP ViT-H/14 - Vision encoder

Operation	**Input → output dim.**	**Num. param.**
Embeddings		
Patch embedding (convolution)	$(16 \times 16) \times (14 \times 14 \times 3) \rightarrow 256 \times 1280$	$1280 \times (14 \times 14 \times 3)$
Concat. special token embed.	$256 \times 1280 \rightarrow 257 \times 1280$	1280
Add position embedding	$257 \times 1280 \rightarrow 257 \times 1280$	257×1280
Layer normalisation	257×1280	2×1280
Encoder $\times 32$		
*Self-attention (*16 *heads)*		
Layer normalisation	257×1280	2×1280
Key projection & reshape	$257 \times 1280 \rightarrow 257 \times 16 \times 80$	$1280^2 + 1280$
Query projection & reshape	$257 \times 1280 \rightarrow 257 \times 16 \times 80$	$1280^2 + 1280$
Value projection & reshape	$257 \times 1280 \rightarrow 257 \times 16 \times 80$	$1280^2 + 1280$
Scaled dot-product attention	$257 \times 16 \times 80 \rightarrow 257 \times 16 \times 80$	0
Concatenate heads	$257 \times 16 \times 80 \rightarrow 257 \times 1280$	0
Output projection	$257 \times 1280 \rightarrow 257 \times 1280$	$1280^2 + 1280$
Residual addition	$257 \times 1280 \rightarrow 257 \times 1280$	0
Point-wise MLP		
Layer normalisation	257×1280	2×1280
Fully connected	$257 \times 1280 \rightarrow 257 \times 5120$	$1280 \times 5120 + 5120$
Activation (GELU)	257×5120	0
Fully connected	$257 \times 5120 \rightarrow 257 \times 1280$	$5120 \times 1280 + 1280$
Residual addition	$257 \times 1280 \rightarrow 257 \times 1280$	0
Output head		
Layer normalisation	257×1280	2×1280
Select special token	$257 \times 1280 \rightarrow 1 \times 1280$	0
Project special token (no bias)	$1 \times 1280 \rightarrow 1 \times 1024$	1280×1024

Figure 16: OpenCLIP ViT-H/14 - Text encoder

Operation	**Input → output dim.**	**Num. param.**
Embeddings (incl. special token)		
Token embedding	77×1024	49408×1024
Add position embedding	$77 \times 1024 \to 77 \times 1024$	77×1024
Layer normalisation	77×1024	2×1024
Encoder $\times 24$		
*Self-attention (*16 *heads)*		
Layer normalisation	77×1024	2×1024
Key projection & reshape	$77 \times 1024 \to 77 \times 16 \times 64$	$1024^2 + 1024$
Query projection & reshape	$77 \times 1024 \to 77 \times 16 \times 64$	$1024^2 + 1024$
Value projection & reshape	$77 \times 1024 \to 77 \times 16 \times 64$	$1024^2 + 1024$
Scaled dot-product attention	$77 \times 16 \times 64 \to 77 \times 16 \times 64$	0
Concatenate heads	$77 \times 16 \times 64 \to 77 \times 1024$	0
Output projection	$77 \times 1024 \to 77 \times 1024$	$1024^2 + 1024$
Residual addition	$77 \times 1024 \to 77 \times 1024$	0
Point-wise MLP		
Layer normalisation	77×1024	2×1024
Fully connected	$77 \times 1024 \to 77 \times 4096$	$1024 \times 4096 + 4096$
Activation (GELU)	77×4096	0
Fully connected	$77 \times 4096 \to 77 \times 1024$	$4096 \times 1024 + 1024$
Residual addition	$77 \times 1024 \to 77 \times 1024$	0
Output head		
Layer normalisation	77×1024	2×1024
Select special token	$77 \times 1024 \to 1 \times 1024$	0

Figure 17: Eigenspectrum of the PCA decomposition.



Notes: Screen plot of PCA components on CLIP embeddings. The linear scale (left) shows that the first 50 components account for the bulk of variance; the log scale (right) reveals a sharp discontinuity around component 650, after which explained variance drops by several orders of magnitude.

D Full Regression Tables

Figure 18: Hedonic regressions

	log(Price sqm.)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Attributes								
Apartment		0.0273		-0.1950***		-0.1750***		-0.2050***
log(Unit area)		-0.1570***		-0.1720***		-0.1730***		-0.1750***
log(Land area)		0.0343**		0.0299***		0.0319***		0.0297***
log(Annex area)		-0.0301***		0.0456***		0.0563***		0.0552***
Age		0.0000***		0.0000		-0.0000		-0.0000
Age^2		0.0000***		-0.0000**		-0.0000		-0.0000***
N. annex 1		-0.0349***		0.0093**		0.0129***		0.0151***
N. annex 2		-0.0679***		0.0183***		0.0252***		0.0289***
N. annex 3+		-0.0752***		0.0193***		0.0296***		0.0337***
N. bathroom 1		-0.0208***		0.0268***		0.0272***		0.0293***
N. bathroom 2		0.1340***		0.0798***		0.0695***		0.0598***
N. bathroom 3+		0.2490***		0.0925***		0.0718***		0.0537***
N. bedroom 1		-0.0067		0.0368***		0.0417***		0.0438***
N. bedroom 2		0.1580***		0.0777***		0.0901***		0.0772***
N. bedroom 3+		0.2090***		0.0683***		0.0829***		0.0665***
N. floor 1-2		0.0805***		0.0598***		0.0599***		0.0610***
N. floor 3-5		0.1100***		0.0690***		0.0688***		0.0707***
N. floor 6+		0.0188**		0.0367***		0.0171***		0.0367***
Embeddings	-	-	✓	✓	-	-	✓	✓
Year FE	✓	✓	✓	✓	✓	✓	✓	✓
Neighbourhood FE	-	-	-	-	✓	✓	✓	✓
Adjusted R^2	0.0193	0.0974	0.2600	0.2770	0.2790	0.2990	0.2980	0.3150
Observations	601,220	601,220	601,220	601,220	601,220	601,220	601,220	601,220

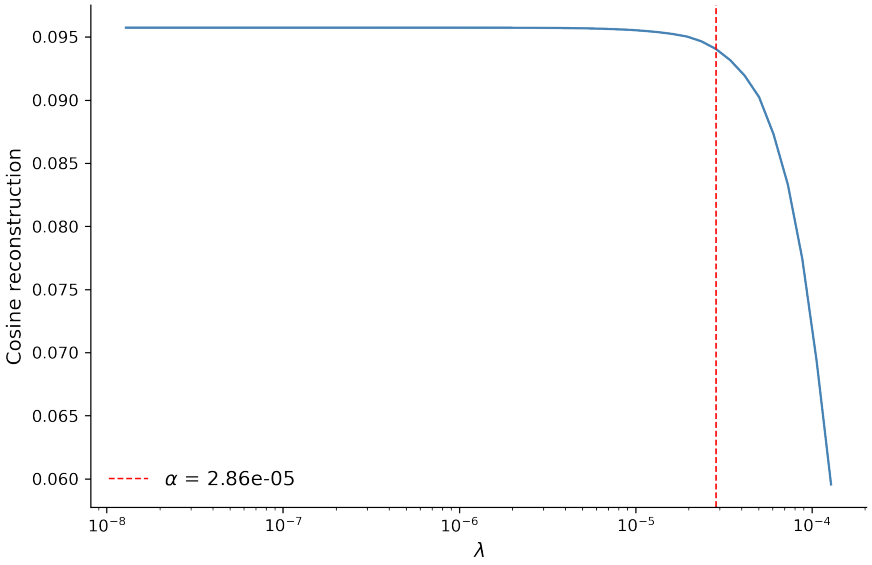
Notes: Standard errors clustered at the hexagon level.

E Validation: OpenStreetMap Data Construction

Spatial distribution of object scores A first validation exercise examines whether the spatial distribution of object scores across Paris hexagons aligns with what would be expected from independent geographic knowledge of the city. Figure 20 displays heatmaps of average scores at the hexagon level for the trees, Haussmannian building, and cobblestone street objects.

The trees heatmap (Figure 20a, left) shows elevated scores along tree-lined boulevards,

Figure 19: Grid search over the LASSO penalty.

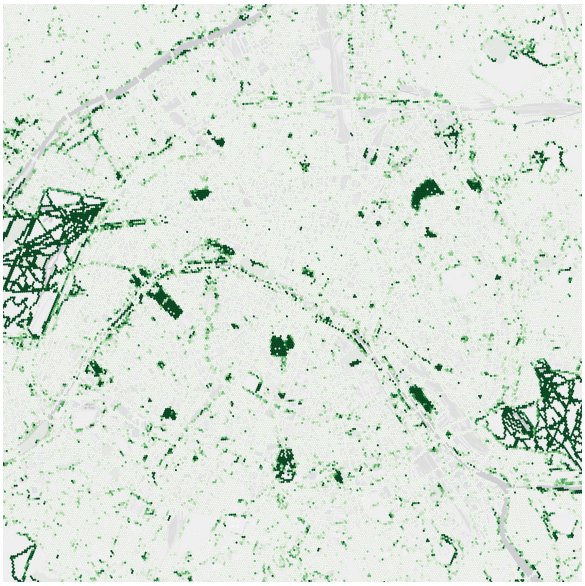


Notes: We select the values of λ via grid search to minimize reconstruction error, measured by the cosine similarity between observed and reconstructed embeddings. The estimation is performed on a random sample of 10,000 images. Both image and word embeddings are L2-normalized to ensure comparability of cosine similarity scores.

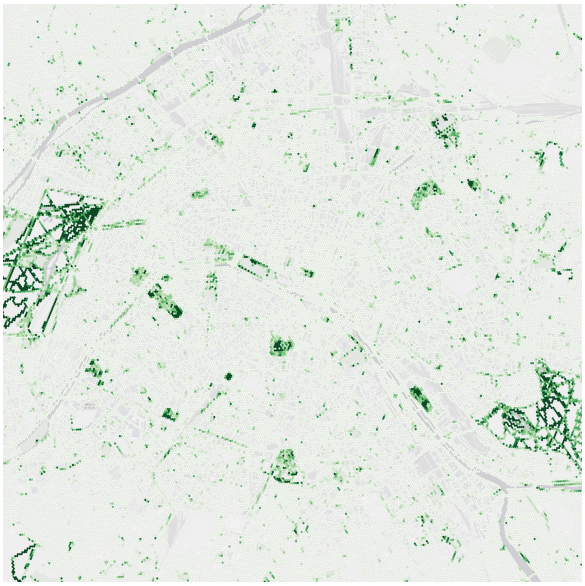
in the Bois de Boulogne and Bois de Vincennes (not shown, lying outside the *arrondissements*), and in hexagons near parks and river banks. The OpenStreetMap vegetation benchmark (Figure 20b, right) captures designated green areas but misses street trees lining the major avenues and vegetation along the Seine embankments. The object scores successfully recover both major parks and the more diffuse urban tree cover that is absent from the administrative green space data, illustrating a strength of the visual measurement approach.

The Haussmannian building object (Figure 20c) shows scores that are densest in the inner *arrondissements* and along the great Haussmann-era boulevards, consistent with the historical geography of 19th-century urban renewal in Paris. Walkability scores (Figure 20d) peak in Montmartre, the Marais, and other districts known for their pedestrian-friendly streets and cobblestone character, and are lower in more car-oriented peripheral areas.

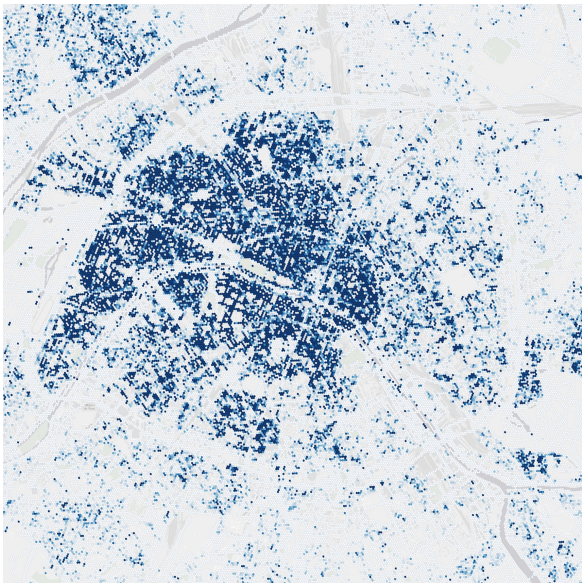
Figure 20: Semantic object score heatmaps and OpenStreetMap validation.



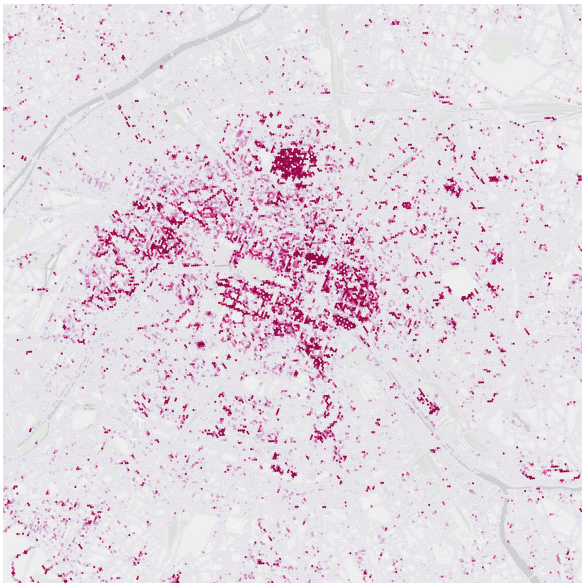
(a) Trees: object scores



(b) Trees: OpenStreetMap



(c) Haussmanian buildings: object scores



(d) Cobblestone streets: object scores

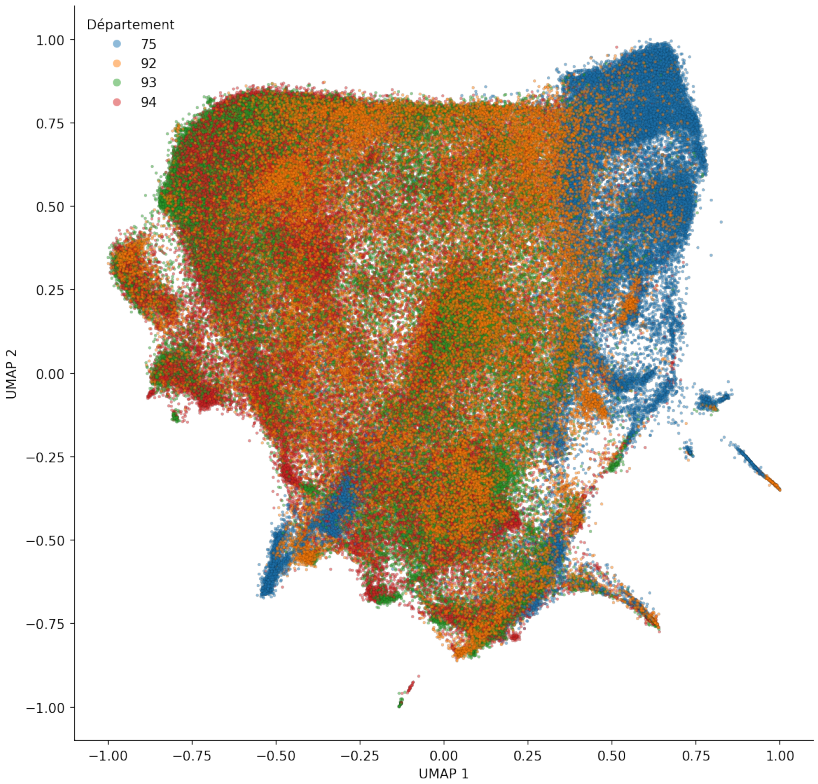
Notes: Panels a, c, and d report average sparse group LASSO object scores, while panel b presents validation measures based on OpenStreetMap data. Vegetation scores capture major parks and urban tree cover that are only partially reflected in administrative sources. Haussmannian scores are concentrated in the historical *arrondissements*, whereas walkability scores are highest in pedestrian-oriented areas such as Montmartre.

UMAP visualization of neighborhood types Figure 21 shows a UMAP projection McInnes et al. (2018) of the hexagon-level average embeddings, where each point

represents the average embedding of one hexagon over all years. UMAP preserves local structure in the high-dimensional embedding space while approximately respecting global structure, so that nearby points in the two-dimensional projection correspond to hexagons with similar visual environments. Points are colored by arrondissement.

The projection reveals a clear spatial clustering: hexagons from the same arrondissement tend to cluster together, reflecting the spatial autocorrelation of neighborhood visual characteristics within administrative units. At the same time, there is substantial within-arrondissement heterogeneity, and hexagons from different *arrondissements* with similar visual environments—for example, comparable Haussmannian streetscapes in different parts of the city—appear close together across arrondissement boundaries. Outlier clusters in the projection correspond to hexagons near parks, major traffic interchanges, or distinctive architectural landmarks. This structure validates that the embeddings encode meaningful neighborhood characteristics and not merely broad location.

Figure 21: UMAP projection of hexagon-level average CLIP embeddings.



Notes: Each point represents the average embedding of one hexagon, projected into two dimensions using UMAP McInnes et al. (2018). Colors correspond to *départements*. Hexagons from inner Paris same arrondissement tend to cluster together.

F Data: Objects

Figure 22: Objects used in the sparse coding.

Category	Objects
Building	Commercial building, Derelict building, Haussmannian building, Heritage site monument, High-rise building, Historical building, Industrial building, Modern building, Religious building, Residential building
Cycling	Bicycle, Cycling lane, Electric scooter
Greenery	Other greenery, Trees
Hospitality	Bar cafe, Fast food, Hotel, Restaurant
Infrastructure	Bridge, Exterior staircase, Retaining wall, Underpass
Item	Statue, Water fountain
Other	Graffiti, Homeless, Trash
Outlet	Cinema theatre concert hall, Museum gallery exhibition, Sports facility
Pedestrian	Cobblestone street, Pavement, Pedestrian, Pedestrian crossing
Public services	Administrative building, Cemetery, Education building, Police fire station, Post office, Public transport
Retail	Bakery, Clothing shop, Cultural retail, Florist shop, Grocery store, Health retail, Luxury retail, Market stall, Shopping mall, Speciality retail food, Speciality retail other
Services	Beauty services, Financial services, Health services, Laundry services, Travel agency
Street furniture	Bench, Fence, Newspaper kiosk, Panel poster, Public toilet, Streetlight, Traffic signs, Trash can
Traffic	Road, Vehicle
View	Construction, Countryside field crops, Water river canal

Notes: To enhance semantic separation, CLIP embeddings are constructed from descriptive phrases rather than object names. For instance, the embedding for a Haussmannian building is obtained by averaging representations of terms such as "19th century building", "limestone facade", "balconies", "mansard roof", and "sculpted window details". Similarly, the public transport embedding is computed from terms including "bus", "bus stop", "bus lane", "metro", "metro station", "tram", and "tram stop".

G Model Appendix

G.1 Parametric aggregate amenity effects with attention weights

We now verify that the parametric specifications of $B(\mathbf{b}, \mathbf{m})$ used in the main text satisfy the assumptions imposed on the aggregate amenity function.

Case 1: Power function. Suppose $\widetilde{B}(\mathbf{b}) = b_1^{\gamma_1} \cdots b_K^{\gamma_K}$ and $B(\mathbf{b}, \mathbf{m}) = b_1^{m_1\gamma_1} \cdots b_K^{m_K\gamma_K}$, with $b_k \in \left[1, \overline{b}_k\right]$.

Indeed, for any amenity k for which $m_k = 0$, the specific amenity does not affect the aggregate amenity component since $b_{i,1}^{0\cdot\gamma_k} = 1$. Moreover,

$$\frac{\partial B(\mathbf{b}, \mathbf{m})}{\partial b_k} = \frac{\gamma_k B(\mathbf{b}, \mathbf{m})}{b_k} > 0,$$

$$\frac{\partial B(\mathbf{b}, \mathbf{m})}{\partial m_k} = \gamma_k \log b_k \cdot B(\mathbf{b}, \mathbf{m}) > 0 \text{ for } b_k > 1 \quad \text{and} \quad \frac{\partial B(\mathbf{b}, \mathbf{m})}{\partial m_k} = 0 \text{ for } b_k = 1,$$

and

$$\frac{\partial^2 B(\mathbf{b}, \mathbf{m})}{\partial m_k \partial b_k} = \frac{m_k \gamma_k^2 \log b_k}{b_k} \cdot B(\mathbf{b}, \mathbf{m}) > 0 \text{ for } b_k > 1 \quad \text{and} \quad \frac{\partial^2 B(\mathbf{b}, \mathbf{m})}{\partial m_k \partial b_k} = 0 \text{ for } b_k = 1.$$

Finally, $b_k^{1\cdot\gamma_k} = b_k^{\gamma_k}$ yields $B(\mathbf{b}, \mathbf{1}_K) = \widetilde{B}(\mathbf{b})$, and checking boundedness is straightforward.

Case 2: Exponential function. Properties for the aggregate amenity effect with attention weights are also verified when $\widetilde{B}(\mathbf{b}) = \exp\left(\sum_{k=1}^K \gamma_k b_k\right)$ and $B(\mathbf{b}, \mathbf{m}) = \exp\left(\sum_{k=1}^K m_k \gamma_k b_k\right)$, with $b_k \in \left[0, \overline{b}_k\right]$.

Indeed, for any amenity k for which $m_k = 0$, the specific amenity does not affect the aggregate amenity component since $0 \cdot \gamma_k b_{i,k} = 0$. Moreover,

$$\frac{\partial B(\mathbf{b}, \mathbf{m})}{\partial b_k} = m_k \gamma_k B(\mathbf{b}, \mathbf{m}) \geq 0,$$

$$\frac{\partial B(\mathbf{b},\mathbf{m})}{\partial m_k} = \gamma_k b_k B(\mathbf{b},\mathbf{m}) > 0 \text{ for } b_k > 0 \quad \text{and} \quad \frac{\partial B(\mathbf{b},\mathbf{m})}{\partial m_k} = 0 \text{ for } b_k = 0,$$

and

$$\frac{\partial^2 B(\mathbf{b},\mathbf{m})}{\partial m_k \partial b_k} = \gamma_k B(\mathbf{b},\mathbf{m}) + m_k \gamma_k^2 b_k B(\mathbf{b},\mathbf{m}) > 0.$$

Finally, $1 \cdot \gamma_k b_k = \gamma_k b_k$ yields $B(\mathbf{b}, \mathbf{1}_K) = \widetilde{B}(\mathbf{b})$, and checking boundedness is straightforward.

These specifications therefore satisfy the selection, monotonicity, complementarity, and full-attention consistency properties imposed in the main text, and meet the regularity conditions required for the existence of corner solutions.

G.2 Computation of the indirect utility $V(\mathbf{m})$

This appendix derives equation 14 in Section 5.

The indirect utility verifies:

$$\begin{aligned} V(\mathbf{m}) &= E_{\mathbf{z}|v(\mathbf{m})}\left(z_{i(\mathbf{z},\mathbf{m}),j(\mathbf{z},\mathbf{m})} v_{i(\mathbf{z},\mathbf{m}),j(\mathbf{z},\mathbf{m})}\right) \\ &= E_{\mathbf{z}|v(\mathbf{m})}\left(\sum_{i,j=1}^{J} z_{i,j} v_{i,j}\, \mathbf{1}_{\{i(\mathbf{z},\mathbf{m})=i,\, j(\mathbf{z},\mathbf{m})=j\}}\right) \\ &= \sum_{i,j=1}^{J} v_{i,j}\, E_{\mathbf{z}|v(\mathbf{m})}\left(z_{i,j}\, \mathbf{1}_{\{i(\mathbf{z},\mathbf{m})=i,\, j(\mathbf{z},\mathbf{m})=j\}}\right) \end{aligned}$$

We have:

$$\begin{aligned} E_{\mathbf{z}|v(\mathbf{m})}\left(z_{i,j}\, \mathbf{1}_{\{i(\mathbf{z},\mathbf{m})=i,\, j(\mathbf{z},\mathbf{m})=j\}}\right) &= E_{z_{i,j}}\left[z_{i,j}\, E_{\mathbf{z}|z_{i,j},v(\mathbf{m})}\left(\mathbf{1}_{\{i(\mathbf{z},\mathbf{m})=i,\, j(\mathbf{z},\mathbf{m})=j\}}\right)\right] \\ &= E_{z_{i,j}}\left[z_{i,j}\, P\left(i(\mathbf{z},\mathbf{m}) = i,\, j(\mathbf{z},\mathbf{m}) = j \mid z_{i,j}, v(\mathbf{m})\right)\right] \end{aligned}$$

Making use of the independence of $z_{k,\ell}$ terms:

$$\begin{aligned}
P\left(i\left(\mathbf{z},\mathbf{m}\right)=i,\, j\left(\mathbf{z},\mathbf{m}\right)=j \mid z_{i,j}, v\left(\mathbf{m}\right)\right) &= P\left(\max_{(r,s)\neq(i,j)} V_{r,s}\left(\mathbf{m}\right) < V_{i,j}\left(\mathbf{m}\right) \mid z_{i,j}, v\left(\mathbf{m}\right)\right) \\
&= P\left(\max_{(r,s)\neq(i,j)} z_{r,s}v_{r,s}\left(\mathbf{m}\right) < z_{i,j}v_{i,j}\left(\mathbf{m}\right) \mid z_{i,j}\right) \\
&= \prod_{(r,s)\neq(i,j)} P\left(z_{r,s}v_{r,s}\left(\mathbf{m}\right) < z_{i,j}v_{i,j}\left(\mathbf{m}\right) \mid z_{i,j}\right) \\
&= \prod_{(r,s)\neq(i,j)} F_{r,s}\left(z_{i,j}\frac{v_{i,j}\left(\mathbf{m}\right)}{v_{r,s}\left(\mathbf{m}\right)}\right)
\end{aligned}$$

Hence, we have:

$$\begin{aligned}
V\left(\mathbf{m}\right) &= \sum_{i,j=1}^{J} v_{i,j}\int_0^{+\infty} x\, f_{i,j}(x) \prod_{(r,s)\neq(i,j)} F_{r,s}\left(x\frac{v_{i,j}\left(\mathbf{m}\right)}{v_{r,s}\left(\mathbf{m}\right)}\right) dx \\
&= \sum_{i,j=1}^{J} \varepsilon v_{i,j}\int_0^{+\infty} x\, T_i E_j x^{-\varepsilon-1}\exp\left(-T_i E_j x^{-\varepsilon}\right) \prod_{(r,s)\neq(i,j)} \exp\left(-T_r E_s\left[\frac{x v_{i,j}\left(\mathbf{m}\right)}{v_{r,s}\left(\mathbf{m}\right)}\right]^{-\varepsilon}\right) dx \\
&= \sum_{i,j=1}^{J} \varepsilon v_{i,j}\int_0^{+\infty} x\, T_i E_j x^{-\varepsilon-1}\exp\left[-\left(\sum_{r,s=1}^{J} T_r E_s\left[\frac{v_{r,s}\left(\mathbf{m}\right)}{v_{i,j}\left(\mathbf{m}\right)}\right]^{\varepsilon}\right) x^{-\varepsilon}\right] dx \\
&= \sum_{i,j=1}^{J} \varepsilon v_{i,j}\int_0^{+\infty} x\, T_i E_j x^{-\varepsilon-1}\exp\left(-A_{i,j}^{\varepsilon} x^{-\varepsilon}\right) dx
\end{aligned}$$

where $f_{i,j}(\cdot)$ is the density of the Fréchet distribution for respective residential and workplace locations i and j, and:

$$A_{i,j} = \left[\sum_{r,s=1}^{J} T_r E_s\left[\frac{v_{r,s}\left(\mathbf{m}\right)}{v_{i,j}\left(\mathbf{m}\right)}\right]^{\varepsilon}\right]^{1/\varepsilon}$$

We have:

$$V\left(\mathbf{m}\right) = \sum_{i,j=1}^{J} \varepsilon T_i E_j v_{i,j} A_{i,j}^{1-\varepsilon}\int_0^{+\infty}\left(\frac{x}{A_{i,j}}\right)^{-\varepsilon}\exp\left(-\left(\frac{x}{A_{i,j}}\right)^{-\varepsilon}\right) d\left(\frac{x}{A_{i,j}}\right)$$

Making the change of variable $(x/A_{i,j})^{-\varepsilon} \to x$, we get:

$$\begin{aligned}
V(\mathbf{m}) &= \sum_{i,j=1}^{J} T_i E_j v_{i,j} A_{i,j}^{1-\varepsilon} \int_0^{+\infty} x^{-1/\varepsilon} \exp(-x)\, dx \\
&= \gamma \sum_{i,j=1}^{J} T_i E_j v_{i,j} A_{i,j}^{1-\varepsilon} \\
&= \gamma \sum_{i,j=1}^{J} T_i E_j \frac{v_{i,j}}{v_{i,j}(\mathbf{m})^{1-\varepsilon}} \left[\sum_{r,s=1}^{J} T_r E_s v_{r,s} (\mathbf{m})^{\varepsilon} \right]^{(1-\varepsilon)/\varepsilon} \\
&= \gamma \sum_{i,j=1}^{J} v_{i,j} \left[T_i E_j v_{i,j} (\mathbf{m})^{\varepsilon} \right]^{(\varepsilon-1)/\varepsilon} \left[\sum_{r,s=1}^{J} T_r E_s v_{r,s} (\mathbf{m})^{\varepsilon} \right]^{-(\varepsilon-1)/\varepsilon} \\
&= \gamma \sum_{i,j=1}^{J} (T_i E_j)^{1/\varepsilon} v_{i,j} \left[\pi_{i,j}(\mathbf{m}) \right]^{(\varepsilon-1)/\varepsilon} \\
&= \gamma \sum_{i,j=1}^{J} \left(T_i E_j v_{i,j}^{\varepsilon} \right)^{1/\varepsilon} \left[\pi_{i,j}(\mathbf{m}) \right]^{(\varepsilon-1)/\varepsilon} \\
&= \widetilde{V}(\mathbf{1}_K) \sum_{i,j=1}^{J} \pi_{i,j}^{1/\varepsilon} \left[\pi_{i,j}(\mathbf{m}) \right]^{1-1/\varepsilon}
\end{aligned}$$

where $\gamma = \Gamma\left(\frac{\varepsilon-1}{\varepsilon}\right)$ with $\Gamma(t) = \int_0^{+\infty} x^{t-1} \exp(-x)\, dx$ and:

$$\widetilde{V}(\mathbf{m}) = \gamma \left[\sum_{r,s=1}^{J} T_r E_s v_{r,s} (\mathbf{m})^{\varepsilon} \right]^{1/\varepsilon}$$

Note that we have:

$$V(\mathbf{1}_K) = \widetilde{V}(\mathbf{1}_K) \sum_{i,j=1}^{J} \pi_{i,j}^{1/\varepsilon} \pi_{i,j}^{1-1/\varepsilon} = \widetilde{V}(\mathbf{1}_K)$$

which is the formula obtained by Ahlfeldt et al. (2015) in the case of perfect attention.

G.3 Optimal attention weights

We have:

$$\begin{aligned} v_{i,j}(\mathbf{m}) &= \frac{B(\mathbf{b}_i,\mathbf{m})\,w_j}{d_{i,j}Q_i^{1-\beta}} \\ \pi_{i,j}(\mathbf{m}) &= \frac{T_iE_jv_{i,j}(\mathbf{m})^{\varepsilon}}{\sum_{r,s=1}^{J}T_rE_sv_{r,s}(\mathbf{m})^{\varepsilon}} \end{aligned}$$

Hence:

$$\begin{aligned} \frac{\partial \log v_{r,s}(\mathbf{m})}{\partial m_k} &= \frac{\partial \log B(\mathbf{b}_r,\mathbf{m})}{\partial m_k} > 0 \\ \frac{\partial \pi_{i,j}(\mathbf{m})}{\partial m_k} &= \frac{\varepsilon}{v_{i,j}(\mathbf{m})}\frac{\partial v_{i,j}(\mathbf{m})}{\partial m_k}\pi_{i,j}(\mathbf{m}) - \varepsilon\sum_{r,s=1}^{J}\frac{\partial v_{r,s}(\mathbf{m})}{\partial m_k}T_rE_sv_{r,s}(\mathbf{m})^{\varepsilon-1}\frac{T_iE_jv_{i,j}(\mathbf{m})^{\varepsilon}}{\left[\sum_{r,s=1}^{J}T_rE_sv_{r,s}(\mathbf{m})^{\varepsilon}\right]^2} \\ &= \varepsilon\frac{\partial \log v_{i,j}(\mathbf{m})}{\partial m_k}\pi_{i,j}(\mathbf{m}) - \varepsilon\sum_{r,s=1}^{J}\frac{\partial \log v_{r,s}(\mathbf{m})}{\partial m_k}T_rE_sv_{r,s}(\mathbf{m})^{\varepsilon}\frac{T_iE_jv_{i,j}(\mathbf{m})^{\varepsilon}}{\left[\sum_{r,s=1}^{J}T_rE_sv_{r,s}(\mathbf{m})^{\varepsilon}\right]^2} \\ &= \varepsilon\left[\frac{\partial \log B(\mathbf{b}_i,\mathbf{m})}{\partial m_k} - \sum_{r=1}^{J}\pi_r(\mathbf{m})\frac{\partial \log B(\mathbf{b}_r,\mathbf{m})}{\partial m_k}\right]\pi_{i,j}(\mathbf{m}) \end{aligned}$$

where:

$$\pi_i(\mathbf{m}) = \sum_{j=1}^{J}\pi_{i,j}(\mathbf{m})$$

We get:

$$\begin{aligned} \frac{\partial V(\mathbf{m})}{\partial m_k} &= \frac{\varepsilon-1}{\varepsilon}\widetilde{V}(\mathbf{1}_K)\sum_{i,j=1}^{J}\frac{\partial \pi_{i,j}(\mathbf{m})}{\partial m_k}\pi_{i,j}^{1/\varepsilon}\left[\pi_{i,j}(\mathbf{m})\right]^{-1/\varepsilon} \qquad (23) \\ &= (\varepsilon-1)\widetilde{V}(\mathbf{1}_K)\sum_{i,j=1}^{J}\left[\frac{\partial \log B(\mathbf{b}_i,\mathbf{m})}{\partial m_k} - \sum_{r=1}^{J}\pi_r(\mathbf{m})\frac{\partial \log B(\mathbf{b}_r,\mathbf{m})}{\partial m_k}\right]\pi_{i,j}^{1/\varepsilon}\left[\pi_{i,j}(\mathbf{m})\right]^{1-1/\varepsilon} \end{aligned}$$

Note that we have:

$$\begin{aligned}\frac{\partial V(\mathbf{1}_K)}{\partial m_k} &= (\varepsilon-1)\widetilde{V}(\mathbf{1}_K)\sum_{i,j=1}^{J}\left[\frac{\partial \log B(\mathbf{b}_i,\mathbf{1}_K)}{\partial m_k}-\sum_{r=1}^{J}\pi_r(\mathbf{1}_K)\frac{\partial \log B(\mathbf{b}_r,\mathbf{1}_K)}{\partial m_k}\right]\pi_{i,j}^{1/\varepsilon}\left[\pi_{i,j}(\mathbf{1}_K)\right]^{1-1/\varepsilon}\\ &= (\varepsilon-1)\widetilde{V}(\mathbf{1}_K)\sum_{i,j=1}^{J}\left[\frac{\partial \log B(\mathbf{b}_i,\mathbf{1}_K)}{\partial m_k}-\sum_{r=1}^{J}\pi_r(\mathbf{1}_K)\frac{\partial \log B(\mathbf{b}_r,\mathbf{1}_K)}{\partial m_k}\right]\pi_{i,j}\\ &= 0\end{aligned}$$

Differentiating equation 23, we also get:

$$\begin{aligned}\frac{\partial^2 V(\mathbf{m})}{\partial m_k^2} &= \frac{\varepsilon-1}{\varepsilon}\widetilde{V}(\mathbf{1}_K)\sum_{i,j=1}^{J}\left[\frac{\partial^2\pi_{i,j}(\mathbf{m})}{\partial m_k^2}-\frac{1}{\varepsilon\pi_{i,j}(\mathbf{m})}\left[\frac{\partial\pi_{i,j}(\mathbf{m})}{\partial m_k}\right]^2\right]\pi_{i,j}^{1/\varepsilon}\left[\pi_{i,j}(\mathbf{m})\right]^{-1/\varepsilon}\\ &= (\varepsilon-1)\widetilde{V}(\mathbf{1}_K)\sum_{i,j=1}^{J}\left[\varepsilon\pi_{i,j}(\mathbf{m})\frac{\partial^2\pi_{i,j}(\mathbf{m})}{\partial m_k^2}-\left[\frac{\partial\pi_{i,j}(\mathbf{m})}{\partial m_k}\right]^2\right]\pi_{i,j}^{1/\varepsilon}\left[\pi_{i,j}(\mathbf{m})\right]^{-1-1/\varepsilon}\end{aligned}$$

G.4 Indirect utility as a function of location choice probabilities is maximum at $\pi_{i,j}$

We denote $V^P(\theta)=\widetilde{V}(\mathbf{1}_K)\sum_{i,j=1}^{J}\pi_{i,j}^{1/\varepsilon}\theta_{i,j}^{1-1/\varepsilon}$ where $\theta=(\theta_{1,1},\ldots,\theta_{J,J})$. We want to solve:

$$\theta^*=\arg\max V^P(\theta)\quad \text{s.t.}\quad \sum_{i,j=1}^{J}\theta_{i,j}=1$$

The Lagrangian can be written as:

$$\mathcal{L}=\widetilde{V}(\mathbf{1}_K)\sum_{i,j=1}^{J}\pi_{i,j}^{1/\varepsilon}\theta_{i,j}^{1-1/\varepsilon}+\varkappa\left(1-\sum_{i,j=1}^{J}\theta_{i,j}\right)$$

where $\varkappa$ is the Lagrange multiplier. The first-order conditions are:

$$\frac{\partial\mathcal{L}}{\partial\theta_{i,j}}=\frac{\varepsilon-1}{\varepsilon}\widetilde{V}(\mathbf{1}_K)\pi_{i,j}^{1/\varepsilon}\theta_{i,j}^{-1/\varepsilon}-\varkappa=0$$

or equivalently:

$$\theta_{i,j} = \left(\frac{\varepsilon - 1}{\varepsilon \varkappa} \widetilde{V}(\mathbf{1}_K) \right)^{\varepsilon} \pi_{i,j} \tag{24}$$

Summing over all i and j and using $\sum_{i,j} \theta_{i,j} = 1$, we get:

$$\begin{aligned} \left(\frac{\varepsilon - 1}{\varepsilon \varkappa} \widetilde{V}(\mathbf{1}_K) \right)^{\varepsilon} &= 1 \\ \varkappa &= \frac{\varepsilon - 1}{\varepsilon} \widetilde{V}(\mathbf{1}_K) \end{aligned}$$

Substituting this expression into equation (24), we get $\theta_{i,j} = \pi_{i,j}$.

We also have:

$$\begin{aligned} \frac{\partial^2 \mathcal{L}}{\partial \theta_{i,j}^2} &= -\frac{1}{\varepsilon} \frac{\varepsilon - 1}{\varepsilon} \widetilde{V}(\mathbf{1}_K) \pi_{i,j}^{1/\varepsilon} \theta_{i,j}^{-1/\varepsilon - 1} < 0 \\ \frac{\partial^2 \mathcal{L}}{\partial \theta_{i,j} \partial \theta_{r,s}} &= 0 \quad \text{for } (i,j) \neq (r,s) \end{aligned}$$

Hence the Lagrangian is strictly concave and the maximum is global. This confirms that $V^P(\theta)$ is maximized uniquely at $\theta_{i,j} = \pi_{i,j}$, which implies that $V(\mathbf{m})$ is maximized at $\mathbf{m} = \mathbf{1}_K$ with $V(\mathbf{1}_K) = \widetilde{V}(\mathbf{1}_K)$.

G.5 Corner solutions for attention weights

Under Assumption 6 of Section 5, $\partial V(\mathbf{m}) / \partial m_k$ is bounded above by a quantity proportional to $\underline{B}_{1,k}(b)$. Hence it is possible that:

$$\frac{\partial V(\mathbf{m})}{\partial m_k} < \lambda \quad \text{for all } \mathbf{m} \in [0,1]^K$$

in which case the optimal choice is the corner $m_k^* = 0$. The intuition is that, if changing attention through a change in m_k does not affect much utility, in particular because the level of amenity b_k is small or amenity k does not matter much, there is no point in paying attention to amenity k. For the parametric cases of Appendix G.1, $\partial B(\mathbf{b}, \mathbf{m}) / \partial m_k$ becomes arbitrarily small when b_k or γ_k becomes arbitrarily small, so that corner solutions $m_k^* = 0$ arise when amenity k is sufficiently unimportant or has a sufficiently low

baseline level.