Building-scale functional mixing reveals a decarbonization window in cities

Jiayi Li[1,†], Zhen Zhang[1,†], Xiaofeng Pan[2], Song Ma[2], Xiangyun Xiong[2], Liangpei Zhang[3,*] and Xin Huang[1,*]

[1]School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China.
[2]Shenzhen Ecological and Environmental Monitoring Center of Guangdong Province, Shenzhen 518049, China.
[3]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan 430079, China.

[†]These authors contributed equally: Jiayi Li, Zhen Zhang.
[*]e-mail: xhuang@whu.edu.cn (Xin Huang); zlp62@whu.edu.cn (Liangpei Zhang).

Contents

Supplementary Tables

Supplementary Note 1: Why harmonization and a two-tier functional system are necessary

SN1.1. Coverage bias and label sparsity in crowd-sourced geodata

Crowd-sourced geodata (e.g., AOIs and POIs) provide useful semantic cues, but their coverage is systematically uneven across city tiers and within-city space **(Supplementary Fig. S1a, b)**. In Tier-1 urban cores, AOI and POI coverage ratios reach ~33% and ~18%, respectively, whereas in Tier-3 cities or peripheral zones they fall to <18% and <3% **(Supplementary Fig. S1c, d)**. Consequently, direct POI-to-building matching leaves ~40–60% of buildings nationwide unlabeled. This missingness is spatially structured, concentrating in low-density neighborhoods, industrial areas, and urban fringes, so direct overlay-based labeling would understate functional diversity where mixing is least documented and bias inference on the mixing–carbon relationship.

SN1.2. Data quality and taxonomy inconsistencies

Where tags exist, their raw forms are not designed for building-scale quantification and can distort estimates of functional intensity. POIs are point records that do not encode built volume; e.g., a 10,000 m^2 factory and a 50 m^2 kiosk can both appear as a single POI, making point counts a poor proxy. Taxonomies also vary across providers and often mix physical building uses with non-physical artifacts (e.g., bus stops tagged on rooftops). Furthermore, matching buildings to distant POIs presents a dilemma. Extending the search radius increases data availability, especially in sparse peripheral zones (**Supplementary Fig. S1e**), but traditional methods struggle to validate the semantic relevance of these remote signals and may introduce noise rather than information.

To resolve this, we implemented a rigorous harmonization pipeline (**Supplementary Methods 1**) to standardize crowd-sourced inputs. Crucially, our proposed BHENet framework expands the effective search radius (**Supplementary Methods 3**) not by simple proximity matching, but by learning the spatial compositional patterns of POIs. This allows the model to intelligently aggregate distant semantic cues—e.g., recognizing that a "Residential Estate" or "Industrial Park" POI located 300 m away, when fused with visual and road network features, implies a corresponding residential or industrial function for the target building—to infer missing labels. This strategy maximizes data utilization and accuracy even in sparse environments where conventional overlay methods fail.

SN1.3. Functional ambiguity across scales: context versus intrinsic use

Beyond data limitations, function itself is scale-dependent. Conventional single-label schemes implicitly assume that each building has a unique and unambiguous function, an assumption that fails in mixed and institutionally managed settings. A university campus illustrates this problem **(Supplementary Fig. S2)**. At the contextual scale, the parcel is designated as education land, whereas at the intrinsic scale it contains dormitories (residential), canteens and shops (commercial), and academic buildings (education). A single-label assignment would either collapse dormitories and canteens into education, masking residential and commercial operational-carbon signatures, or relabel them as residential or commercial and lose the co-location mechanisms that arise within the educational context. This semantic mismatch helps explain why prior multiscale evidence on mixed use can be inconsistent, since the meaning of mixing changes with the scale at which functions are defined.

SN1.4. Rationale and benefits of a two-tier functional system

To address both sparse, biased tags and functional ambiguity, we propose a two-tier functional ontology that separates contextual labels, capturing the dominant land-use setting and planning intent of a district, from intrinsic labels, capturing the most plausible building-level use within that context. This separation is essential for scale-consistent analysis. It allows areas that appear homogeneous at the district scale to exhibit building-level mixing explicitly, such as industrial parks that contain dormitories or supporting retail, where operational frictions and local capacity limits may become consequential. It also makes harmonization operational by enabling decision rules that integrate AOI geometry, POI proximity, and conflict-resolution priorities to extract a curated set of high-confidence building labels for supervision, which can then be generalized to the national building stock through model-based inference. Crucially, the two-tier framework is what enables nuanced functional combinations to be identified explicitly rather than absorbed as noise. Without this framework, nationwide mapping would either inherit the biases of sparse raw tags or rely on indiscriminate overlays that cannot separate context from instance, weakening both interpretability and transferability.

Supplementary Note 2: City-center identification and ring-based urban structure profiling

SN2.1. Definition of city centers for center-periphery analysis

Center–periphery gradients in functional mixing and urban operational carbon depend on the definition of “city center”. In China, functional density and infrastructure investment remain strongly shaped by institutional and historical inertia, so the primary administrative and commercial core often persists as a stable organizing node despite suburbanization. To avoid arbitrary anchors such as geometric centroids of administrative polygons and to ensure cross-city comparability, we define city centers using an evidence-based protocol designed for stable, interpretable ring-based profiling.

SN2.2. Two-step center identification: planning anchors corroborated by spatial evidence

We identify city centers through a two-step validation that combines planning anchors with independent spatial evidence. First, we extract candidate centers from statutory planning documents, including Territorial Spatial Planning and municipal master plans, which specify central districts and typically coincide with the administrative core or historic CBD. Second, we corroborate these candidates using multi-source spatial signals that reflect de facto activity, including peak night-time light clusters (VIIRS/DMSP composites), high centrality in the road network (e.g., dense intersections and convergence of arterial roads), and consistency with the historic urban core for mature cities. This procedure anchors the center in planning intent while ensuring alignment with functional reality, improving robustness for nationwide comparisons (**Supplementary Fig. S3**).

SN2.3. Polycentric cities and distance-to-center assignment

Most cities can be treated as predominantly monocentric at the scale of our analysis, but some mega-cities exhibit polycentric structures shaped by geography and historical development. In such cases, using a single anchor can inflate distances for large parts of the urban area and obscure localized gradients. We therefore adopt a nearest-center strategy. For cities classified as polycentric based on statutory plans and multiple activity peaks, we define a small set of center anchors and compute the distance of each grid cell i as the Euclidean distance to the nearest anchor:

$$d_i = \min_k \|x_i - C_k\|$$

where x_i is the centroid of grid cell i and C_k denotes center anchors. This assignment yields localized catchment regions and ensures that center–periphery profiling reflects the relevant nearby core rather than an artifact of selecting a single city-wide centroid.

SN2.4. Ring-based profiling and metric aggregation

Using the identified center anchors, we construct a standardized ring-based framework to characterize how functional composition and mixing vary from cores to fringes. We partition each city’s urban area into concentric distance bands at 5-km intervals from the center to the urban boundary. Within each ring, we aggregate building-resolved volumes to compute function-specific volume shares and mean BFMI, calculated under contextual and intrinsic labels where applicable. This consistent profiling captures heterogeneous trajectories across cities, including monotonic distance decay and non-monotonic patterns with mid-ring troughs, and provides the structural basis for interpreting national-scale center–periphery contrasts in the main text.

Supplementary Note 3: Hierarchical functional taxonomy and aggregation strategy

SN3.1. Hierarchical taxonomy for multiscale functional organization

To represent urban functional organization across scales while maintaining statistical stability for carbon modeling, we use a hierarchical taxonomy with three complementary granularities **(Extended Data Fig. 1a)**. At the neighborhood-context scale, we follow China's planning standard (GB50137-2011) and define eight macro-classes (Residential, Industrial, Commercial, Education, Public Open Space, Transport, Medical, Administrative) to characterize dominant neighborhood context. We then intersect neighborhood-context labels with intrinsic building uses predicted by BHENet to obtain 29 logically valid context–intrinsic combinations (e.g., Industrial–Commercial), which preserves mixed-use diversity but is long-tailed. For building-scale BFMI computation, we aggregate these 29 combinations into 12-class analytical aggregation by retaining the 11 most prevalent combinations and grouping rare combinations, which together contribute ~1.2% of total building volume, into Other to reduce sparsity-driven noise **(Supplementary Table S1)**. The 29-to-12 aggregation is motivated by both statistical and semantic considerations. Entropy-based mixing indices are sensitive to long-tail categories because rare combinations expand the normalization denominator while contributing little effective diversity in most grid cells, mechanically lowering BFMI without indicating meaningful functional change. Aggregation therefore improves signal stability for nationwide comparisons.

SN3.2 Robustness of BFMI thresholds to taxonomic aggregation

While the main text contrasts the mixing effects between neighborhood-scale BFMI based on the 8-class neighborhood context and building-scale BFMI based on the 12-class analytical aggregation, a critical methodological question remains: Does the aggregation from the 29-class context–intrinsic combination set to the 12-class analytical aggregation distort the measurement of mixing intensity? We rigorously tested this by comparing the normalized BFMI distributions under both schemas:

Quantifying Sparsity Bias (**Extended Data Fig. 1b, c**): The 29-class BFMI distribution exhibits a systematic leftward shift (Mean = 0.352; 95% CI, 0.352–0.353) relative to the 12-class baseline (Mean = 0.476; 95% CI, 0.475–0.477). This attenuation is attributable to sparsity bias (or denominator inflation): the inclusion of rare functional combinations (e.g., Education–Industrial) significantly increases the entropy normalization denominator ($\log_2 29 \approx 4.86$ vs. $\log_2 12 \approx 3.58$). Since these rare categories are absent in most grid cells, they dilute the index value without reflecting a meaningful increase in effective diversity.

Structural Stability: Despite this magnitude shift, the indices calculated under the 12-class and 29-class schemas demonstrate an exceptionally strong linear correlation (Pearson's $r > 0.999$; **Extended Data Fig. 1d**). This confirms that the rank order of functional mixing intensity—and thus the structural patterns of the urban system—is invariant to the aggregation strategy.

Threshold Robustness: Consequently, we evaluated the quantitative stability of the "decarbonization window" thresholds identified in the SDM marginal analysis (**Extended Data Fig. 1e**). Ideally, robust structural thresholds should remain stable or shift predictably despite the systematic magnitude changes caused by denominator inflation. Our results confirm this stability:

The Friction Phase End-point (Lower Bound): The threshold marking the onset of the decarbonization window demonstrated exceptional consistency, occurring at BFMI $\approx 0.145 \pm 0.022$ (95% CI, 0.120–0.163) in the 29-class model, statistically indistinguishable from the baseline 12-class threshold of 0.141 ± 0.027 (95% CI, 0.112–0.166). The substantial overlap of these confidence intervals confirms that the transition from friction to synergy is structurally invariant to taxonomic granularity.

The Saturation Point (Upper Bound): The turning point where benefits plateau or reverse appeared at BFMI $\approx 0.679 \pm 0.012$ (95% CI, 0.669–0.692) in the 29-class model, relative to 0.723 ± 0.013 (95% CI, 0.714–0.739) in the 12-class model. This slight leftward shift ($\Delta \approx 0.044$) is mathematically consistent with the overall distribution shift induced by sparsity bias (denominator inflation). Crucially, under both schemas, the threshold maps to the same high-intensity mixing regime relative to the respective population distributions.

Conclusion: The 12-class analytical schema is therefore methodologically superior, as it filters out sparsity-induced noise to provide a measure of "effective diversity" while preserving the intrinsic signal and critical thresholds of the fine-grained data. Complementary robustness checks confirming the stability of these thresholds across varying spatial grid resolutions (1 km, 2 km, 4 km) are detailed in **Supplementary Methods 4.3**.

Supplementary Note 4: Spatial structure and magnitude of decarbonization potential

Applying the aggregation protocol described in **Supplementary Methods 7** across the regression-valid nationwide grid sample, we estimate a total theoretical abatement potential of ≈ 310.4 Mt C yr^{-1} under the ideal alignment (95% CI, 267.9–355.1 Mt C yr^{-1}; 1,138.1 Mt CO_2 yr^{-1} equivalent, 95% CI, 982.3–1,302.0 Mt CO_2 yr^{-1}), representing a ≈20.9% reduction in total urban operational carbon. Under the capacity-constrained cost-effective alignment, the aggregate potential is ≈ 97.4 Mt C yr^{-1} (95% CI, 90.7–104.2 Mt C yr^{-1}; 357.0 Mt CO_2 yr^{-1} equivalent, 95% CI, 332.7–382.0 Mt CO_2 yr^{-1}), corresponding to a ≈6.6% reduction. Within cities, we further decompose contributions by the same 5-km annular rings used in the main analysis, yielding center–periphery abatement profiles for visualization and comparison.

We visualize these spatial gradients and aggregate magnitudes in **Supplementary Fig. S4**. The distance-decay profiles (**Supplementary Fig. S4a, b**) reveal distinct structural constraints across the urban hierarchy. Tier 1 cities exhibit a non-monotonic efficiency pattern, where per-grid abatement potential is suppressed in the saturated core and peaks in intermediate suburbs (20–25 km, ≈ 2.52 kt C yr^{-1} $grid^{-1}$ under ideal alignment). In contrast, Tier 3 cities display consistently high abatement efficiency across all distance bands, peaking at ≈ 3.67 kt C yr^{-1} $grid^{-1}$ in the 10–20 km rings, reflecting pervasive functional deficits. Aggregated to national totals (**Supplementary Fig. S4c**), Tier 3 cities dominate the mitigation wedge, contributing ≈ 227.8 Mt C yr^{-1} under the ideal scenario (≈ 73% of the national total) and ≈ 72.8 Mt C yr^{-1} under the cost-effective scenario. This confirms that while optimizing Tier 1 cores offers localized efficiency gains, the bulk of the national decarbonization potential resides in correcting widespread functional imbalances in lower-tier cities.

Supplementary Methods 1: Data preprocessing and harmonization pipeline

To construct a consistent spatial baseline for the 2020 reference year, we integrated three primary datasets (details provided in **Supplementary Table S8**): (1) building footprints, using the open-access sub-meter building footprint product for China released in our previous work; (2) areas of interest (AOIs), land-use polygons from OpenStreetMap (OSM) comprising ~4.7 million records; and (3) points of interest (POIs), ~42 million records from Amap (Gaode Map) spanning diverse functional categories. A key harmonization step was coordinate alignment. Because crowd-sourced datasets in China commonly use the GCJ-02 coordinate system, we applied a non-linear rectification algorithm to transform all coordinates to WGS-84, ensuring strict spatial consistency for overlays among footprints, AOIs, POIs, and satellite imagery.

We standardized AOIs to address semantic inconsistency and topological errors. First, we filtered the raw OSM dataset by interpreting key–value semantics (e.g., *landuse*, *amenity*, *leisure*) to exclude AOIs unrelated to physical building or parcel functions. We then repaired invalid geometries by correcting self-intersections and unclosed polygons. We cleaned and spatially associated POIs to mitigate noise in the massive POI dataset, including duplicates and non-physical tags (*e.g.*, transport stops). We first screened granular POI types to exclude categories irrelevant to architectural function (e.g., bus stops, road intersections), and filtered anomalies by removing low-confidence records or entries lacking functional attributes (e.g., place names without use information). We then assigned POIs to building footprints using a distance-based rule designed for dense urban environments: a POI was associated with a building if its Euclidean distance to the building footprint boundary was less than half the minimum inter-building distance, where the minimum inter-building distance is defined as the shortest distance from the building edge to any adjacent building.

Supplementary Methods 2: Rule-based labeling: extracting high-confidence building samples

We derive high-confidence supervision anchors through a rule-based labeling pipeline that integrates AOI geometry with POI evidence under a two-tier representation. Contextual (neighborhood) labels are assigned in a top-down manner from standardized AOIs: a building receives a preliminary contextual label if the overlap between its footprint and an AOI category exceeds 50% of footprint area. When a building intersects multiple AOI types, we enforce a priority order reflecting institutional encapsulation and functional dominance: Education > Public Open Space > Residential > Commercial > Industrial > Transport > Administrative > Medical. Under this rule, a building overlapping both Education and Residential AOIs is assigned the Education context, consistent with campus-contained functions.

Intrinsic (building) labels are refined within the contextual setting using multi-source fusion rules (**Supplementary Table S2**). If AOIs provide explicit building-level tags (e.g., *building=dormitory*) and overlap a footprint by more than 90%, the intrinsic label is assigned directly from the AOI tag. Otherwise, intrinsic use is inferred from proximal POIs subject to an adaptive consistency filter: we consider only POIs within a distance threshold defined as half the minimum inter-building distance, and apply context-aware mapping logic to resolve ambiguity. For example, within an Education context, POIs tagged as restaurants or supermarkets assign an intrinsic Commercial label (Education–Commercial), whereas dormitory tags assign an intrinsic Residential label (Education–Residential). In Industrial contexts, life-service POIs identify Industrial–Commercial support facilities.

To train BHENet, we extract a high-confidence subset by retaining only buildings with an unambiguous contextual label and intrinsic labels corroborated by either explicit AOI building tags or coherent, high-confidence POI evidence. This procedure yields approximately 30 million high-confidence labeled buildings nationwide. As shown in **Supplementary Fig. S5**, the dataset achieves comprehensive geographic coverage across all city tiers (**Supplementary Fig. S5c**) and sufficient sample volumes even for long-tail minority categories (**Supplementary Fig. S5a, b**). We mitigate class imbalance via stratified sampling across the 29 Context-Intrinsic combinations, using 70% of samples for training and 30% for testing. As an internal audit of the high-confidence supervision anchors, manual validation against a benchmark of 300 randomly selected buildings using Google Earth imagery, Baidu Street View, and Amap details yields accuracies of 96% for contextual labels and 90% for intrinsic labels, supporting the reliability of the supervision anchors.

Supplementary Methods 3: BHENet architecture, training, and evaluation

BHENet is a multimodal classifier that predicts the neighborhood context (8 classes) and context–intrinsic building-function labels (29 classes) for each target building using three complementary inputs. For the textual modality, we retrieve POIs within a 500 m radius of the building centroid; to reduce redundancy and computation, we retain the two nearest POIs per functional category and tokenize their names, types, and distances as the text sequence. For the visual modality, we crop a high-resolution aerial image patch of 1024 × 1024 pixels centered on the building centroid using a fixed 200 m window and normalize it for spectral consistency. For auxiliary vector features, we embed local road-network and building-footprint geometry descriptors to provide morphological context.

The network comprises three parallel feature extractors followed by a fusion module and dual classification heads (**Fig. 4**). The text branch uses a LLaMA-based encoder to derive semantic representations from the POI sequence, and the image branch uses the Segment Anything Model (SAM) image encoder to extract spatial and textural features from the aerial patch. We fuse modalities using a cross-attention mechanism that aligns text and image features through relative position encoding and learns modality weights adaptively; auxiliary vector features are embedded via a CNN and integrated into the fused representation. The resulting multimodal embedding is passed to two heads that jointly predict the contextual class and the intrinsic class.

To effectively integrate heterogeneous modalities, we designed a specialized multi-stage fusion architecture (**Supplementary Fig. S6**). First, to ensure geometric correspondence between Textual Features (from POIs) and Visual Features (from imagery), we apply Position Embedding Fusion. Spatially explicit relative position encodings are injected into both feature sets to align their coordinate reference frames, enabling the model to capture the relative spatial context of semantic signals. The aligned features are then processed via a Vision-Language Feature Fusion block using a cross-attention mechanism. Here, visual features serve as the Query Mapping (QM), while textual semantics provide the Key Mapping (KM) and Value Mapping (VM). This structure dynamically weights textual cues based on their relevance to visual patterns, extracting shared information and mitigating the limitations of unimodal representation.

Subsequently, to bridge the semantic gap between raw vector data and high-level features, Auxiliary Data (e.g., road networks, footprints) undergoes Data Embedding via a convolutional neural network (CNN) layer, transforming it into a compatible latent space. These embedded features are integrated through a dedicated Auxiliary Data Fusion module. We employ a parallel decision strategy where each auxiliary data type is processed by an independent Auxiliary Decision Layer (1 to n). This design allows the model to optimize the contribution of each specific data source (Decision Support Data, DSD) independently before aggregation, ensuring that critical morphological constraints are preserved without being overwhelmed by dominant visual or textual signals.

The aggregated multimodal representation is finally refined by a Self-Attention Module. By modeling global dependencies and recalibrating feature importance across all modalities, this module enhances the model's focus on discriminative signals while suppressing noise. Finally, the reconstructed feature vector is passed to the classification head to generate the Output, predicting both contextual and intrinsic functional classes. This hierarchical fusion strategy combines semantic, visual and morphological inputs in a coordinated manner to improve classification accuracy.

To accommodate spatial heterogeneity, we train BHENet with a regional distributed strategy. We first pre-train a national model on the full dataset, fine-tuning the LLaMA and SAM backbones with LoRA to adapt them to the domain while limiting compute, and training the fusion module and classification heads from scratch. We then fine-tune the model within each of the 34 provincial-level regions to capture localized architectural patterns and semantic conventions. Training uses AdamW with a learning rate of 1e−4 and batch size 4, with early stopping to mitigate overfitting.

An internal held-out supervision-anchor validation set, comprising 30% of the rule-derived high-confidence labels ($N \approx 9$ million), was reserved from model training and used only for model selection and diagnostic evaluation. On this non-training set, BHENet achieved accuracies of 95.85% for the neighborhood-context prediction and 96.09% for the building-scale context–intrinsic prediction. These internal validation metrics characterize generalization within the rule-derived supervision system rather than the final mapped product. The final nationwide map was independently assessed using a stratified national validation sample drawn from the provincial building-function outputs. Candidate buildings were sampled across BF1–BF2 functional strata, cities and spatial bins, after excluding very small footprints (<15 m^2), and 3,024 buildings covering all provincial-level regions were manually interpreted using high-resolution imagery, street-view, map evidence, and POI data and field survey findings. This validation yielded overall accuracies of 90.77% at the neighborhood scale and 86.97% at the building scale, with province-level accuracy estimates further computed to assess spatial consistency (**Extended Data Fig. 2**).

To dissect the contribution of each data modality, we conducted a perturbation-based occlusion analysis on the internal held-out supervision-anchor test set (**Supplementary Table S3**). Results indicate that textual inputs are the dominant predictor for most categories (mean contribution ∼75.7%), confirming the semantic richness of POI data. However, visual imagery proves critical for morphology-dependent categories, most notably for Residential buildings (contribution ∼72.7%), where distinctive architectural patterns (e.g., repeating slab blocks) offer stronger cues than sparse POI signals.

Furthermore, using the same internal held-out supervision-anchor test set, we validated the architectural design through an ablation study of the fusion mechanism (**Supplementary Table S4**). Replacing simple feature concatenation with our proposed cross-attention fusion yielded a performance gain of ~1.0% (OA increasing from 91.81% to 92.84%). The full architecture—integrating position embedding, cross-attention, and auxiliary data fusion—achieved the highest internal held-out OA (95.85%), underscoring the necessity of spatially aware multimodal integration for resolving complex urban functional semantics.

Supplementary Methods 4: BFMI computation and scale divergence tests

SM4.1 BFMI definition and weighting

We quantify functional mixing on regular grid cells and use 2 km × 2 km grids in the main analysis, consistent with the main text. For each grid cell g, we aggregate function-specific weights $w_{g,i}$ for categories $i = 1, \dots, F$ and compute shares $p_{g,i} = w_{g,i} / \sum_{i=1}^{F} w_{g,i}$. BFMI is defined as normalized Shannon entropy,

$$BMFI_g = -\sum_{i=1}^{F} p_{g,i} \, log_2 \, p_{g,i} \Big/ log_2 \, F \quad \in [0,1]$$

where absent categories take $p_{g,i} = 0$ with the convention $0 \log_2 0 = 0$. To represent functional intensity rather than counts, we use volume-weighted shares as the baseline.

For building b, volume is $V_b = A_b \times H_b$, where A_b is footprint area and H_b is height; the category weight in cell g is $w_{g,i} = \sum_{b \in (g,i)} V_b$.

To verify the necessity and robustness of this volumetric approach, we conducted a rigorous sensitivity analysis comparing the baseline volume-weighted index ($BFMI_{Vol}$) against an area-weighted alternative ($BFMI_{Area}$), computed using footprint area alone ($w_{g,i} = \sum_{b \in (g,i)} A_b$). Globally, the two metrics exhibit exceptional consistency, with a Pearson correlation coefficient of 0.993, confirming that height uncertainty introduces minimal systematic noise to the relative ordering of functional intensity (**Supplementary Fig. S7a**). However, the scatter distribution reveals a distinct height-dependent deviation: grids with lower mean building heights tend to cluster above the 1:1 diagonal ($BFMI_{Vol} > BFMI_{Area}$), while those with significant vertical development shift downward.

This structural divergence follows a systematic negative gradient with building height, revealing two critical morphological mechanisms that justify the use of volume weighting (**Supplementary Fig. S7b**). In low-rise zones (mean height < 16 m), volume weighting provides a "density correction": the mean divergence is consistently positive, peaking at +0.014 (95% CI 0.013–0.014) in the 6–8 m range. This indicates that area-based metrics often underestimate mixing in industrial or peri-urban districts by allowing large-footprint structures (e.g., single-story factories) to mask the functional contribution of multi-story amenities (e.g., dormitories), which volume weighting correctly amplifies.

Conversely, in high-rise zones (> 16 m), a "vertical specialization" effect emerges. The divergence turns negative beyond 16 m ($\Delta \approx -0.002$) and deepens to -0.010 (95% CI -0.011 to -0.009) in grids with mean heights of 20–22 m. This reflects the tendency of vertical urbanization to concentrate dominant functions in mono-functional towers (e.g., pure CBD office blocks or high-rise residential slabs), creating a volume-dominant signal that suppresses the superficial diversity of low-rise podiums. The 3D spatial profiling of Xi'an and Hefei (**Supplementary Fig. S7c, d**) further corroborates this pattern, where negative divergence precisely maps to highly specialized CBD cores. By capturing these opposing forces—density correction in the periphery and vertical specialization in the core—volume weighting offers a more physically representative measure of functional intensity than area-based proxies.

SM4.2 Neighborhood versus building mixing: workflows and scale-divergence tests

To quantify whether “mixing” captures distinct mechanisms across functional scales, we compute neighborhood-scale and building-scale BFMI on the same grid system. Neighborhood-scale mixing uses the 8-class neighborhood context and yields $BFMI_g^N$ by aggregating weights across neighborhood-context classes within each grid cell. Building-scale mixing uses the 12-class analytical aggregation derived from the context–intrinsic combinations and yields $BFMI_g^B$ by aggregating weights across analytical building-scale categories. We define scale divergence as $\Delta BFMI_g = BFMI_g^B - BFMI_g^N$ and evaluate it through correlations, quantile agreement, and center–periphery ring profiles to identify where district-level homogeneity masks substantial building-level mixing and whether this divergence follows systematic spatial gradients. Ring profiling follows a grid-first workflow: we compute $BFMI_g$ at the grid level and then average it within distance bands from the city center.

Results reveal that building-scale mixing systematically exceeds neighborhood-scale mixing, confirming that coarse-grained land-use metrics significantly underestimate urban functional complexity. The hexbin density analysis (**Supplementary Fig. S8d**) shows that while the two indices are strongly correlated (Pearson's r = 0.842), a substantial proportion of grid cells (70.2%) exhibit positive scale divergence (building-scale mixing greater than neighborhood-scale mixing), with a global mean divergence of 0.061. This structural deviation indicates that functionally homogeneous neighborhoods—such as industrial parks or residential estates—frequently host diverse, fine-grained activities (e.g., internal dormitories or retail) that are only capturable at the building resolution.

The spatial profiling of scale divergence uncovers striking structural differences in urban development patterns across city tiers (**Supplementary Fig. S8a-c**). In major metropolitan areas (Tier 1), high levels of positive divergence persist from the core to the periphery. The mean divergence remains robust (approximately 0.09–0.10) even at distances of 30–35 km from the center, and the proportion of grids with negative divergence remains consistently low (below 20%). This pattern reflects a high-density development mode where vertical functional integration is pervasive, mitigating mono-functional sprawl even in suburban zones. In contrast, lower-tier cities exhibit a rapid decay in divergence. While central areas in Tier 3 cities show moderate hidden mixing (mean divergence of 0.077 at 2.5 km), this value drops sharply to 0.044 beyond 7.5 km. Crucially, this decay is driven by functional polarization: the proportion of grids with negative divergence nearly doubles from the core (22.5%) to the suburbs (approximately 40%). This suggests that suburban expansion in lower-tier cities is characterized by coarser, segregated land-use parcels (e.g., pure industrial zones or sleeping towns) lacking the micro-scale complementarity found in Tier 1 cities.

The systematic prevalence of positive divergence and its distinct spatial gradients confirm that neighborhood-scale metrics alone are insufficient to characterize the "effective mixing" relevant to carbon performance. The building-scale approach is therefore essential, particularly for capturing the hidden mixing in Tier 1 peripheries and identifying the functional hollowing-out in lower-tier suburbs.

SM4.3 Sensitivity and robustness across taxonomic and spatial granularities

We rigorously assessed the stability of the functional mixing index (BFMI) and its associated decarbonization thresholds across taxonomic granularity, comparing the 8-class neighborhood context with the 12-class analytical aggregation, and across spatial resolution, comparing 1 km, 2 km and 4 km grid cells. This dual-sensitivity analysis aims to quantify the Modifiable Areal Unit Problem (MAUP) and verify that the observed "decarbonization window" reflects an intrinsic property of the urban system rather than an artifact of spatial discretization.

We first examined how changing the grid resolution alters the statistical distribution of mixing intensity (**Supplementary Fig. S9a, b and Supplementary Table S5**). At the finest resolution (1 km), the urban fabric is often partitioned into functionally homogeneous fragments. This "over-segmentation" results in a significantly lower BFMI across both schemas. For the 12-class analytical aggregation, the 1 km grid yields a mean BFMI of 0.360 (Median 0.371, IQR 0.234), which is systematically lower than the coarser scales (difference in mean ≈ -0.12). The left-skewed distribution confirms that small windows (1 km²) struggle to capture diverse functional combinations, artificially inflating the proportion of low-entropy (single-function) cells. In contrast, the BFMI distributions at 2 km and 4 km resolutions exhibit remarkable consistency. For the 12-class analytical aggregation, the mean BFMI stabilizes at 0.476 (2 km) and 0.485 (4 km), with similar dispersion patterns (standard deviation ≈ 0.15–0.16). The strong overlap in kernel density estimates (KDE) between these two scales suggests that the 2 km grid serves as an optimal analytical unit, effectively balancing the capture of local functional diversity with the smoothing of micro-scale fragmentation noise.

The relationship between neighborhood-scale mixing (8-class) and carbon emissions remains structurally robust across resolutions (**Supplementary Fig. S9c**). Consistent with the main text, the fitted response curves at all three scales demonstrate a predominantly monotonic decarbonization trend across the valid data range. Increasing BFMI consistently correlates with reduced emissions, validating the general substitution effect of neighborhood functional balance. A minor deviation is observed at the 1 km scale, where marginal effects turn positive at extreme mixing levels (BFMI > 0.8). However, this rebound is likely negligible for policy interpretation, as the 1 km grid contains sparse data in this high-entropy tail owing to the fragmentation effect described above, making the estimation in this range less structurally representative.

Crucially, the tri-phasic response and the "decarbonization window" identified in the building-scale (12-class) analysis demonstrate high stability across spatial scales (**Supplementary Fig. S9d**). The threshold marking the transition from friction to synergy remains statistically consistent. The lower bound appears at BFMI ≈ 0.141 ± 0.027 (95% CI, 0.112–0.166) in the 2 km baseline. At 4 km, this threshold shifts slightly to 0.121 ± 0.038 (95% CI, 0.078–0.154), and at 1 km to 0.090 ± 0.406 (95% CI, 0.005–0.817). The substantial overlap of confidence intervals confirms that the onset of agglomeration benefits is not an artifact of scale. The critical turning point where benefits plateau or reverse is equally robust. The upper bound is identified at BFMI ≈ 0.723 ± 0.013 (95% CI, 0.714–0.739) in the 2 km baseline. At 4 km, it shifts to 0.768 ± 0.027 (95% CI, 0.745–0.798), and at 1 km to 0.691 ± 0.079 (95% CI, 0.636–0.794). The slight rightward shift at 4 km (difference ≈ +0.045) is consistent with the MAUP smoothing effect (larger units naturally encapsulate more diversity), yet the threshold consistently maps to the same high-intensity mixing regime relative to the respective distributions. These multi-scalar validations confirm that the "decarbonization window" is a scalable, intrinsic property of the urban functional system, robust to both taxonomic definitions and spatial aggregation units.

Supplementary Methods 5: Localized jobs–housing separation metrics (EAD and JSD)

SM5.1 Definitions using a moving window approach

To capture the realized jobs–housing balance within a daily activity space, we construct two spatially explicit separation metrics—Local EAD and Local JSD—for each 2 km × 2 km grid cell using a local moving window approach, rather than traditional city-wide aggregates. For a focal grid cell i, let S_i be the set of neighboring grid cells within a 10 km radius (representing a typical commuting catchment). Within this local window S_i, we aggregate residential volume R_k and employment volume J_k for all neighbor cells $k \in S_i$, and normalize them into localized spatial distributions:

$$p_R(k) = R_k / \left(\sum_{m \in S_i} R_m \right), \quad p_J(k) = J_k / \left(\sum_{m \in S_i} J_m \right) \quad \text{for } k \in S_i$$

Local EAD (Expected Average Commuting Distance): This metric captures the expected physical separation between housing and jobs within the neighborhood, representing the local circulation demand (movement channel).

$$\mathrm{EAD}_i = \sum_{k \in S_i} \sum_{l \in S_i} d_{k,l} \cdot p_R(k) \cdot p_J(l) \cdot \frac{1}{d_{\max}} \in [0,1]$$

where $d_{k,l}$ is the distance between grid centroids, and $d_{max} = 10km$ is the window radius used for normalization.

Local JSD (Jensen–Shannon Divergence): This metric captures the morphological dissimilarity between the residential distribution $P_i = \left(p_R(k)\right)_{k \in S_i}$ and the employment distribution $Q_i = \left(p_J(k)\right)_{k \in S_i}$ within the window, representing spatial polarization (morphology channel) independent of metric distance. Let $M_i = \frac{1}{2}(P_i + Q_i)$. We define:

$$\mathrm{JSD}_i = \frac{1}{2} D_{\mathrm{KL}}(P_i \parallel M_i) + \frac{1}{2} D_{\mathrm{KL}}(Q_i \parallel M_i)$$

where $D_{\mathrm{KL}}(P \parallel M) = \sum_{k \in S_i} p_R(k) \log_2\big(p_R(k)/m(k)\big)$. Prior to regression, both metrics are standardized to improve coefficient comparability.

SM5.2 Distance computation, stability, and sensitivity

We define $d_{k,l}$ as the Euclidean distance between grid centroids within the local projection. While network distance is conceptually superior, Euclidean distance serves as a robust proxy for spatial impedance at the 2 km grid scale. For numerical stability in JSD, we use the convention $0 \log 0 = 0$ and add a smoothing factor $\epsilon = 10^{-12}$ to avoid division-by-zero.

Employment volume J is defined by aggregating non-residential functional volumes (Industrial, Commercial, Education, Administrative, Transport, Public Open Space, and Medical). To verify that our results reflect general structural constraints rather than artifacts of functional classification, we conducted sensitivity checks by restricting the employment-volume term J to Industrial and Commercial volumes only, representing a core market-employment definition. Comparative analysis revealed strong spatial concordance between the baseline metrics (using all non-residential functions) and the alternative metrics (restricted to Industrial and Commercial volumes). The grid-level Pearson correlation coefficient was $r = 0.925$ ($P < 0.001$) for EAD and $r = 0.742$ ($P < 0.001$) for JSD, indicating that the core spatial structure of jobs–housing separation is robust to functional categorization. Furthermore, substituting these alternative metrics into the SDM regression yielded structurally identical coefficients. The marginal effect of EAD on urban operational carbon remained remarkably stable, shifting marginally from 6,095.2 t C yr^{-1} under the baseline employment definition (95% CI, 5,496.0–6,694.4) per unit of normalized EAD to 6,096.9 t C yr^{-1} under the core market-employment definition (95% CI, 5,497.7–6,696.1). This corresponds to approximately 609.5 t C yr^{-1} per +1 km in EAD, as reported in the main text. Similarly, the JSD coefficient showed negligible variation, from 2,484.1 t C yr^{-1} (95% CI, 1,977.7–2,990.6) per unit of JSD to 2,484.6 t C yr^{-1} (95% CI, 1,978.2–2,991.1), corresponding to 248.4 t C yr^{-1} per +0.1 in JSD. This quantitative stability confirms that the identified mobility penalties are driven by the fundamental spatial disconnect between living and working spaces, rather than artifacts of specific functional definitions.

Supplementary Methods 6: Spatial econometric model, inference, and diagnostics

We estimate the association between functional mixing and ODIAC-derived annual grid-level urban operational carbon using a 2 km × 2 km grid system defined over each city's built-up area, while explicitly accounting for cross-grid interactions. The outcome y is annual grid-level territorial urban operational carbon, expressed in t C yr^{-1}. To jointly capture outcome spillovers and covariate spillovers, we adopt a Spatial Durbin panel model (SDM) with fixed effects,

$$y = \rho W y + X\beta + WX\theta + \alpha_{city} + \alpha_{ring} + \varepsilon$$

where W is the spatial-weights matrix, ρ is the spatial lag parameter for the dependent variable, and WX allows regressors to propagate through spatial spillovers.

City fixed effects α_{city} absorb time-invariant differences across cities (*e.g.*, energy systems, industrial bases, and reporting heterogeneity). Annular ring fixed effects α_{ring} absorb systematic center–periphery gradients and unobserved spatial heterogeneity common across cities (*e.g.*, infrastructure intensity and urban hierarchy). Rings are defined as 5-km distance bands anchored at the identified city center(s); for polycentric cities, ring membership and distance are assigned using the nearest-center rule to preserve locally experienced center–periphery structure.

Functional mixing (BFMI) is the core regressor. To allow for saturation and reversal, BFMI enters the model through an orthogonal polynomial basis to flexibly recover non-linear responses while limiting multicollinearity. Let x denote centered and standardized BFMI. We represent non-linearity using a cubic Legendre basis $P_1(x), P_2(x), P_3(x)$ and allow the corresponding spatially lagged terms to capture spillovers of mixing,

$$y = \rho W y + \sum_{m=1}^{3} \beta_m P_m(x) + \sum_{m=1}^{3} \theta_m P_m(x) + Z\gamma + WZ\eta + \alpha_{city} + \alpha_{ring} + \varepsilon$$

where Z denotes the remaining covariates. As a robustness check on functional form, we replace the orthogonal-polynomial representation with natural cubic splines for BFMI and verify the stability of the key turning points and marginal-effect profiles.

The baseline spatial-weights matrix W is constructed using a k-nearest-neighbor specification (k = 8) to capture local interactions and near-range accessibility, ensuring consistent connectivity density across the grid system regardless of geometric irregularities. To rigorously test sensitivity to this assumed spatial structure, we re-estimated the SDM using alternative specifications, including Queen contiguity (adjacency-based) weights and Inverse Distance Weighting (IDW) with a 10 km cutoff. All matrices are row-standardized for comparability, and alternatives are evaluated under the identical fixed-effects structure and covariate set to isolate the influence of spatial connectivity assumptions on spillover estimates.

The covariate vector Z controls for functional composition, transport infrastructure, built-environment intensity, and topographic constraints (**Extended Data Table 1**). Specifically, we include class-specific functional volumes to control for scale and composition effects; road network density and population density as proxies for infrastructure and demographic intensity; night-time light intensity (NTL) to proxy overall socio-economic activity; and elevation, slope, and relief to account for terrain-induced energy penalties. Crucially, we incorporate Local EAD and Local JSD to operationalize system-level jobs–housing separation constraints along movement and morphology channels, respectively. Including both metrics isolates the marginal effect of functional mixing from broader jobs–housing mismatches, reducing omitted-variable bias related to unobserved commuting structures. All continuous covariates are standardized (z-score) to facilitate coefficient comparability and convergence.

Since SDM coefficients do not directly represent marginal impacts due to spatial feedback loops, we computed the Direct, Indirect, and Total marginal effects implied by the simultaneous system (reported in **Extended Data Table 2**). We derived the non-linear marginal effect curve of BFMI from the estimated Legendre polynomial coefficients, adopting the Average Marginal Effects (AME) as the primary summary metric for both the baseline analysis and all robustness experiments to capture the representative system-level response. We identified the critical thresholds defining the tri-phasic response as the two zero-crossings of the BFMI total marginal-effect curve, corresponding to the lower-bound friction end-point and the upper-bound saturation point. Statistical inference for marginal effects and threshold locations relies on the delta method, with standard errors clustered at the city level to robustly accommodate within-city dependence and heteroskedasticity. Crucially, we verify that the estimated response profiles are structurally invariant to the choice of marginal-effect summary; comparing the baseline AME against Marginal Effects at the Mean (MEM) and at Representative values (MER) reveals negligible deviations, with identified turning-point thresholds differing by < 0.01 across metrics. The lower and upper thresholds of the abatement window therefore correspond to the two zero-crossings of the BFMI total marginal-effect curve. In addition, the point at which the derivative of the marginal-effect curve reaches its most negative value is used to identify the steepest strengthening of the marginal-abatement response; in the baseline model, this occurs at BFMI ≈ 0.437.

To validate model integrity, we rigorously diagnosed multicollinearity and residual spatial dependence. Variance Inflation Factors (VIFs) confirmed that correlations among functional-volume controls and activity proxies remained within acceptable bounds (VIF < 5 for all core regressors). Post-estimation diagnostics, including Moran's I on residuals, verified that the Spatial Durbin specification effectively absorbed latent spatial autocorrelation ($P > 0.1$ for residual dependence). We further conducted a systematic battery of robustness checks (**Supplementary Table S6**) spanning four dimensions to ensure the findings reflect intrinsic system-level regularities rather than modeling artifacts: spatial connectivity (comparing KNN-8 against Queen and IDW matrices); functional form (replacing Legendre polynomials with restricted cubic splines); spatial scale (replicating the analysis at 1 km and 4 km resolutions to address MAUP); and structural heterogeneity (stratifying the sample by city tier). Results across all specifications consistently reproduced the characteristic friction–synergy–saturation response, confirming the structural stability of the reported thresholds.

Supplementary Methods 7: Counterfactual alignment and abatement estimation

We translate the main-text BFMI–urban operational carbon marginal-effect function into counterfactual abatement potentials on the same 2 km × 2 km built-up grid system. For each grid cell g, $BMFI_g$ denotes the observed 2020 building-scale BFMI, computed using the volume-weighted 12-class analytical schema described in **Supplementary Methods 4.1**. $BFMI_g^{cf}$ denotes the counterfactual BFMI assigned to the same grid cell under a specified alignment scenario.

Under ideal alignment, $BFMI_g^{cf}$ is set to the upper turning point of the fitted response curve before saturation, corresponding to BFMI ≈ 0.723 in the baseline model. Under cost-effective alignment, $BFMI_g^{cf}$ is defined as:

$$BFMI_g^{cf} = \begin{cases} BFMI_g + 0.30, & BFMI_g < 0.30, \\ BFMI_g - 0.25, & BFMI_g > 0.75, \\ BFMI_g, & \text{otherwise.} \end{cases}$$

This rule raises under-mixed grids into the synergy window and pulls hyper-mixed grids away from the saturation regime, while leaving intermediate grids unchanged.

Given a shift $BFMI_g \rightarrow BFMI_g^{cf}$, we compute the implied change in urban operational carbon by discretely integrating the marginal-effect function along the BFMI path. We bin BFMI over [0, 1], compute the average marginal effect $\widehat{AME}(b)$ within each bin b, and approximate grid-level abatement by a Riemann sum over traversed bins:

$$\Delta \mathrm{C}(g) \approx \sum_{b \in \mathcal{B}(g)} \widehat{\mathrm{AME}}(b)\, \Delta \mathrm{BFMI}_b$$

where $B(g)$ denotes the set of bins crossed between observed and counterfactual BFMI, and $\Delta BFMI_b$ is the traversed width within bin b. Uncertainty is propagated from the SDM parameters to the bin-specific marginal effects $\widehat{AME}(b)$ and the integrated ΔC estimates using the delta method, with standard errors clustered at the city level to account for within-city dependence. To rigorously bound these counterfactuals, we perform sensitivity checks using Marginal Effects at Representative values (MER) and at the Sample Median (MEM). These alternative summaries yield aggregate abatement totals that deviate by less than 12.2% from the baseline AME estimates (e.g., varying the national ideal-alignment potential from 272.4 to 340.5 Mt C yr^{-1}), providing a compact bracket on counterfactual potentials and ensuring that the identified mitigation wedge is not an artifact of tail-sensitive averaging. Grid-level changes are aggregated to city and national totals and summarized by city tier, with detailed results and spatial patterns discussed in **Supplementary Note 4**. Aggregate

CO_2-equivalent values reported for interpretation are obtained by multiplying carbon-mass estimates by the molecular-weight ratio 44/12.

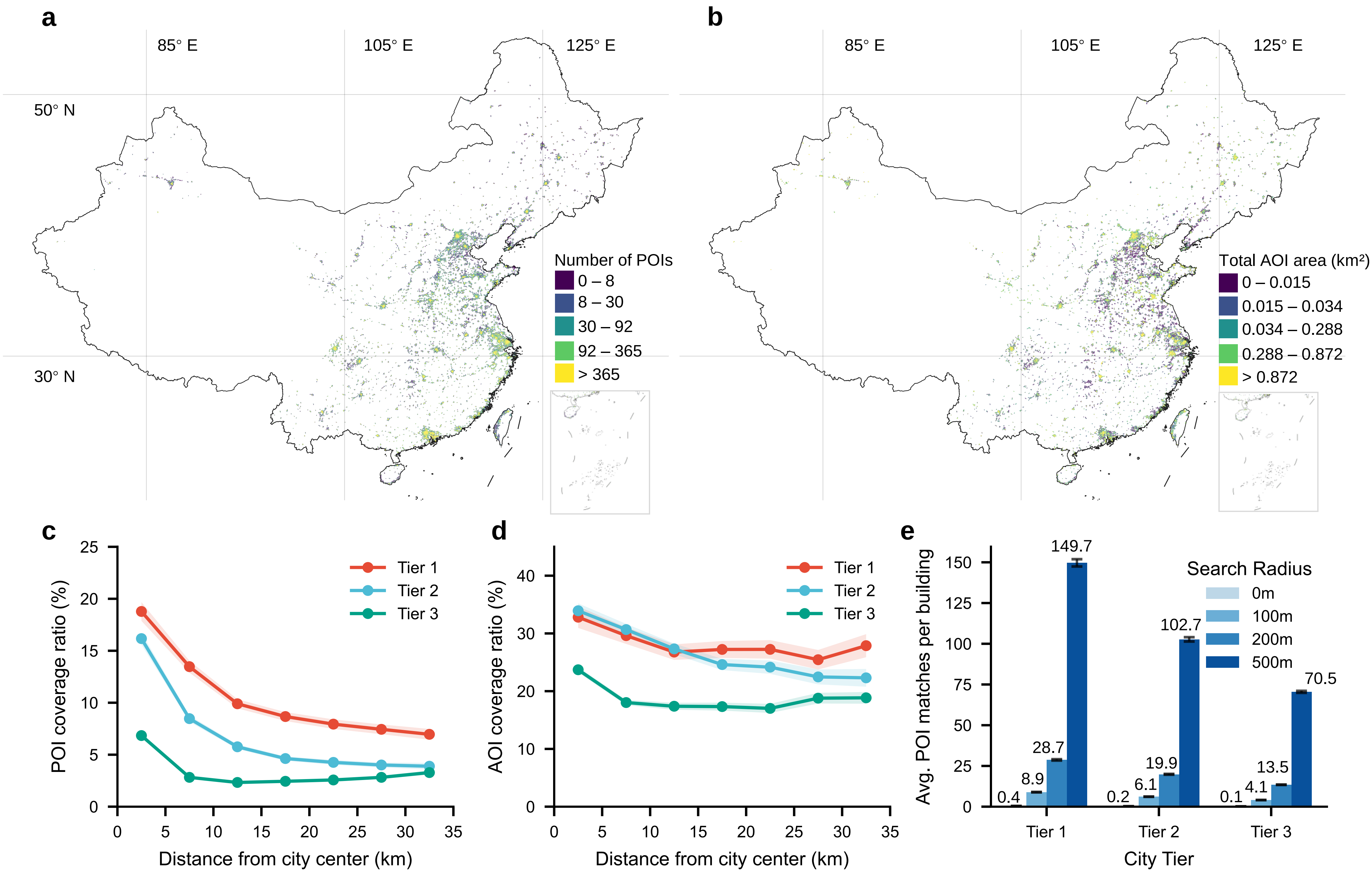
Supplementary Methods 8: Global comparative analysis (31 cities): harmonization and comparability

To assess the external validity of the building-scale mixing mechanism beyond the high-density Chinese context, we compiled a comparative dataset covering 31 major global cities spanning multiple continents (**Supplementary Table S7**). The sample was strategically stratified to maximize morphological and developmental diversity, spanning: (i) established compact metropolises (e.g., London, Paris and New York); (ii) automobile-dependent sprawling regions (e.g., Los Angeles, Dallas); and (iii) smaller-scale urban settlements (e.g., Canberra, Zürich) that offer a counterpoint to mega-city dynamics. This structural heterogeneity allows us to rigorously test whether the "decarbonization window" recurs across contrasting urban contexts or is an artifact of China's specific urbanization pathway.

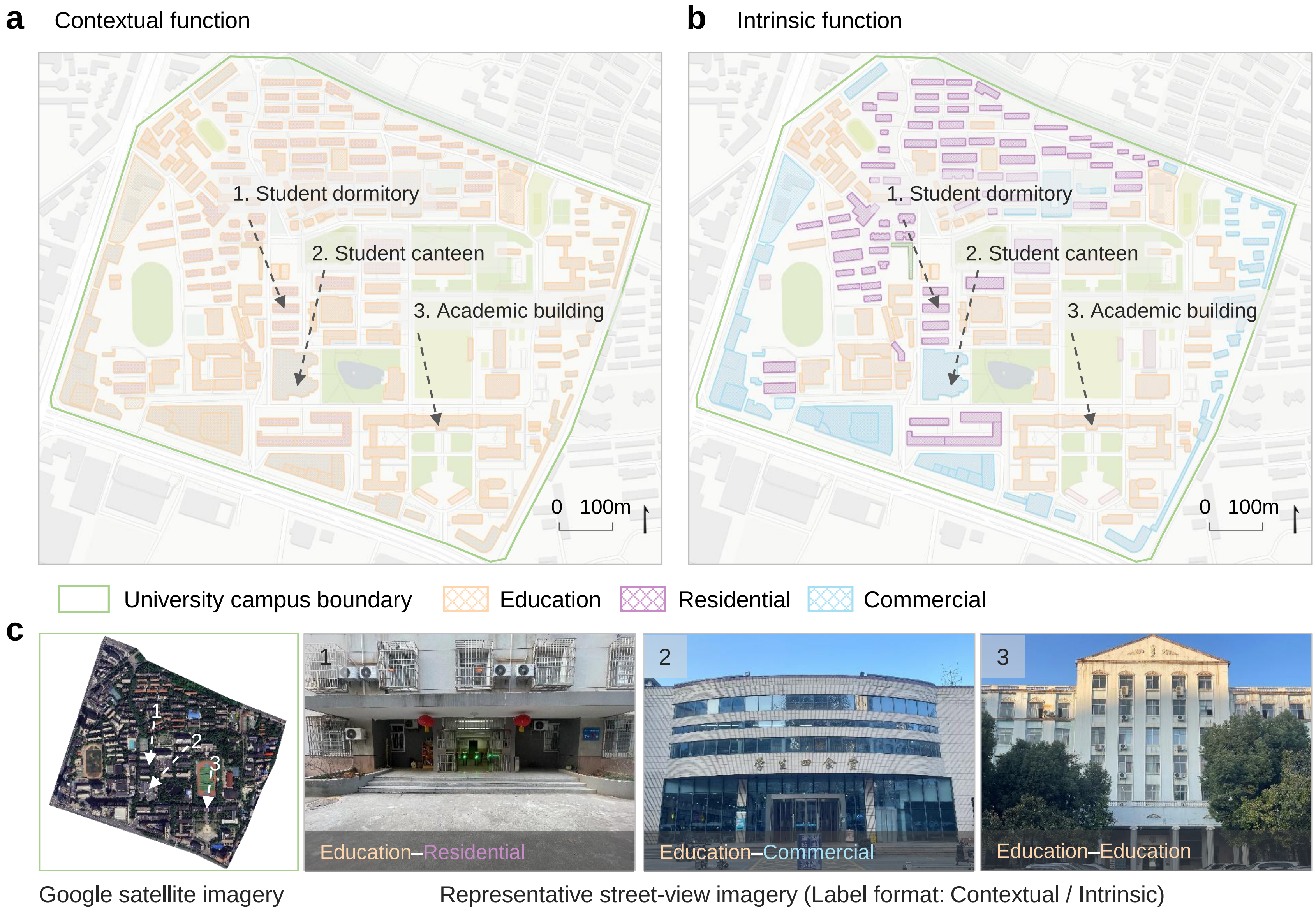
While the analytical framework mirrors the main study, the restricted availability of custom surveying data for global cities necessitated the integration of high-quality global open-source alternatives. We implemented a rigorous harmonization protocol to ensure structural comparability with the Chinese dataset. Specifically, building geometries and demographic data from global open repositories were aggregated to the standardized 2 km × 2 km grid system to mitigate resolution-induced biases (MAUP) and align with the Chinese baseline. The corresponding data sources are listed in

Supplementary Table S8. Crucially, the dependent variable was derived from the same ODIAC2024 product and reported in the same carbon-mass units (t C yr^{-1}), ensuring consistency in territorial land-based carbon measurement. Functional labels were generated using the BHENet model; to accommodate cross-cultural variations, the model (pre-trained on China's multi-modal data) underwent city-specific fine-tuning using local samples from each of the 31 target cities to rigorously adapt to distinct regional architectural styles and POI naming conventions before full-scale deployment.

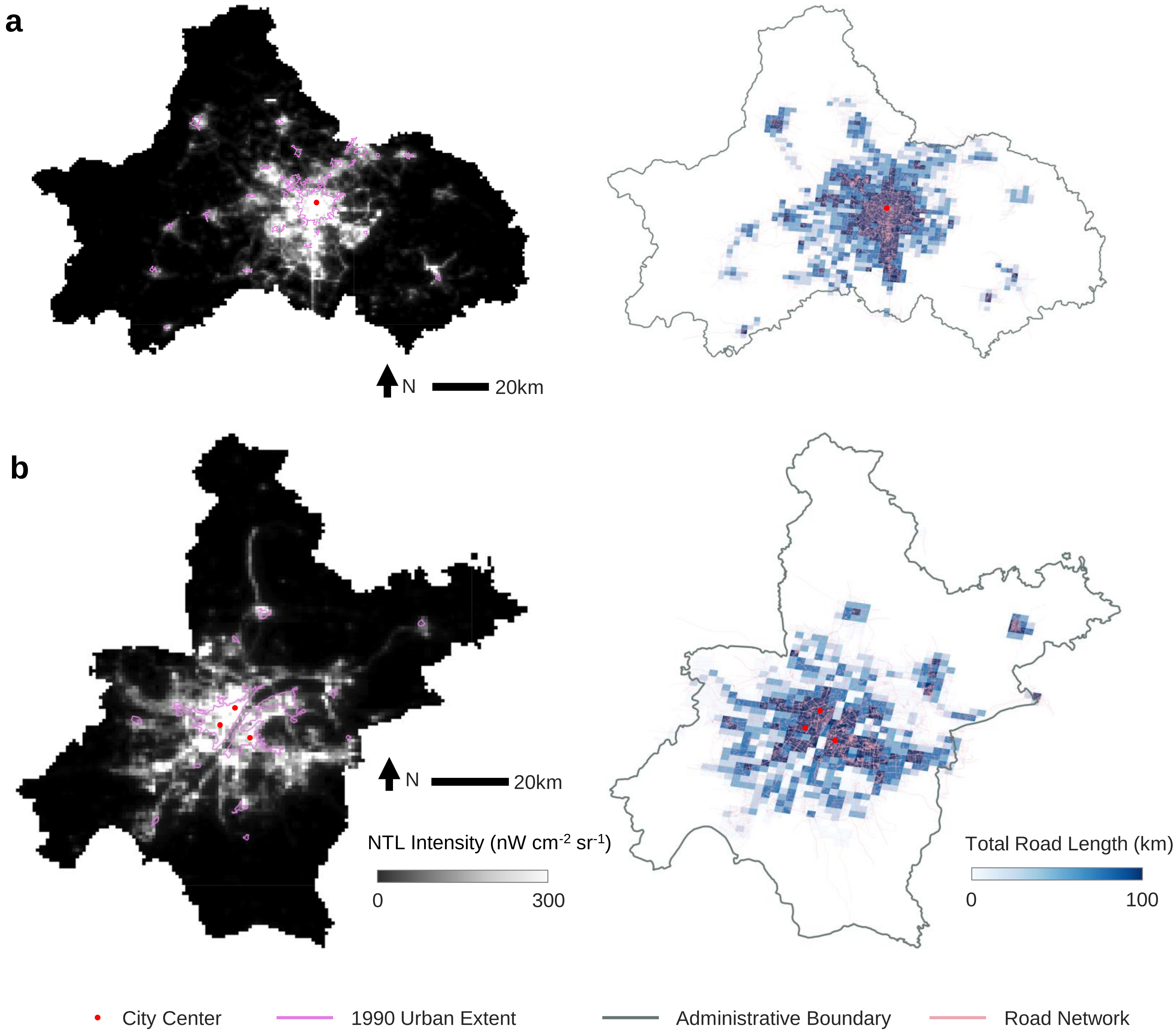
Comparing the marginal effect curves reveals a striking structural convergence: global cities exhibit the same characteristic tri-phasic (S-shaped) response as Chinese cities, supporting the broader cross-context recurrence of the "friction–synergy–saturation" sequence rather than a regional artifact. However, quantitative thresholds reflect distinct urban substrates. While the upper-bound saturation point is remarkably consistent between contexts (BFMI ≈ 0.703 globally vs. 0.723 in China), the lower-bound activation threshold is higher in global cities (0.276; 95% CI, 0.257–0.294) than in China (0.141; 95% CI, 0.112–0.166). This delayed onset of synergy likely reflects the "density-activation gap": the global sample has substantially lower baseline density (mean building volume ≈ 83.5 vs. 441.8 $\times 10^4$ m^3 per grid) and lower urban operational carbon flux (7,833 vs. 20,803 t C yr^{-1}) (see **Supplementary Table S7**). In lower-density environments, a higher critical mass of functional mixing is required to overcome spatial impedance and trigger agglomeration benefits, whereas China's hyper-dense fabric catalyses synergy at much lower mixing levels. Additionally, the absolute magnitude of marginal effects differs, scaling with the respective baseline emission intensities.



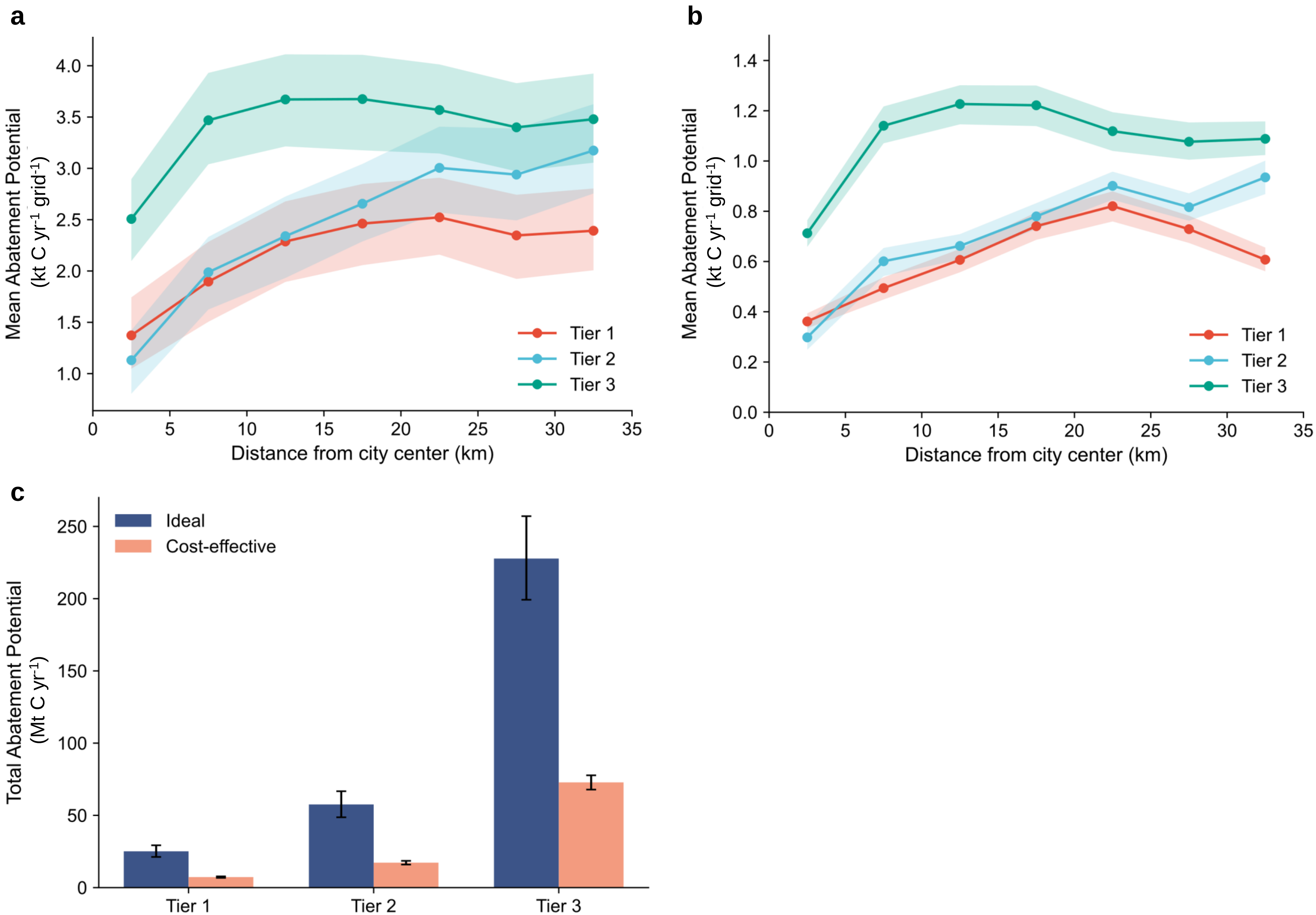
Supplementary Fig. S1 | Spatial heterogeneity and representativeness bias in crowd-sourced geographical data.
a, b, Nationwide spatial distribution of semantic data density, aggregated at the 2 km × 2 km grid scale. Maps display the absolute **Points of Interest (POI) counts** (**a**) and total **Areas of Interest (AOI) coverage** (km^2) (**b**) per grid cell, highlighting the pronounced concentration in eastern coastal regions and urban cores compared with western and peripheral sparsity. **c, d,** Center-periphery gradients of data completeness across the urban hierarchy. Curves illustrate the **coverage ratio**—defined as the percentage of building footprints successfully matched with at least one POI (**c**) or intersecting with an AOI polygon (**d**)—as a function of distance from the city center. Cities are stratified by urban built-up area: **Tier 1** (>3,000 km^2), **Tier 2** (800–3,000 km^2), and **Tier 3** (<800 km^2). The consistent decay in coverage ratios with increasing distance across all tiers confirms a systematic center-periphery bias, which is notably exacerbated in lower-tier cities. **e,** Sensitivity of label availability to spatial matching thresholds. Grouped bars show the average POI matching intensity (number of POIs per building) across city tiers under increasing buffer radii (0 m, 100 m, 200 m, and 500 m). The persistent gap between Tier 1 and Tier 3, even at a 500 m radius, underscores the structural scarcity of semantic data in lower-tier cities. In **c** and **d**, data points represent grid-level means; shaded bands indicate **95% confidence intervals**. In e, error bars denote **standard errors of the mean (s.e.m.)**.



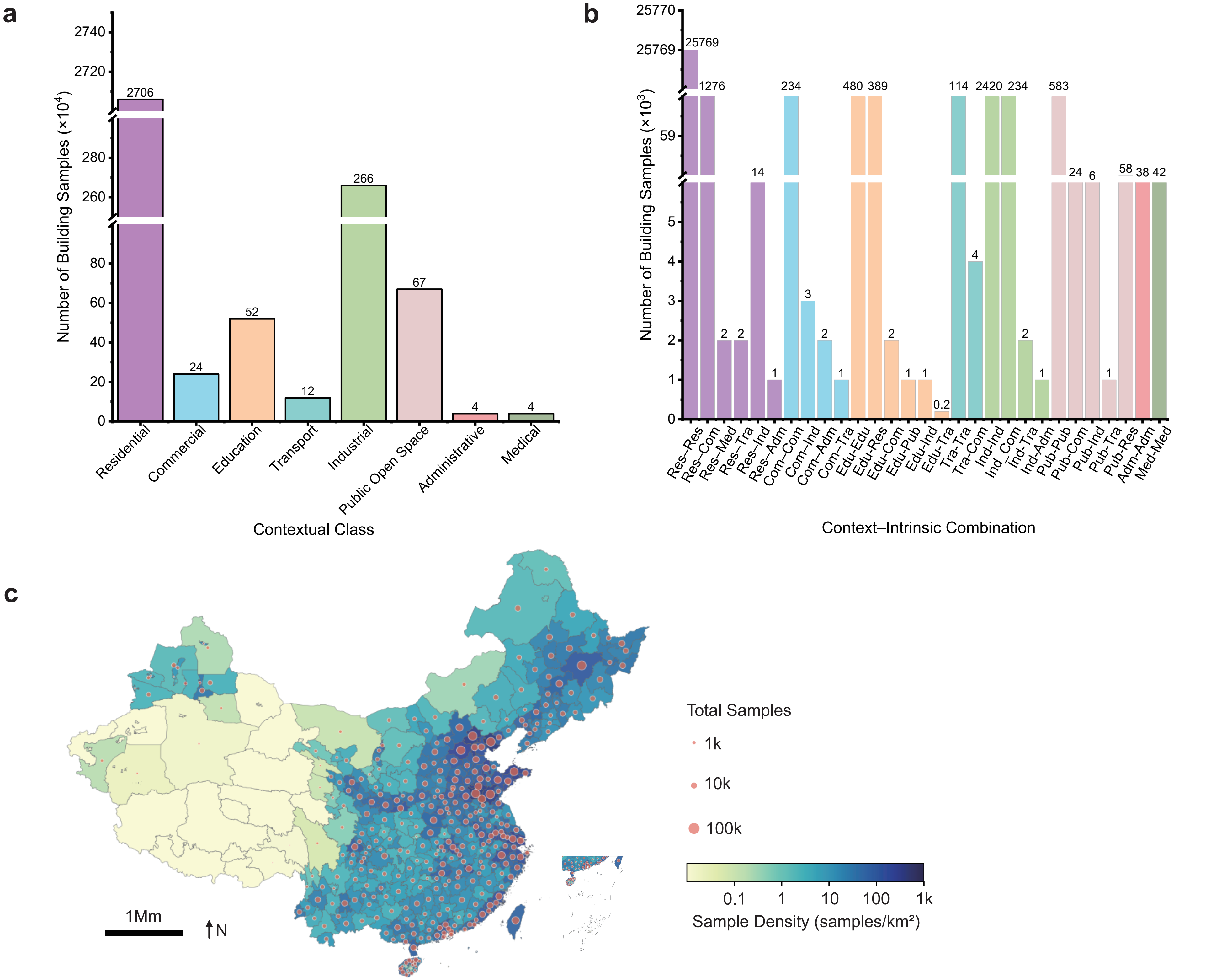
Supplementary Fig. S2 | Multiscale divergence between contextual and intrinsic functional labels.
a, b, Comparison of functional labeling schemas applied to a representative university campus (located at 30.53° N, 114.36° E). **a,** Under the **contextual (neighborhood) schema**, the entire plot is uniformly classified as **Education**, reflecting its dominant land-use designation. **b,** Under the **intrinsic (building) schema**, individual structures exhibit functional differentiation, revealing uses that may either **diverge from** (e.g., **Residential** dormitories, **Commercial** canteens) or **align with** (e.g., **Education** academic buildings) the broader contextual context. This comparison illustrates the limitation of single-label systems in capturing fine-grained functional heterogeneity. **c,** Ground-truth validation using multi-source imagery. Left: **Google satellite imagery** of the campus. Right: **Street-view imagery** of specific structures labeled **1–3** in **a** and **b**, confirming their intrinsic functions: **Building 1** (Student dormitory; Contextual: Education / Intrinsic: Residential), **Building 2** (Canteen; Edu./Com.), and **Building 3** (Academic building; Edu./Edu.). Building footprints are shaded by function according to the color key. Label format in **c** follows the **Contextual/Intrinsic** convention.



Supplementary Fig. S3 | Multi-source spatial validation of city center anchors for monocentric and polycentric urban structures. a, b, Validation of identified city centers (red dots) for a representative monocentric city (Chengdu, **a**) and a polycentric city (Wuhan, **b**). Left panels display the night-time light intensity (background gradient) overlaid with the 1990 urban extent (purple outline), derived from historical impervious surface data. The identified anchors spatially coincide with the historical cores and peak luminosity clusters, confirming their long-standing dominance as organizing nodes. Right panels map the grid-level road density (visualized as total road length per 2-km grid, km grid^{-1}) overlaid with the major road network. High-density road infrastructure converges precisely at the selected anchors. For Wuhan (b), the "Three Towns" structure is captured by three distinct functional anchors (Hankou, Wuchang, and Hanyang) separated by the Yangtze and Han rivers, validating the necessity of the polycentric, nearest-center profiling strategy described in **Supplementary Note 2**.

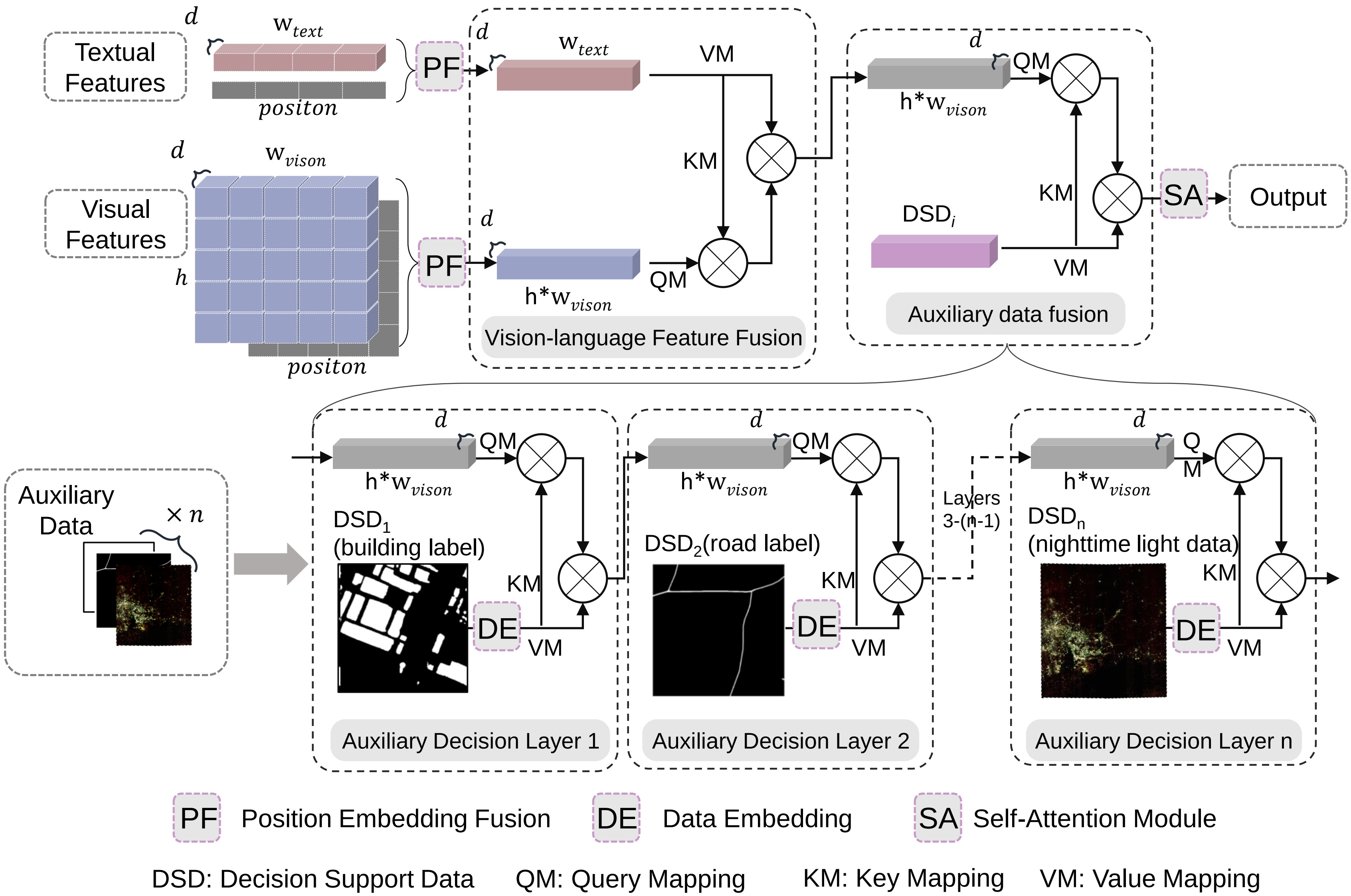


Supplementary Fig. S4 | Spatial distribution and magnitude of counterfactual abatement potentials across city tiers.
a, b, Distance-decay profiles of mean grid-level abatement potential (kt C yr^{-1} $grid^{-1}$) stratified by city tier (Tier 1, red; Tier 2, blue; Tier 3, green). Panels show results for the **Ideal** alignment (**a**), where functional mixing is aligned to the upper turning point before saturation (BFMI ≈ 0.723), and the **Cost-effective** alignment (**b**), which targets only the low- (BFMI < 0.30) and high-mixing (BFMI > 0.75) tails. Data points are plotted at the midpoint of 5-km annular rings (e.g., 2.5 km for the 0–5 km ring); shaded ribbons indicate 95% confidence intervals. **c,** Aggregated total abatement potential (Mt C yr^{-1}) by city tier under both scenarios. Tier 3 cities exhibit the largest aggregate potential due to their extensive stock of under-mixed grids, whereas Tier 1 cities show higher efficiency per unit of adjustment but limited total volume. Error bars represent 95% confidence intervals derived from Spatial Durbin Model parameters using the delta method. All estimates are based on the grids retained for abatement estimation after applying the carbon-data and model-validity filters (N = 91,111).

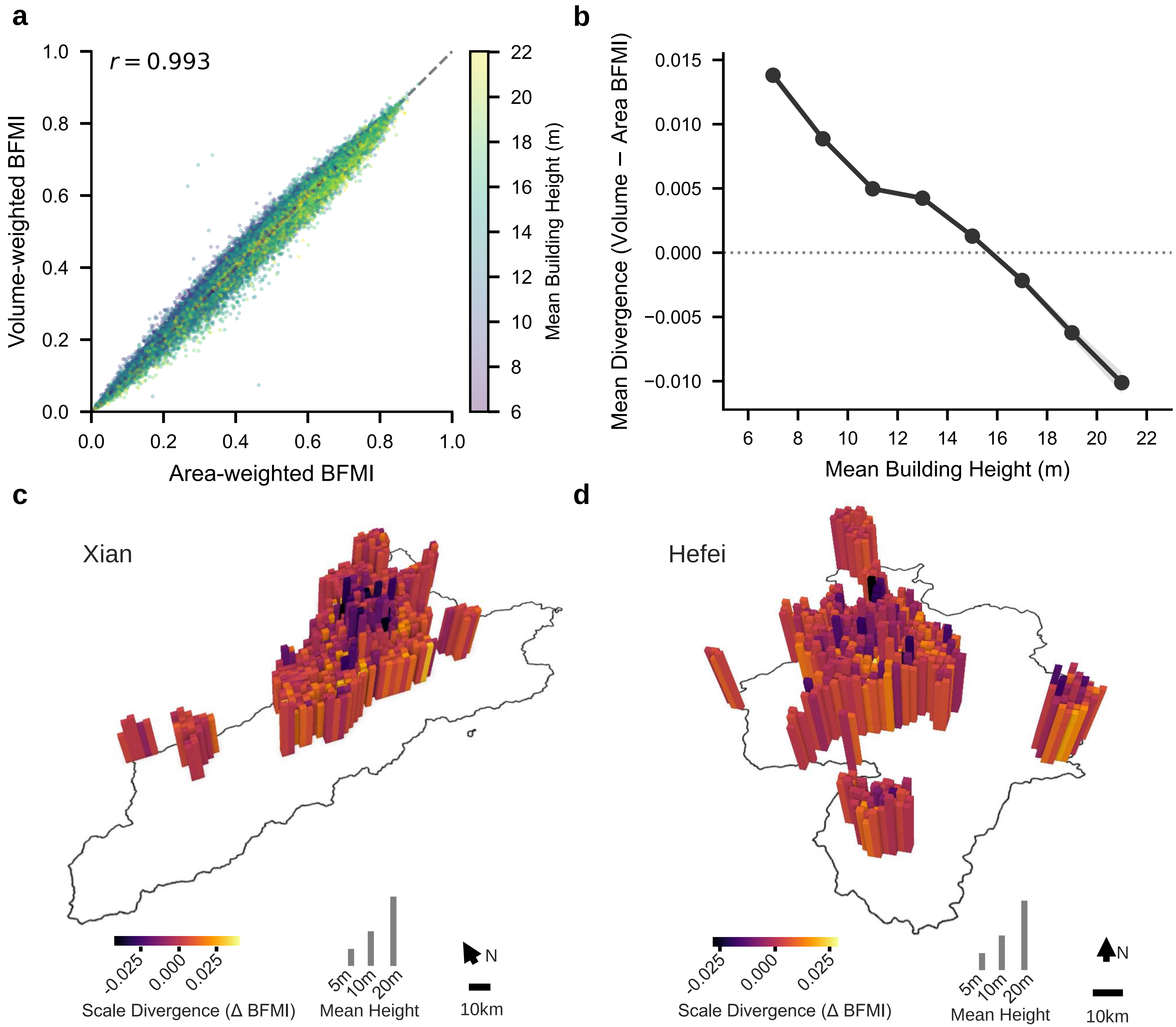


Supplementary Fig. S5 | Nationwide distribution and typological representativeness of the high-confidence building sample dataset.

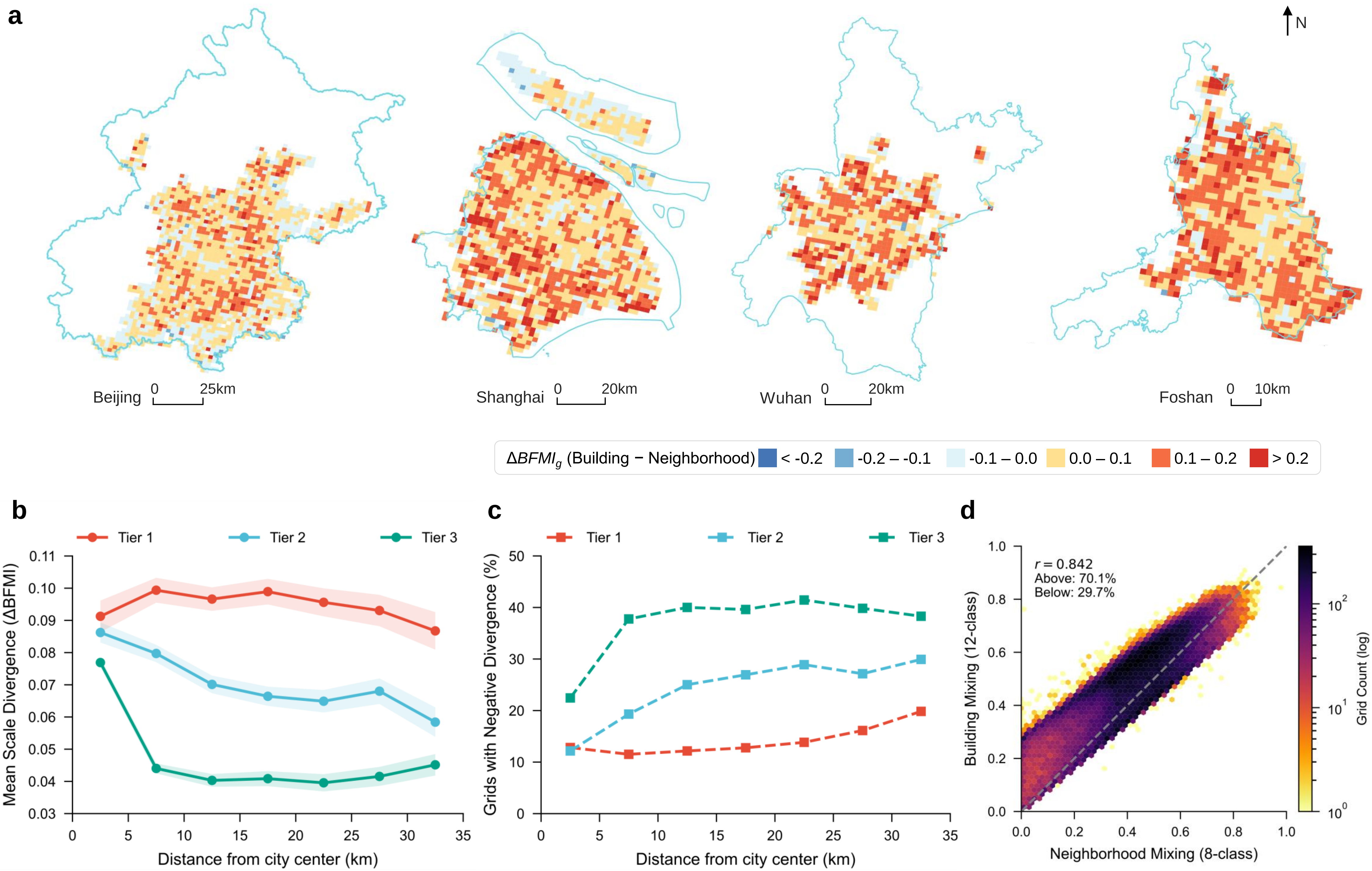
a, Distribution of sample counts across the 8 contextual (neighborhood-scale) functional categories. **b,** Distribution of sample counts across the 29 fine-grained Context–Intrinsic combinations. Even long-tail minority categories retain sufficient sample volumes to support robust feature learning in the BHENet model. **c,** Spatial distribution of sample density and volume across 372 Chinese cities. The choropleth base map displays the log-transformed sample density (samples per km²), demonstrating continuous coverage across all provinces despite the east-west population gradient. Overlaid proportional circles (orange bubbles) represent the absolute sample volume per city, highlighting the major data contributions from high-density metropolitan clusters. The dataset comprises approximately 30 million labeled buildings, ensuring typological saturation across diverse urban morphologies.



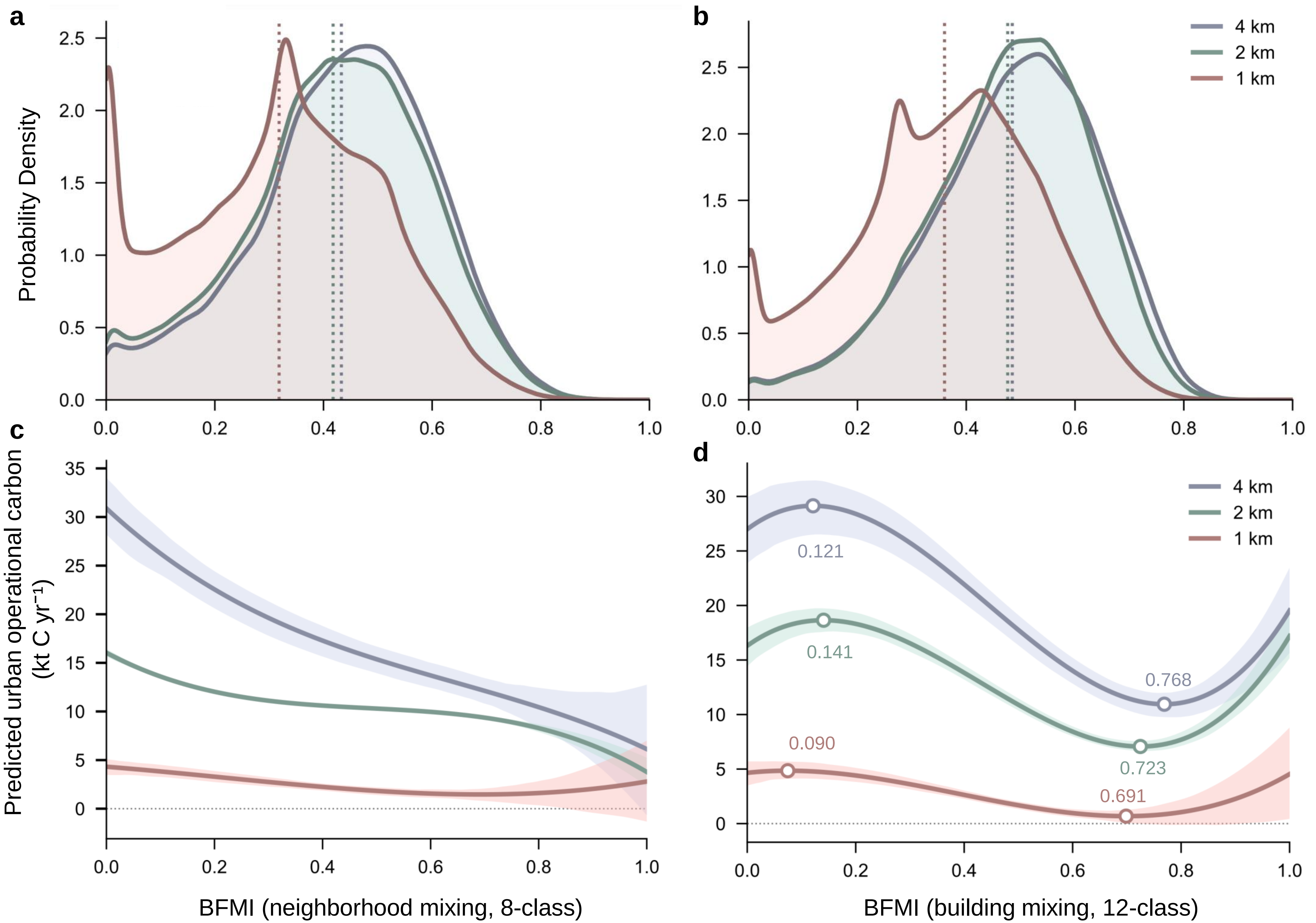
Supplementary Fig. S6 | Detailed architecture of the multimodal feature fusion module.
The fusion module integrates spatially aligned Visual Features and Textual Features with Auxiliary Data (vector context) through a hierarchical attention mechanism. Position Embedding Fusion (PE) injects relative spatial encodings to align the coordinate frames of image and text inputs. The core Vision-Language Feature Fusion block utilizes cross-attention, where visual features serve as the Query Mapping (QM) and textual semantics serve as the Key Mapping (KM) and Value Mapping (VM), dynamically weighting semantic cues based on visual relevance. Auxiliary Data is transformed via Data Embedding (DE) and processed through parallel Auxiliary Decision Layers (1 to *n*), which independently optimize contributions from distinct Decision Support Data (DSD) sources (e.g., road networks, footprints). The aggregated representation is refined by a Self-Attention Module (SA) to model global dependencies before the final classification Output.



Supplementary Fig. S7 | Sensitivity of the building-scale mixing index (BFMI) to volumetric weighting.
a, Scatter plot comparing BFMI calculated using volume weights (y-axis) versus footprint-area weights (x-axis) across 102,838 grid cells nationwide. The exceptionally strong linear correlation (Pearson's r = 0.993) suggests that area-weighted BFMI serves as a viable proxy when height data is unavailable. However, systematic deviations (data points colored by mean building height) reveal structural biases masked by global correlation. **b,** Mean scale divergence (ΔBFMI = Volume BFMI − Area BFMI) as a function of grid-level building height. The downward gradient quantifies these biases: volume weighting corrects mixing scores upward in low-rise zones (positive Δ) and downward in high-rise zones (negative Δ). Shaded bands represent 95% confidence intervals. **c, d,** 3D spatial visualization of scale divergence in representative Tier 2 cities, Xi'an (c) and Hefei (d), confirming the patterns identified in b. Grid elevation represents mean building height, and colors indicate ΔBFMI. Bright yellow/orange zones (low-rise) highlight areas where large-footprint structures (e.g., factories) mask multi-story amenities, while dark purple zones (high-rise) identify vertically specialized cores. These results demonstrate that volume weighting provides a more physically rigorous measure of functional intensity, reducing uncertainty in capturing vertical urbanism compared to area-based metrics.



Supplementary Fig. S8 | Spatial structure and systematic divergence between neighborhood-scale and building-scale functional mixing. **a,** Spatial distribution of scale divergence ($\Delta BFMI_g = BFMI_g^B - BFMI_g^N$) in representative cities (Beijing, Shanghai, Wuhan, Foshan). Red zones ($\Delta > 0$) indicate building-level diversity masked by neighborhood homogeneity; blue zones ($\Delta < 0$) indicate mono-functional buildings within mixed districts. **b, c,** Center–periphery profiles of mean divergence (**b**) and the proportion of grids with negative divergence (**c**), stratified by city tier. Tier 1 cities maintain high positive divergence across all distances, whereas lower-tier cities exhibit a rapid suburban decay and increased polarization (rising negative proportion), reflecting coarser peripheral development. Shaded bands in **b** denote **95% confidence intervals**. **d,** Hexbin scatter plot (n = 102,838) showing the systematic positive deviation of building-scale mixing from the neighborhood baseline (Mean Δ = 0.060, Pearson's r = 0.842), confirming the capture of fine-grained functional complexity.



Supplementary Fig. S9 | Sensitivity analysis of functional mixing distributions and decarbonization thresholds across taxonomic and spatial scales.

a, b, Kernel density estimation (KDE) of the normalized Building Function Mixed Index (BFMI) across three spatial resolutions (1 km, 2 km, and 4 km grid cells). Distributions are calculated using the 8-class neighborhood context (**a**) and the 12-class analytical aggregation (**b**). The systematic leftward shift and lower mean values observed at the 1 km resolution (red curves) reflect the "fragmentation effect" of the Modifiable Areal Unit Problem (MAUP), where finer grids artificially segment functional clusters into homogeneous units. Conversely, the 2 km (green) and 4 km (blue) distributions show high consistency, validating the 2 km grid as a robust analytical baseline. **c, d,** Fitted BFMI–urban operational carbon response curves across spatial scales, estimated via Spatial Durbin Models (SDM). **c,** Under the 8-class neighborhood context, mixing is associated with monotonic reductions in urban operational carbon across all scales, with a minor rebound at extreme values (>0.8) in the 1 km model due to data sparsity in the high-entropy tail. **d,** Under the 12-class analytical aggregation, the tri-phasic "decarbonization window" is structurally robust across resolutions. Shaded bands indicate **95% confidence intervals**. Colored dots mark the statistically identified turning points for each scale, with adjacent text labels denoting the specific BFMI values. Despite the expected smoothing effect at coarser resolutions (shifting the 4 km threshold slightly rightward), the relative position of the window remains stable, confirming that the capacity constraint is a robust structural feature of the urban functional system.

Supplementary Table S1 | Hierarchical functional taxonomy, aggregation status, and summary statistics.

Neighborhood context (8 classes)	Context–intrinsic combination (29 classes)	Analytical aggregation (12 classes)	Volume (10^4 m^3)	Share of Total (%)	Aggregation Status
Residential	Residential–Residential	Residential–Residential	14,205,074.7	30.92	Retained
	Residential–Commercial	Residential–Commercial	8,661,818.1	18.86	Retained
	Residential–Industrial	Other	27,129.9	0.06	Merged
	Residential–Medical	Other	7,486.1	0.02	Merged
	Residential–Transport	Other	3,297.8	<0.01	Merged
	Residential–Administrative	Other	1,202.9	<0.01	Merged
Industrial	Industrial–Industrial	Industrial–Industrial	9,149,525.7	19.92	Retained
	Industrial–Commercial	Industrial–Commercial	6,109,897.4	13.30	Retained
	Industrial–Transport	Other	13,303.1	0.03	Merged
	Industrial–Administrative	Other	2,102.4	<0.01	Merged
Public Open Space	Public–Public	Public–Public	981,715.3	2.14	Retained
	Public–Commercial	Other	178,870.4	0.39	Merged
	Public–Residential	Public–Residential	38,109.6	0.08	Retained
	Public–Industrial	Other	8,561.7	0.02	Merged
	Public–Transport	Other	904.7	<0.01	Merged
Transport	Transport–Transport	Transport–Transport	656,477.7	1.43	Retained
	Transport–Commercial	Other	125,757.4	0.27	Merged
Education	Education–Education	Education–Education	2,705,755.0	5.89	Retained
	Education–Residential	Other	132,948.7	0.29	Merged
	Education–Commercial	Other	9,626.2	0.02	Merged
	Education–Public	Other	3,296.9	<0.01	Merged
	Education–Industrial	Other	1,465.8	<0.01	Merged
	Education–Transport	Other	241.1	<0.01	Merged
Commercial	Commercial–Commercial	Commercial–Commercial	2,242,401.0	4.88	Retained
	Commercial–Industrial	Other	13,997.9	0.03	Merged
	Commercial–Administrative	Other	9,470.8	0.02	Merged
	Commercial–Transport	Other	3,902.5	<0.01	Merged
Medical	Medical–Medical	Medical–Medical	393,121.8	0.86	Retained
Administrative	Administrative–Administrative	Administrative–Administrative	247,548.5	0.54	Retained
Total	29 Types	12 Types	45,935,011.1	100.00	--

Note:

Values represent the total building volume (10^4 m^3) and share of total volume (%) across all surveyed cities. **Contextual Class Definitions** refer to the dominant land-use function of the neighborhood containing the buildings: **Residential**, housing, apartments, and dormitories; **Industrial**, manufacturing plants, warehouses, and logistics parks; **Education**, schools, universities, and research institutes; **Commercial**, retail, business offices, hotels, and entertainment; **Public Open Space**, buildings located within parks, plazas, greenbelts, and scenic areas; **Transport**, stations, airports, and transport hubs; **Medical**, hospitals and clinics; **Administrative**, government and civic administration offices. **Aggregation Status: Retained**, category preserved as an independent class in the analytical schema; **Merged**, category aggregated into "Other". Values marked as <0.01% indicate a negligible share.

Supplementary Table S2 | Heuristic rules for intrinsic functional refinement and conflict resolution.

Mechanism	**Contextual Context (Dominant)**	**Trigger Condition (Secondary Signal)**	**Inferred Intrinsic Label**	**Rationale / Example**
A. AOI Conflict Resolution	**All Contexts**	**Intersecting AOI (Lower Priority)**	**Secondary AOI Class**	Resolves overlapping land-use boundaries.
(Example)	*Education*	*Intersecting AOI: Residential*	*Residential*	Faculty housing or dorms within campus.
	Industrial	*Intersecting AOI: Commercial*	*Commercial*	Business parks within industrial zones.
B. Semantic Refinement	Education	Tag: *dormitory, apartments*	Residential	Explicit geometry tag.
		POI: *Catering*, *Shopping*, *Store*	Commercial	Canteens and campus retail.
		Default (No trigger)	*Education*	Academic buildings.
	Industrial	POI: *Dining*, *Life Services*, *Hotel*	Commercial	Worker support services.
		POI: *Logistics*, *Warehouse*	Transport	Logistics hubs.
		Default	*Industrial*	Factories/Workshops.
	Residential	POI: *Retail*, *Dining*, *Bank*	Commercial	Ground-floor commercial podiums.
		POI: *Kindergarten*, *School*	Education	Embedded community education.
		Default	*Residential*	Pure housing units.
	Public Open Space	POI: *Retail*, *Ticket Office*	Commercial	Tourist services.
		Default	*Public Open Space*	Scenic/Civic structures.
	Transport	POI: *Retail*, *Catering*	Commercial	Station/Airport commerce.
		Default	*Transport*	Terminals/Hubs.
	Commercial	Tag: *apartments, residential*	Residential	Mixed-use apartments.
		Default	*Commercial*	Offices/Malls.

Note:

This table outlines the hierarchical logic for determining the Intrinsic (Building) label. Mechanism A (AOI Conflict): Applied when a building footprint intersects multiple AOI polygons. Contextual assignment follows an Institutional Encapsulation Priority (Education > Public Open Space > Residential > Commercial > Industrial > Transport > Administrative > Medical). The intersecting

AOI category with higher priority defines the Context (neighborhood); the category with lower priority is assigned as the Intrinsic label, capturing nested functional zones (e.g., a dormitory in overlapping Education and Residential AOIs becomes Education–Residential). Mechanism B (Semantic Refinement): Applied to buildings within a single context. Intrinsic labels are refined using explicit building tags (highest precedence) or high-confidence POI spatial associations. If no trigger is present, the intrinsic label defaults to the contextual category.

Supplementary Table S3 | Contribution of data modalities to classification accuracy via occlusion analysis.

Functional Category	Textual Contribution (%)	Visual Contribution (%)	Auxiliary Contribution (%)	Dominant Modality
Industrial	84.61	14.38	1.00	Textual
Transport	86.53	10.89	2.58	Textual
Residential	15.60	72.66	11.73	Visual
Education	82.62	17.26	0.11	Textual
Public Open Space	86.41	12.73	0.86	Textual
Medical	87.20	12.25	0.55	Textual
Commercial	88.07	11.82	0.12	Textual
Administrative	74.83	25.05	0.12	Textual
Overall	75.72	22.12	2.13	Textual

Note: Contribution scores are derived from the relative drop in classification accuracy when the specific input modality is masked (occluded). While text is the primary signal for most functions, visual imagery is the decisive feature for identifying residential structures.

Supplementary Table S4 | Ablation study on feature fusion strategies and model components.

Exp. ID	Vision-Language Fusion Strategy	Position Embedding	Auxiliary Data Fusion	Mean Accuracy (OA)	Improvement
1	Concatenation (Baseline)	×	×	0.9181	--
2	Concatenation	Concatenation	×	0.9142	-0.39%
3	Cross-Attention	×	×	0.9284	+1.03%
4	Cross-Attention	Concatenation	×	0.9248	+0.67%
5	Cross-Attention	Distance-Encoding	×	0.9332	+1.51%
6	Cross-Attention	Cross-Attention	×	0.9426	+2.45%
7	Cross-Attention	Cross-Attention	Cross-Attention	0.9585	+4.04%

Note: "Concatenation" refers to simple feature stacking; "Cross-Attention" refers to the proposed attention-based fusion mechanism. Exp. 7 represents the final BHENet architecture. The significant gain from Exp. 1 to Exp. 3 highlights the efficacy of attention mechanisms in aligning heterogeneous modalities, while the leap in Exp. 7 confirms the value of auxiliary vector data. These ablation results are based on the internal held-out supervision-anchor set and are used to compare model components; independent final-map validation is reported in **Extended Data Fig. 2**.

Supplementary Table S5 | Summary statistics of BFMI distributions across taxonomic schemas and spatial resolutions.

Grid Resolution	BFMI taxonomy	Count (*N*)	Mean	Median	Std. Dev.	IQR
4 km	Analytical aggregation (12 classes)	25,710	0.485	0.500	0.158	0.212
	Neighborhood context (8 classes)	25,710	0.433	0.448	0.167	0.219
2 km (Baseline)	Analytical aggregation (12 classes)	102,838	0.476	0.491	0.151	0.202
	Neighborhood context (8 classes)	102,838	0.418	0.430	0.169	0.224
1 km	Analytical aggregation (12 classes)	399,094	0.360	0.371	0.170	0.234
	Neighborhood context (8 classes)	399,094	0.318	0.330	0.186	0.280

Supplementary Table S6 | Robustness checks and sensitivity analysis across model specifications.

Model Specification	Description	Marginal Effect Shape	Decarbonization Window (BFMI)	EAD Total Effect (t C yr^{-1})	JSD Total Effect (t C yr^{-1})	Log-Likelihood
(1) Baseline	12-class, 2 km Grid, Legendre P_3, KNN-8	Tri-phasic	[0.141, 0.723]	6,095.2***	2,484.1***	-940,816.3
Alternative Weights						
(2) Queen Contiguity	Adjacency-based connectivity	Tri-phasic	[0.139, 0.718]	4,882.7***	2,216.2***	-919,154.7
(3) IDW Matrix	Inverse distance, 10 km cutoff	Tri-phasic	[0.177, 0.672]	10,012.9***	3,835.0***	-936,364.5
Functional Form						
(4) Legendre P_4	Tests for higher-order non-linearity	Tri-phasic	[0.163, 0.790]	5990.0***	2557.2***	-940,653.8
(5) Splines (3 knots)	Restricted cubic splines ($df = 3$)	Tri-phasic	[0.155, 0.718]	5,423.5***	2,452.5***	-940,152.5
(6) Splines (4 knots)	Restricted cubic splines ($df = 4$)	Tri-phasic	[0.165, 0.780]	5,677.5***	2,245.5***	-940,755.5
Spatial Scale						
(7) 4 km Grid	Coarser aggregation	Tri-phasic	[0.121, 0.768]	9,983.1***	11,689.6***	-515,136.2
(8) 1 km Grid	Finer aggregation	Tri-phasic	[0.090, 0.691]	2,608.2***	1,516.0***	-3,067,261.9
Taxonomy						
(9) Context–intrinsic combinations (29 classes)	Fine-grained combinations	Tri-phasic	[0.145, 0.679]	6,130.1***	2,537.0***	-940,785.5
(10) Neighborhood context (8 classes)	Contextual labels	**Monotonic**	N.A. (Monotonic decay)	6,807.3***	2,997.7***	-940,211.1
City Stratification						
(11) Tier 1 Cities	Mega-cities (>3,000 km^2)	Tri-phasic	[0.289, 0.804]	21,166.1***	13,209.0***	-102,941.8
(12) Tier 2 Cities	Large cities (800–3,000 km^2)	**Bi-phasic**	[0.000, 0.664]†	**-89.5 (ns)**	3,298.0***	-209,498.8
(13) Tier 3 Cities	Small/Medium (<800 km^2)	Tri-phasic	[0.145, 0.689]	5,671.5***	2,270.9***	-581,963.9

Note:

Summary of sensitivity analyses validating the structural robustness of the tri-phasic BFMI–urban operational carbon relationship. Model Specification: All models include the full set of controls and fixed effects as in the Baseline (**Extended Data Table 2**), unless otherwise specified. Decarbonization Window: Defines the BFMI range where the marginal effect on urban operational carbon is significantly negative. N.A. (Monotonic decay) indicates a continuous reduction in urban operational carbon without turning points. EAD/JSD Total Effect: coefficients represent the total marginal impact on urban operational carbon, reported in t C yr^{-1}; EAD is expressed in normalized 10-km-window units, so the +1-km interpretation used in the main text is obtained by dividing the EAD coefficient by 10. Significance: *** indicates $P < 0.01$; (ns) indicates not significant ($P > 0.1$). Specifically, the nonsignificant EAD effect in Tier 2 cities likely reflects high internal heterogeneity (e.g., polycentric manufacturing hubs vs. monocentric capitals) confounding the distance signal. †

Tier 2 cities exhibit a bi-phasic response (monotonic decrease from BFMI = 0 until the upper threshold at 0.664).

Supplementary Table S7 | Characteristics of the 31 international cities and comparison with Chinese city tiers.

Region	City	Country	Class[a]	Pop. Density[b] (per km^2)	Mean BFMI[c] (12-class)	Mean Volume[d] (10^4 m^3)	Mean urban operational carbon[e] (t C yr^{-1})
N. America	New York	USA	L	2,196	0.53	452.7	26,885
	Mexico City	Mexico	L	2,820	0.56	395.2	15,323
	Los Angeles	USA	L	785	0.30	145.3	12,365
	Chicago	USA	L	461	0.45	94.5	19,185
	Dallas–Fort Worth	USA	L	239	0.39	41.2	6,727
	Toronto	Canada	L	392	0.37	79.1	7,671
	Phoenix	USA	L	460	0.44	67.0	7,925
	Montréal	Canada	L	666	0.49	107.2	16,621
	Seattle	USA	M	313	0.45	51.1	3,964
	Denver	USA	M	357	0.38	50.1	10,674
	Vancouver	Canada	M	395	0.39	62.3	4,712
	Raleigh	USA	S	171	0.41	15.2	3,374
S. America	São Paulo	Brazil	L	2,066	0.63	272.3	9,021
	Quito	Ecuador	M	968	0.56	75.0	3,986
	Florianópolis	Brazil	S	479	0.54	35.4	2,828
Europe	London	UK	L	337	0.51	46.9	3,614
	Paris	France	L	1,448	0.64	389.5	15,477
	Madrid	Spain	L	402	0.40	59.0	3,710
	Berlin	Germany	L	160	0.45	28.8	5,653
	Rome	Italy	L	256	0.51	28.6	4,574
	Warsaw	Poland	M	201	0.49	27.1	7,905
	Rotterdam–Hague	Netherlands	M	369	0.41	121.8	22,338
	Hamburg	Germany	M	151	0.49	30.1	2,620
	Valencia	Spain	M	238	0.46	42.3	4,194
	Lyon	France	M	193	0.50	40.6	3,712
	Zürich	Switzerland	S	276	0.53	48.3	4,186
Africa	Lagos	Nigeria	L	1,498	0.48	125.4	4,170
	Cairo	Egypt	L	1,194	0.48	39.7	12,343
Oceania	Sydney	Australia	L	297	0.51	100.1	6,990
	Brisbane	Australia	M	178	0.36	53.1	8,124
	Canberra	Australia	S	187	0.30	30.4	6,780
China (Ref.)	China Tier 1	--	L	2,218	0.51	703.0	39,606
	China Tier 2	--	M	1,724	0.50	507.3	25,170
	China Tier 3	--	S	1,195	0.46	374.9	16,077

Note:

Sample: 31 international cities selected to represent diverse urban morphologies across 5 continents, stratified by metropolitan scale (Class: L = Large, M = Medium, S = Small).

Metric Definitions:

[a]Class: International classification based on metropolitan population; Chinese classification based on built-up area (Tier 1 > 3,000 km^2, Tier 2 800–3,000 km^2, Tier 3 < 800 km^2).

[b]Pop. Density: Average residential population density per grid cell (people km^{-2}).

[c]Mean BFMI: Grid-level average of the 12-class volume-weighted Building Function Mixed Index.

[d]Mean Volume: Average total building volume per 2 km grid cell (10^4 m^3). Note the structural difference: Chinese Tier 1 cities (703.0) exhibit functional volumes an order of magnitude higher than most international large cities (e.g., London: 46.9, New York: 452.7), underpinning the density-dependent boundary conditions discussed in the main text.

[e]Mean urban operational carbon: average annual ODIAC-derived territorial land-based carbon emissions per 2 km grid cell, reported in carbon-mass units (t C yr^{-1}). CO_2-equivalent values can be obtained using the molecular-weight ratio 44/12.

Data Sources: Functional and morphological metrics are derived from the unified functional atlas constructed in this study, whereas demographic and carbon variables are derived from the external datasets listed in Supplementary Table S8 and aggregated to the same 2 km grid system. For specific data sources (e.g., population grids, building heights) and processing details, please refer to the Data Dictionary in **Supplementary Table S8**.

Supplementary Table S8 | Data sources, access routes, and preprocessing

Dataset / variable group	Source (provider / publication)	Version / vintage	License / restrictions	Access (route)	Preprocessing (summary)	Spatial resolution / unit	Temporal coverage	Citation
Urban operational carbon outcome (China and global 31 cities)	ODIAC Fossil Fuel CO_2 Emissions Dataset (ODIAC2024), National Institute for Environmental Studies (NIES), Japan	ODIAC2024 monthly 1 km GeoTIFF land product; 2020 monthly layers used (ODIAC2024 release covers 2001–2023)	ODIAC terms; attribution required	ODIAC portal (U1); NIES DOI page (U2)	The 12 monthly layers for 2020 were summed to annual carbon mass and aggregated to the 2 km × 2 km analysis grid within the urban built-up mask using area-weighted zonal summation. No bilinear resampling was used for emission totals. Aggregate CO_2-equivalent values, when reported, were obtained from carbon-mass estimates using the molecular-weight ratio 44/12.	Native: 1 km GeoTIFF land rasters, tonnes of carbon per cell per month. Used: t C yr^{-1} per 2-km grid cell; CO_2 equivalents reported only for aggregate interpretation.	ODIAC2024 product coverage: monthly series for 2001–2023; analysis uses the 2020 monthly layers.	Oda, T., Maksyutov, S. & Andres, R. J. The Open-source Data Inventory for Anthropogenic CO_2, version 2016 (ODIAC2016): a global monthly fossil fuel CO_2 gridded emissions data product for tracer transport simulations and surface flux inversions. Earth Syst. Sci. Data 10, 87–107 (2018). Oda, T. & Maksyutov, S. ODIAC Fossil Fuel CO_2 Emissions Dataset (ODIAC2024) (Center for Global Environmental Research, National Institute for Environmental Studies, 2025); https://doi.org/10.17595/20170411.001.
OSM AOIs (Contextual labels; China and global 31 cities)	OpenStreetMap (OSM) contributors (planet dump / extracts)	OSM planet snapshot (2020-01-01)	ODbL v1.0 (attribution + share-alike for derived DB)	OSM planet dump (U3); attribution (U4)	Tag normalization; AOI conflict resolution by priority; building-AOI overlap; topology checks	Native: vector polygons. Used: building-footprint overlap features; aggregated to 2-km grids when needed	Native: continuously updated. Used: 2020-01-01 snapshot	OpenStreetMap contributors. Planet dump 2020-01-01. OpenStreetMap https://planet.openstreetmap.org (2020).
OSM roads & land use (China and global 31 cities)	OpenStreetMap (roads; landuse/parcels)	OSM planet snapshot (2020-01-01)	ODbL v1.0 (attribution + share-alike for derived DB)	OSM planet dump (U3); attribution (U4)	Simplification/snapping; topology repair; road-length / land-use fractions per 2-km cell	Native: vector lines/polygons. Used: aggregated to 2-km grids	Native: continuously updated. Used: 2020-01-01 snapshot	OpenStreetMap contributors. Planet dump 2020-01-01. OpenStreetMap https://planet.openstreetmap.org (2020).
POIs (text semantics; China)	Gaode/AMap POI database via Open Platform APIs	API query window (2020)	Provider terms; redistribution restricted; attribution as required by provider	AMap Open Platform (U5)	Deduplication; category mapping; spatial filtering around buildings; keep nearest POIs per category (model design)	Native: point POIs. Used: building-level semantic features	Native: live database. Used: 2020 query window	Gaode Map (AMap). AMap Web Service API documentation. https://lbs.amap.com (2020).
POIs (text semantics; Global 31 cities)	Overture Maps Foundation (Places theme)	Study tag: 2023-12; schema: JSON Schema draft 2023-12; Overture public data releases begin 2023 (see release notes)	Composite licensing by upstream sources; attribution required per Overture attribution guidance	Docs (U6); release notes (U7)	Deduplication; category mapping; spatial join around buildings; keep nearest POIs per category	Native: point POIs. Used: building-level semantic features	Release-based snapshot (not historical). Used as static covariate aligned to analysis window	Overture Maps Foundation. Places theme documentation and releases. https://docs.overturemaps.org/blog/ (2023).
Building footprints (China)	China's sub-meter building footprints dataset (Zenodo; paper describes product)	Zenodo records (footprints vintage ~2020; released 2024)	CC BY 4.0	Zenodo records (U8, U9)	Geometry cleaning; reprojection; join to height; grid aggregation	Native: footprint polygons (sub-meter mapping). Used: building unit;	Vintage aligned to ~2020	Huang, X., Zhang, Z. & Li, J. China's first sub-meter building footprints derived by

Dataset / variable group	Source (provider / publication)	Version / vintage	License / restrictions	Access (route)	Preprocessing (summary)	Spatial resolution / unit	Temporal coverage	Citation
						aggregated to 2-km grids		deep learning. *Remote Sens. Environ.* **311**, 114274 (2024).
Building height (China)	CNBH-10 m building height product (Zenodo; paper describes product)	CNBH-10 m (10 m raster; reference period ~2020)	CC BY 4.0	Zenodo record (U10)	Reprojection; footprint-height join; building volume = area x height	Native: 10 m raster. Used: building-level height via join; aggregated to 2-km grids	Reference period ~2020	Wu, W. B. et al. A first Chinese building height estimate at 10 m resolution (CNBH-10 m) using multi-source earth observations and machine learning. *Remote Sens. Environ.* **291**, 113578 (2023).
Building height (Global 31 cities)	Global urban built-up height grid (Figshare; paper describes product)	Figshare release (2022-11-01); built-up heights ~2015	CC BY 4.0	Figshare (U11)	Reprojection; resampling; aggregation to 2-km grids	Native: 500 m grid (mean built-up height). Used: aggregated to 2-km grids	Coverage: circa 2015. Used as static height covariate for matched window	Zhou, Y. et al. Satellite mapping of urban built-up heights reveals extreme infrastructure gaps and inequalities in the Global South. *Proc. Natl Acad. Sci. USA* **119**, e2214813119 (2022).
Aerial imagery (visual features; China and global 31 cities)	Google Earth imagery	Tile-specific vintages (~2019-2021; nearest to 2020 where feasible)	Google Earth end-user terms; imagery not redistributable; attribution required	Google Earth (U12); terms (U13)	Patch extraction centered at building; normalization to fixed input size for encoder	~0.5 m GSD (used for feature extraction; varies by location)	Multi-year; tiles near 2020 where feasible	Google. Google Earth. https://earth.google.com (2020).
Administrative boundaries (China)	Standard Map Service System, Ministry of Natural Resources of the People's Republic of China	Standard boundary set; review number GS(2024)0650	Use subject to Standard Map Service terms; include review number GS(2024)0650 on maps; base map not modified	Standard Map Service (U14)	Reprojection; spatial join to city IDs	Vector polygons	Static (used in 2020 analysis window; accessed 2024)	Ministry of Natural Resources of the People's Republic of China. Standard Map Service System (review no. GS(2024)0650). https://www.webmap.cn (2024).
Night-time lights (covariate; China and global 31 cities)	Earth Observation Group (EOG), Payne Institute (VIIRS DNB Annual V2.1; via Google Earth Engine catalog)	V2.1 annual composites (2020 used)	CC BY 4.0 (EOG products); attribution required	GEE catalog (U15)	Reprojection; resampling; aggregation to 2-km grids	Native: 15 arc-second (~500 m) grid. Used: aggregated to 2-km grids	Annual series; used 2020 (or nearest)	Elvidge, C. D. et al. Annual time series of global VIIRS nighttime lights derived from monthly averages: 2012 to 2019. *Remote Sens.* **13**, 922 (2021).
Population density (covariate; China)	China-wide 100 m gridded population dataset (Figshare; paper describes product)	v1 (100 m; 2020 used)	CC BY 4.0	Figshare DOI (U16)	Reprojection; aggregation to 2-km grids	Native: 100 m grid. Used: aggregated to 2-km grids	2020 (or nearest year)	Chen, Y. et al. A 100 m gridded population dataset of China's seventh census using ensemble learning and big geospatial data. *Earth Syst. Sci. Data* **16**, 3705-3718 (2024).

Dataset / variable group	Source (provider / publication)	Version / vintage	License / restrictions	Access (route)	Preprocessing (summary)	Spatial resolution / unit	Temporal coverage	Citation
Population density (covariate; Global 31 cities)	WorldPop Project (global gridded population)	WorldPop population counts/density (2020; 100 m preferred)	CC BY 4.0	WorldPop (U17) or GEE catalog (U18)	Reprojection; aggregation to 2-km grids	Native: ~100 m (3 arc-second) grid. Used: aggregated to 2-km grids	2020 (or nearest year)	Lloyd, C. T., Sorichetta, A. & Tatem, A. J. High resolution global gridded data for use in population studies. *Sci. Data* **4**, 170001 (2017).
Terrain controls (elevation, slope, relief; China and global 31 cities)	Copernicus DEM GLO-30 (Copernicus/ESA; distributed via Copernicus Data Space / GEE)	GLO-30 (public release; static)	Copernicus DEM license; attribution required	Copernicus Data Space (U19) or GEE catalog (U20)	Derive slope/relief; aggregate to 2-km grids	Native: 30 m DEM. Used: aggregated to 2-km grids	Static	Copernicus Programme. Copernicus DEM GLO-30. Copernicus Data Space Ecosystem https://dataspace.copernicus.eu/explore-data/data-collections/copernicus-contributing-missions/collections-description/COP-DEM (2021).
Global cities list (31 cities)	This study (city selection for global comparison)	N/A	N/A	Included in Supplementary and public repository upon publication (DOI to be provided)	City metadata harmonization (continent/size tiers)	City IDs / polygons	Reference year aligned to emissions window	This study.
Building footprints (Global; 31 cities)	This study (harmonized footprints; respecting third-party limits)	v1.0	To be released (respecting third-party limits)	Public repository upon publication (DOI to be provided)	Harmonize footprints; optional height join; compute volume; aggregate to 2-km grids	Vector + optional height; aggregated to 2-km grids	Matched to emissions window	This study.
Derived: China building-function atlas (contextual & intrinsic labels)	This study	v1.0	To be released (respecting third-party limits)	Public repository upon publication (DOI to be provided)	Nationwide inference outputs; QC; optional uncertainty flags	Building-level vectors/IDs	2020 reference	This study.
Derived: BFMI grids + ring summaries	This study	v1.0	To be released	Public repository upon publication (DOI to be provided)	Volume aggregation; Shannon entropy; normalization; ring aggregation	2 km grids; ring panels	2020 reference	This study.
Code (preprocess + BHENet + econometrics)	This study	v1.0	Open-source license (to be specified)	Public repository upon publication	End-to-end scripts, configs, and analysis notebooks	N/A	N/A	This study.

Notes: (i) URL short codes (U1, U2, …) are expanded in the URL key below. (ii) For Standard Map Service products, include the review number GS(2024)0650 in map figure captions as required.

URL key (expanded)

U1: https://odiac.org

U2: https://www.nies.go.jp/doi/10.17595/20170411.001-e.html

U3: https://planet.openstreetmap.org

U4: https://www.openstreetmap.org/copyright

U5: https://lbs.amap.com

U6: https://docs.overturemaps.org

U7: https://docs.overturemaps.org/blog/

U8: https://zenodo.org/records/10473278

U9: https://zenodo.org/records/10475803

U10: https://zenodo.org/records/7064268

U11: https://figshare.com/s/7f2b254ed18fac8eb7a0

U12: https://earth.google.com

U13: https://maps.google.com/help/terms_maps-earth/

U14: https://www.webmap.cn

U15: https://developers.google.com/earth-engine/datasets/catalog/NOAA_VIIRS_DNB_ANNUAL_V21

U16: https://doi.org/10.6084/m9.figshare.24916140.v1

U17: https://www.worldpop.org

U18: https://developers.google.com/earth-engine/datasets/catalog/WorldPop_GP_100m_pop

U19: https://dataspace.copernicus.eu/explore-data/data-collections/copernicus-contributing-missions/collections-description/COP-DEM

U20: https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_DEM_GLO30