Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Capitalization of urban visual morphology: a world-model perspective from the housing market of the Taiwan region

**Juliana Pin-Jen Tseng**
The University of Hong Kong
**Kuo-Kun Tseng**
kktseng@hit.edu.cn
Harbin Institute of Technology, Shenzhen
**Regina Fang-Ying Lin**
Shenzhen Technology University

**Research Article**

**Keywords:** urban visual morphology, hedonic prices, housing market, satellite imagery, world model, Taiwan region, spatial inequality

**Posted Date:** September 9th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-10729214/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Capitalization of urban visual morphology: a world-model perspective from the housing market of the Taiwan region

Juliana Pin-Jen Tseng[1,2], Kuo-Kun Tseng[1*], Regina Fang-Ying Lin[3]

[1*]School of Computer Science and Technology, School of Future Design, Harbin Institute of Technology, Shenzhen, Shenzhen, China.
[2]The University of Hong Kong, Hong Kong, China.
[3]Business School, Shenzhen Technology University, Shenzhen, China.

*Corresponding author(s). E-mail(s): kktseng@hit.edu.cn;
Contributing authors: tseng.juliana@hotmail.com;
linfangying@sztu.edu.cn;

**Abstract**

Whether the measurable morphology of urban visual environments is reflected in housing prices, and how this capitalization varies across time and income, remains an open empirical question, particularly outside Western cities. We study this question with roughly 544,000 geocoded transactions from the Taiwan region's actual-registration records (2012–2026), matched to time-varying Sentinel-2 imagery. In township-by-year fixed-effects hedonic models with township-clustered standard errors, we estimate implicit prices for eight visual-morphology features measured within 500 m neighborhoods. Built-texture complexity commands a sizeable positive premium, whereas the local shares of green, water, bright and dark pixels carry significant discounts. These valuations shifted systematically over the past decade and trace a pronounced income gradient: bright built form is valued mainly by low-income communities, while high-price neighborhoods pay more for brightness and discount complex texture more heavily, so that visual capitalization bears directly on spatial inequality. The results survive scale, urban-heterogeneity and accessibility checks and are corroborated by deep features of the Prithvi-EO-2.0 geospatial foundation model. Treating each neighborhood as a small dynamical system, a world-model analysis shows that temporal visual trajectories carry nearly the same incremental information as the full deep-embedding set, that prices capitalize the visible present

rather than the anticipated future, and that feature-space counterfactual interventions imply economically meaningful revaluations (about NT$350 billion in aggregate).

**Keywords:** urban visual morphology, hedonic prices, housing market, satellite imagery, world model, Taiwan region, spatial inequality

# Introduction

City dwellers experience their neighborhoods partly through what they see: the layout of streets, buildings, vegetation and water as they appear from above. Economists have long suspected that this visual environment shapes how residents value a place. The hedonic price framework turns that intuition into a testable statement by decomposing the price of a dwelling into the implicit prices of its attributes, including locational and environmental amenities [1–3]. Advances in computer vision and satellite imagery now make it possible to measure the visual morphology of urban neighborhoods at scale, and a growing body of work asks how the "look" of a place is priced by markets [4, 5]. Landscape metrics of urban green space, for instance, enter hedonic housing-price models with economically meaningful coefficients [6]; spatially and temporally explicit hedonic specifications are increasingly recommended to avoid biased implicit-price estimates [7].

Beyond green space, street-level imagery has been used to capture the perceived quality, walkability and aesthetic character of neighborhoods and to link these visual attributes to property values [8–11]. Yet much of this work concentrates on a single city or a small set of visual cues, rarely spans more than one time period, and leans on hand-coded or crowd-sourced perceptual scores that are hard to reproduce. Satellite imagery offers an alternative: its coverage of urban form is complete, consistent and time-varying, and economists have increasingly drawn on it to measure local economic activity, growth and welfare from space [12–14]. The same data make it possible to compute morphological descriptors, such as land-cover shares, texture and brightness statistics, for every neighborhood of a national housing market over more than a decade. Whether such reproducible, satellite-derived features are systematically priced, and how their prices have evolved, remains an open empirical question.

Three questions serve as the starting point. Which visual-morphology features are actually capitalized into housing prices once unobserved neighborhood and time heterogeneity is absorbed? How do these capitalized valuations change as markets, preferences and urban form shift? And who captures the benefits: are visual premiums systematically concentrated in high-income or high-price communities in ways that could amplify spatial inequality?

We use the complete real-estate transaction records of the Taiwan region's "actual registration" (buyer–seller) system from 2012 to 2026, matched to time-varying Sentinel-2 imagery and township-level income statistics. The Taiwan region's housing market is dense and mature, with a high volume of transactions; the microdata are complete down to the individual transaction; and the metropolitan areas span a broad

mix of neighborhood morphologies, from old, dense urban cores to newer, low-density suburbs. Our empirical strategy estimates hedonic price models on approximately 544,000 residential transactions, absorbing township-by-year fixed effects, clustering standard errors at the township level, and identifying implicit prices from morphological variation within townships. Because our handcrafted measures could in principle miss visual information that markets price, we also validate them against deep visual features from a large geospatial foundation model (Prithvi-EO-2.0).

We first describe how the eight visual-morphology features are measured from satellite imagery, then identify which of them are capitalized into prices in the baseline hedonic model and trace how those implicit prices have changed over time. We next examine the income heterogeneity of visual premiums, their spatial association with price and income, and their robustness to scale, metropolitan and accessibility specifications. A world-model analysis then estimates per-neighborhood visual trajectories, tests whether prices capitalize anticipated future visual change, and quantifies the incremental information in temporal dynamics. The Discussion draws out the economic interpretation, limitations, counterfactual policy simulations and implications; the Methods describe the data, sample construction, feature-extraction algorithms and the hedonic, validation and world-model specifications.

# Results

## Measuring neighborhood visual morphology

For each transaction, we crop a **500 m-radius neighborhood** around its 100 m grid cell from the satellite image of the corresponding year, and compute eight visual-morphology features:

- Green ratio (`green_ratio`), water ratio (`water_ratio`), bright pixel ratio (`bright_ratio`), and dark pixel ratio (`dark_ratio`);
- Edge density (`edge_density`, a proxy for built-texture complexity) and texture entropy (`texture_entropy`, the entropy of the luminance histogram);
- Mean brightness (`mean_brightness`) and brightness standard deviation (`brightness_std`, a proxy for visual disorder).

Quality control removes neighborhoods with cloud cover exceeding 30%. All features are sample-centered so that interaction terms are easy to interpret. For the temporal analyses (the M3 year-trend model and the early/recent comparison in M5), features are additionally standardized to zero mean and unit variance *within each year* to purge year-to-year measurement drift in the annual composites (e.g., sensor and scene differences); without this step, cross-period coefficient contrasts could reflect changes in the measured feature distributions rather than changes in valuation. Cross-sectional estimates (M2, M4, M6) retain raw units.

## Which visual features are capitalized into housing prices

Table 1 summarizes the model specifications. Relative to a baseline model (M1) with building attributes only, the main model (M2) adds the eight morphological features,

and M3 further adds their linear interactions with time (year − 2019) to trace trends. All models absorb township-by-year fixed effects and cluster standard errors at the township level; the dependent variable is the log price per unit area.

**Table 1** Model specifications and fit statistics

| | M1 Baseline | M2 Main | M3 Trend | M_all (Prithvi) |
|---|---|---|---|---|
| Building attributes | Yes | Yes | Yes | Yes |
| 8 handcrafted morphological features | No | Yes | Yes | Yes |
| Morphology × (year − 2019) | No | No | Yes | No |
| Prithvi spectral + embedding features | No | No | No | Yes |
| $R^2$ | 0.7698 | 0.7755 | 0.7772 | 0.7807 |
| $N$ | 544,232 | 544,232 | 544,232 | 492,691 |

[a] M_all is estimated on the Prithvi-matched subsample; see Table 2 of the main text.

Table 2 reports the morphological main effects from M2. Controlling for building attributes and township-by-year fixed effects, **edge density (neighborhood texture complexity) carries a significant positive premium**, while **higher local shares of bright, dark, green and water pixels are significantly discounted**. The negative associations on green and water ratios at the local (500 m) neighborhood scale reflect their spatial co-variation with older, lower-density neighborhoods within urban cores; their positive amenity value would need to be identified at more aggregated scales or through interactions. Mean brightness, brightness standard deviation (a visual-disorder proxy) and texture entropy are not significant at the 5% level. Figure 1 plots the M2 coefficient forest.

**Table 2** Visual-morphology main effects on log unit price (M2), township-clustered standard errors

| Feature | Coef. | SE | $p$ |
|---|---|---|---|
| Edge density | +0.324 | 0.060 | $< 0.001$ |
| Green ratio | −0.196 | 0.037 | $< 0.001$ |
| Water ratio | −0.146 | 0.055 | 0.009 |
| Bright pixel ratio | −0.141 | 0.046 | 0.002 |
| Dark pixel ratio | −0.129 | 0.063 | 0.040 |
| Texture entropy, brightness std, mean brightness | n.s. | | |

## Temporal evolution of visual premiums

### Continuous trends (M3)

Table 3 reports the morphology–time interactions from M3, estimated on features standardized within each year (coefficients in within-year-SD units per year). The

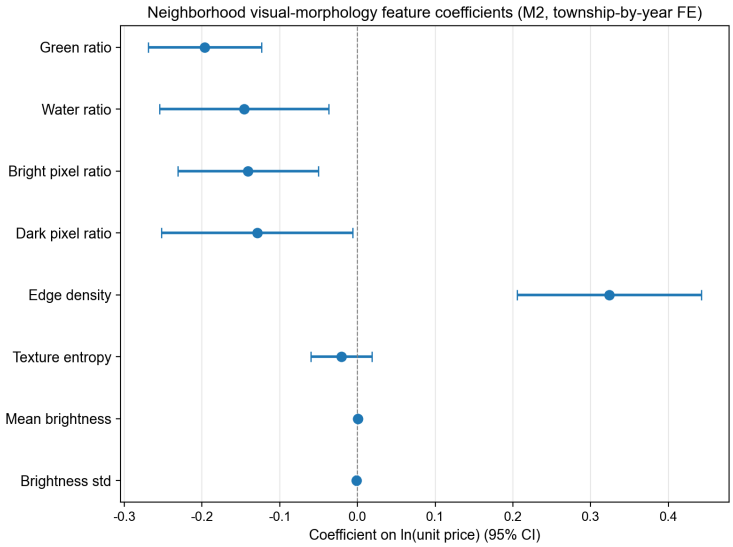



**Fig. 1** Estimated coefficients of the eight visual-morphology features on log unit housing price (M2), with 95% confidence intervals based on township-clustered standard errors.

negative premia on green, water and dark built form all deepen over time ($-0.0050$, $-0.0027$ and $-0.0070$ per year), the premium on bright built form (bright pixel share) rises ($+0.0066$, $p = 0.002$), and the mean-brightness premium declines ($-0.0132$, $p = 0.001$), while the visual-disorder discount weakens ($+0.0041$, $p = 0.027$). The edge-density and texture-entropy trends are insignificant.

**Table 3** Morphology × year-trend coefficients (M3; features standardized within year, coefficients in within-year-SD units per year)

| Morphology × trend | Coef. | $p$ |
|---|---|---|
| Green ratio | $-0.0050$ | 0.003 |
| Water ratio | $-0.0027$ | $< 0.001$ |
| Bright pixel ratio | $+0.0066$ | 0.002 |
| Dark pixel ratio | $-0.0070$ | 0.001 |
| Mean brightness | $-0.0132$ | 0.001 |
| Brightness std | $+0.0041$ | 0.027 |
| Edge density | n.s. | |
| Texture entropy | n.s. | |

## Early ($\leq$2018) versus recent ($\geq$2019) subsamples (M5)

We also split the sample at 2018/2019, contrasting an early period ($\leq$2018; $R^2 = 0.7660$, $N = 217{,}220$) with a recent period ($\geq$2019; $R^2 = 0.7784$, $N = 327{,}012$),

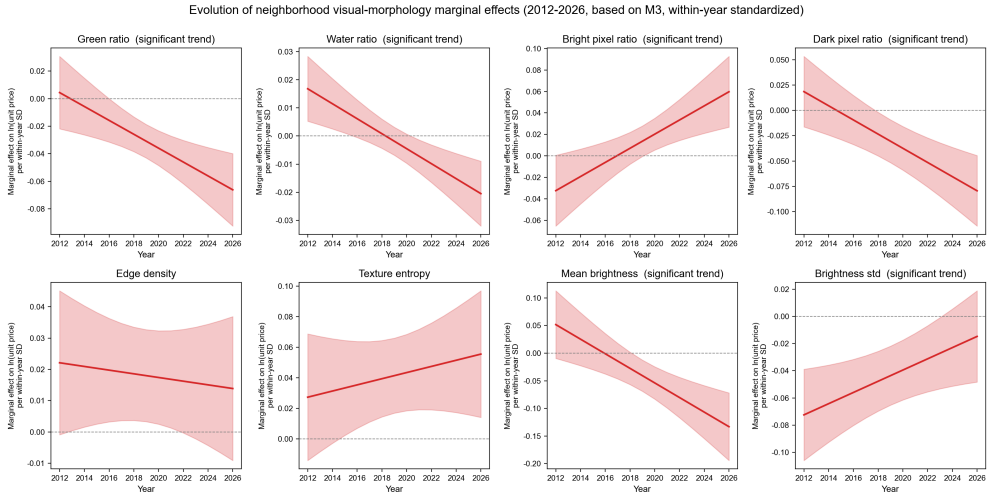



**Fig. 2** Evolution of the marginal effects of visual-morphology features on log unit housing price over time, in within-year-SD units (M3 trends and M5 early/recent comparison).

and pool the interaction of morphology with a "recent" indicator using within-year-standardized features (Table 4). The negative associations of green, water and dark built form deepen ($-0.042$, $-0.015$ and $-0.033$), the mean-brightness premium declines ($-0.065$), and the early edge-density premium fades ($-0.024$), while the bright-pixel-share premium rises ($+0.028$) and the texture-complexity premium edges up ($+0.030$, $p = 0.050$). Figure 2 visualizes the evolution of marginal effects.

**Table 4** Morphology × recent-period interactions (M5; features standardized within year, coefficients in within-year-SD units)

| Morphology × recent | Coef. | $p$ |
|---|---|---|
| Green ratio | $-0.042$ | $< 0.001$ |
| Water ratio | $-0.015$ | 0.003 |
| Bright pixel ratio | $+0.028$ | 0.009 |
| Dark pixel ratio | $-0.033$ | 0.048 |
| Edge density | $-0.024$ | 0.009 |
| Mean brightness | $-0.065$ | 0.008 |
| Texture entropy | $+0.030$ | 0.050 |
| Brightness std | $+0.011$ | 0.392 |

Over the decade, the market's valuation structure shifted in a consistent direction: discounts on green, water and dark built form deepened, the brightness premium declined, the early edge-density premium faded, and value moved toward compact

bright built form. What the visual environment is worth to buyers has therefore changed over time.

## Income heterogeneity of visual premiums

### Proxy variables

We use two complementary proxies for neighborhood income. The first (M4) is the township-level median unit-price tercile (low/middle/high). The second (M6) uses comprehensive income tax statistics from the Taiwan region's Ministry of Finance: township-level average income per household over 2012–2019, transaction-weighted into terciles. The two proxies correlate at $r = 0.708$ at the township level ($N = 284$), confirming that the price proxy is a valid approximation of income.

### Heterogeneity by price tercile (M4, reference = middle)

High-price (affluent) communities pay a premium for "brighter" environments, while imposing a larger discount on complex texture and on high bright-pixel shares. The brightness dimension, in short, is valued very differently from one neighborhood type to the next (Table 5).

**Table 5** Income/wealth heterogeneity by price tercile (M4)

| Feature × group | Low × | High × |
|---|---|---|
| Mean brightness | +0.000 ($p = 0.60$) | +0.0023 ($p = 0.000$) |
| Bright pixel ratio | −0.138 ($p = 0.13$) | −0.166 ($p = 0.008$) |
| Texture entropy | +0.025 ($p = 0.46$) | −0.065 ($p = 0.004$) |
| Other 5 features | n.s. | n.s. |

### Heterogeneity by comprehensive income (M6, $N = 544{,}545$, $R^2 = 0.7807$)

Subsample regressions: low-income $R^2 = 0.754$ ($N = 189{,}863$), middle $R^2 = 0.697$ ($N = 179{,}176$), high $R^2 = 0.762$ ($N = 175{,}506$).

The **bright pixel ratio (share of bright built form) premium is significantly concentrated in low-income communities** (+0.187 against a main effect of −0.194): in low-income, predominantly old and dense built-up areas, a brighter environment carries positive value, whereas in the full sample this feature is negatively priced. The reversal deserves emphasis: low-income communities put more weight on "brightness", plausibly because a brighter built form reads as a tangible sign of improvement. Water ratio, by contrast, commands a significant premium only in high-income townships. Fig. 11 maps income levels and the corresponding premiums.

To assess whether these cross-group valuation differences are statistically meaningful, we run single-constraint Wald tests of the equality of the low- and high-group

**Table 6** Income heterogeneity by comprehensive-income tercile (M6)

| Feature × group | Low × | High × |
|---|---|---|
| Water ratio | +0.076 ($p = 0.650$) | +0.434 ($p = 0.035$) |
| Bright pixel ratio | +0.187 ($p = 0.019$) | +0.030 ($p = 0.691$) |
| Texture entropy | +0.037 ($p = 0.050$) | +0.000 ($p = 0.993$) |
| Mean brightness | −0.001 ($p = 0.078$) | +0.000 ($p = 0.907$) |
| Other 4 features | n.s. | n.s. |

coefficients for each feature (Table 7). In the price-proxy model (M4), the texture-entropy discount ($\chi^2 = 6.31$, $p = 0.012$) and the mean-brightness premium ($\chi^2 = 5.43$, $p = 0.020$) differ significantly between low- and high-price communities. In the comprehensive-income model (M6), the bright-pixel-ratio difference ($\chi^2 = 4.52$, $p = 0.034$), the water-ratio difference ($\chi^2 = 5.81$, $p = 0.016$) and the mean-brightness difference ($\chi^2 = 4.69$, $p = 0.030$) are all significant. The Wald tests therefore place the income gradient in visual valuation on firmer statistical footing and identify bright built form as the main dimension on which low- and high-income communities disagree about the visual environment.

**Table 7** Formal Wald tests of cross-group coefficient differences ($H_0$: low = high)

| | Price proxy (M4) | | Income proxy (M6) | |
|---|---|---|---|---|
| Feature | $\chi^2$ | $p$ | $\chi^2$ | $p$ |
| Mean brightness | 5.43 | 0.020 | 4.69 | 0.030 |
| Texture entropy | 6.31 | 0.012 | 1.12 | 0.289 |
| Bright pixel ratio | 0.09 | 0.760 | 4.52 | 0.034 |
| Water ratio | 0.22 | 0.641 | 5.81 | 0.016 |
| Other features | n.s. | | n.s. | |

**Table 8** Interpretation of the two proxies

| Heterogeneity interpretation |
|---|
| Price proxy (neighborhood affluence): the mean and dispersion of brightness are more sensitive in high-price areas; |
| Income proxy (township income): the bright pixel share is significantly positive in low-income areas. |
| Overall: "brightness" is the core dimension of spatial differentiation in visual premiums, but its sign varies with the income/price level of the reference community. |

## Spatial concentration of premiums: association with price and income

At the grid level (100 m), each cell's premium is the M2-implied dot product of the coefficients and the centered feature values. The grid-level premium is essentially uncorrelated with the grid median unit price ($r = -0.03$), reflecting that pixel-share features (bright, dark, green, water), which are negatively priced, tend to be elevated in high-density, higher-price cores. At the township level, however, visual premiums rise with income: the township premium–income correlation is $r = 0.241$ ($p < 0.001$, $N = 284$), and the mean premium increases monotonically from $-0.047$ in low-income townships to $-0.023$ in middle-income and $-0.001$ in high-income townships. Fig. 3 documents the grid-level premium–price association and how visual premiums rise with township income; Fig. 4 maps the neighborhood visual-morphology premium and the grid price level.

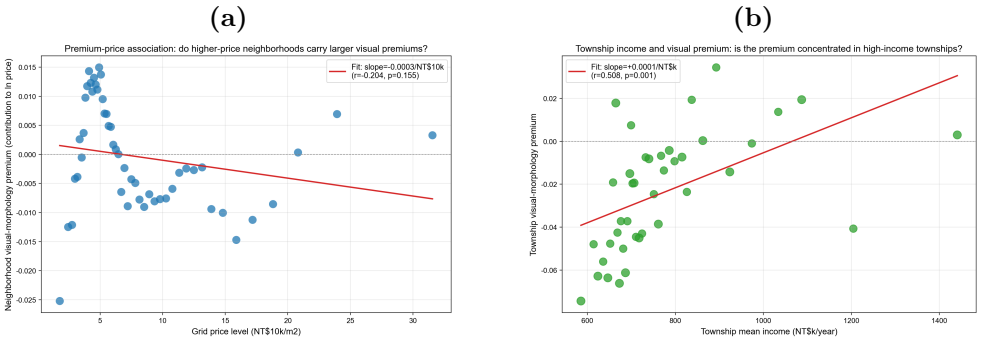



**Fig. 3** Visual premiums and housing-market outcomes. (a) Association between grid-level visual premiums and housing prices. (b) Visual premiums across income terciles; the bright built-form premium is concentrated in low-income communities.

At the township scale, visual-morphology premiums thus rise systematically with income, even as the bright built-form component favored by low-income communities is priced against the market average. **The capitalization benefits of the visual environment accrue disproportionately to high-income communities, while the bright built-form premium accrues mainly to low-income communities**. Visual capitalization thus tracks spatial inequality directly.

## Robustness and limitations

### Scale robustness (re-estimating M2 with 300 m / 500 m / 1,000 m neighborhoods)

We re-extract morphological features within 300 m and 1,000 m neighborhoods and re-estimate M2 on the grids for which imagery is available at those scales ($N \approx 381{,}000$ for the 300 m and 1,000 m columns; the 500 m baseline uses $N = 544{,}232$). Table 9 compares the coefficients across scales.

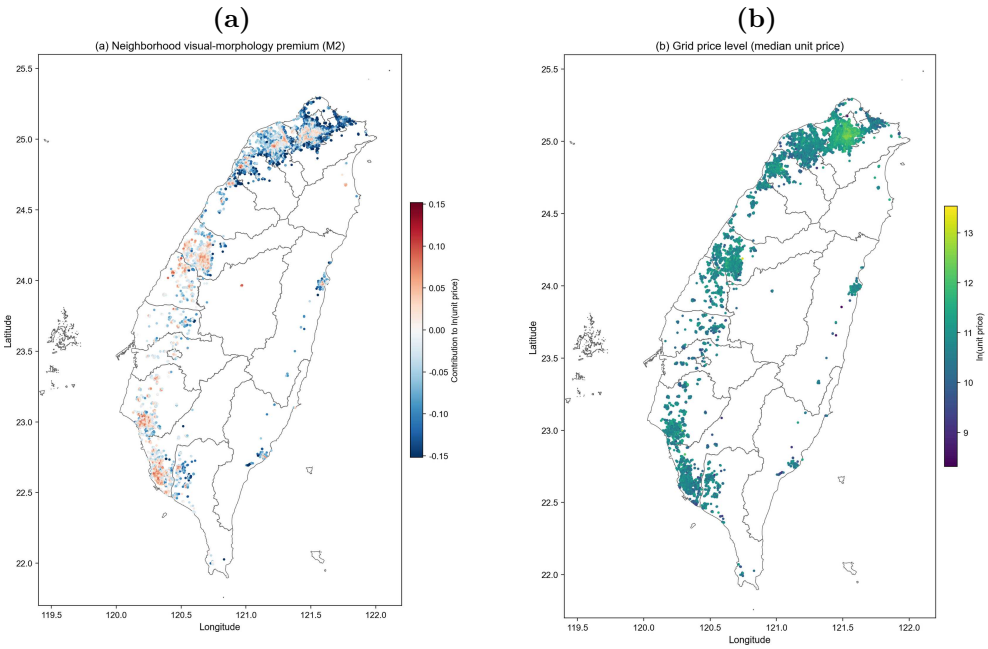



**Fig. 4** Maps of the visual-morphology premium and the grid price level across the Taiwan region. (a) Neighborhood visual-morphology premium implied by the M2 coefficients. (b) Grid price level (median unit price).

The core conclusions are robust across scales: **edge density (positive premium) and green ratio (negative association)** are significant with consistent signs throughout the 300 m–1,000 m range. The edge-density coefficient increases with the neighborhood scale (0.20, 0.32, 0.32), indicating that built-texture complexity is valued strongly at broader scales. The **water ratio** loses significance at the 1,000 m scale ($p = 0.161$) and **brightness standard deviation** is only significant at the 300 m and 1,000 m scales; these features should be interpreted with caution. The **bright pixel ratio** is significant at 300 m and 500 m but not at 1,000 m, suggesting that the bright built-form discount is a localized phenomenon.

## Fixed effects, clustering and sampling

All models absorb township-by-year fixed effects with township-clustered standard errors; income/price-proxy heterogeneity is introduced through interactions, which remain compatible with the fixed effects, and identification therefore comes from morphological variation within townships. Geocoding and satellite matching tilt the sample toward urban areas (New Taipei, Kaohsiung, Taoyuan, Taipei); counties and townships without township-level income statistics (roughly 3.0% of transactions, chiefly Hsinchu City districts and parts of Taitung and Chiayi counties) are excluded from the income-heterogeneity models. Income terciles are transaction-weighted rather than township-counted to avoid unbalanced group sizes.

**Table 9** Scale robustness of morphological main effects (M2)

| Feature | 300 m | 500 m (baseline) | 1,000 m |
|---|---|---|---|
| Edge density | +0.195 ($p = 0.003$) | +0.324 ($p < 0.001$) | +0.319 ($p = 0.002$) |
| Green ratio | −0.198 ($p < 0.001$) | −0.196 ($p < 0.001$) | −0.212 ($p < 0.001$) |
| Brightness std | −0.004 ($p < 0.001$) | −0.001 ($p = 0.113$) | −0.004 ($p = 0.008$) |
| Water ratio | −0.154 ($p = 0.002$) | −0.146 ($p = 0.009$) | −0.134 ($p = 0.161$) |
| Bright pixel ratio | −0.091 ($p = 0.032$) | −0.141 ($p = 0.002$) | −0.075 ($p = 0.211$) |
| Dark pixel ratio | −0.086 ($p = 0.181$) | −0.129 ($p = 0.040$) | −0.001 ($p = 0.989$) |
| Texture entropy, mean brightness | n.s. | n.s. | n.s. |
| $R^2$ | 0.7746 | 0.7755 | 0.7743 |

## Proxy limitations

Comprehensive income tax data are reported on a filing basis (excluding tax-exempt amounts and government transfers) and cover only 2012–2019; the price proxy correlates with the income proxy at $r = 0.708$ rather than perfectly.

## Urban heterogeneity: excluding the Taipei metropolitan area

As a check on whether Taipei's market drives the results, we re-estimate M2 after dropping Taipei City and, in a stricter variant, the entire Taipei metropolitan area (Taipei City and New Taipei) (Table 10, Fig. 5). The main conclusions hold: the green ratio retains a significant negative association (−0.212, $p < 0.001$ excluding Taipei City; −0.253, $p < 0.001$ excluding the Taipei metro), brightness standard deviation retains a significant discount once the Taipei metro is excluded (−0.002, $p = 0.006$), and the edge-density premium remains significantly positive in both variants (+0.305, $p < 0.001$ and +0.259, $p < 0.001$), indicating that the built-texture premium is not an artifact of the Taipei market.

**Table 10** Urban heterogeneity: M2 excluding the Taipei metropolitan area

| Feature | Full sample | Excl. Taipei City | Excl. Taipei metro |
|---|---|---|---|
| Green ratio | −0.196 ($p < 0.001$) | −0.212 ($p < 0.001$) | −0.253 ($p < 0.001$) |
| Edge density | +0.324 ($p < 0.001$) | +0.305 ($p < 0.001$) | +0.259 ($p < 0.001$) |
| Brightness std | −0.001 ($p = 0.113$) | −0.001 ($p = 0.107$) | −0.002 ($p = 0.006$) |
| Water ratio | −0.146 ($p = 0.009$) | −0.170 ($p = 0.004$) | −0.141 ($p = 0.009$) |
| Bright pixel ratio | −0.141 ($p = 0.002$) | −0.040 ($p = 0.411$) | +0.080 ($p = 0.008$) |
| $N$ | 544,232 | 471,647 | 335,950 |
| $R^2$ | 0.7755 | 0.7069 | 0.6382 |

## Accessibility control: distance to the nearest rail station

Because proximity to public transit is a key locational amenity that could confound visual premiums, we add the log distance to the nearest of 545 rail-transit stations (TRA rail, HSR, and metro) as a control in M2 (Table 11, Fig. 5). Proximity to transit is significantly capitalized ($\ln(1 + \text{dist_km}) = -0.127$, $p < 0.001$), yet the visual-morphology coefficients move only slightly: edge density remains significantly positive ($+0.254$, $p < 0.001$), the green-ratio discount remains significant ($-0.149$, $p < 0.001$), and the water-ratio discount remains significant ($-0.157$, $p = 0.006$). No feature that is significant in the baseline changes sign or loses significance. Visual premiums are therefore not merely proxies for transit accessibility.

**Table 11** Accessibility control: M2 with and without the rail-accessibility control

| Feature | M2 | M2 + accessibility |
|---|---|---|
| Green ratio | $-0.196$ ($p < 0.001$) | $-0.149$ ($p < 0.001$) |
| Edge density | $+0.324$ ($p < 0.001$) | $+0.254$ ($p < 0.001$) |
| Water ratio | $-0.146$ ($p = 0.009$) | $-0.157$ ($p = 0.006$) |
| Brightness std | $-0.001$ ($p = 0.113$) | $-0.001$ ($p = 0.181$) |
| Texture entropy | $-0.020$ ($p = 0.307$) | $-0.023$ ($p = 0.226$) |
| $\ln(1 + \text{dist_km})$ | – | $-0.127$ ($p < 0.001$) |
| $N$ | 544,232 | 544,232 |
| $R^2$ | 0.7755 | 0.7788 |

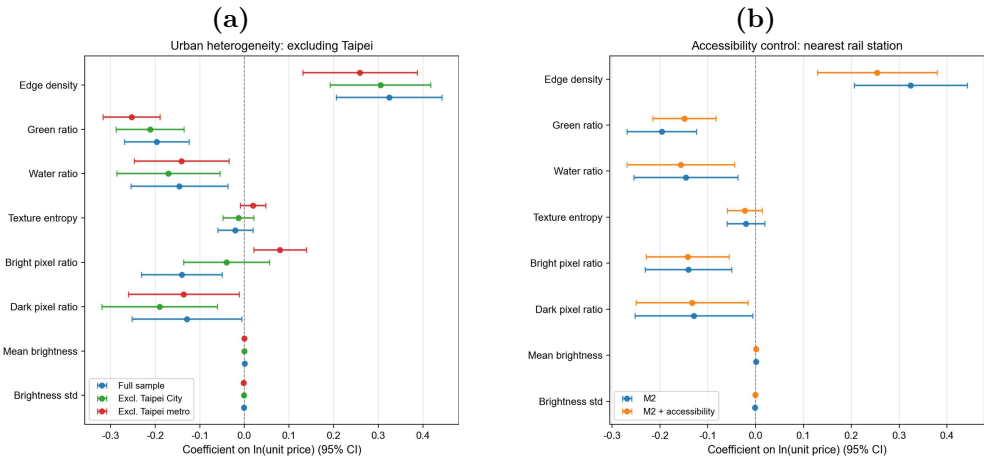



**Fig. 5** Robustness of visual-morphology premiums. (a) M2 coefficients in the full sample and after excluding Taipei City or the entire Taipei metropolitan area. (b) M2 coefficients with and without the rail-accessibility control. Error bars are 95% confidence intervals based on township-clustered standard errors.

## A world-model perspective: visual trajectories and the expectation channel

Knowing which visual features markets price does not tell us whether prices also respond to how neighborhoods change. To address the second half of that question we adopt a world-model perspective [15–17] that treats each 500 m neighborhood as a small dynamical system: for every grid cell we fit linear trajectories of its Prithvi-EO-2.0 visual features over time, then ask whether these temporal descriptors add explanatory power beyond the static snapshot, and whether prices capitalize *anticipated* future visual change.

### Spatiotemporal context adds nearly as much as deep static embeddings

For each of the 13 dimensions of the Prithvi-EO-2.0 vision-transformer foundation model [18–20] (NDVI, NDWI, brightness and ten embedding principal components) we fit a per-grid ordinary-least-squares slope of the feature on calendar year (requiring at least three annual observations), and append the thirteen slopes and thirteen first-differences to the handcrafted model. On the world-model subsample ($N$ = 421,205; grids with sufficient annual coverage), adding these twenty-six temporal descriptors raises $R^2$ from 0.7801 (M2_hand) to 0.7829 (M2_tc), a gain of +0.0028, nearly identical to the +0.0031 contributed by the full deep static embedding set (M2_all; Table 12). The temporal trajectory of a neighborhood's visual features is thus almost as informative, in a marginal-$R^2$ sense, as the entire high-dimensional static embedding, and it does so with a compact set of interpretable dynamics parameters.

**Table 12** Explanatory power of static versus dynamic visual information (world-model subsample, $N$ = 421,205; not directly comparable to Table 2, which does not require repeated annual coverage)

| Model | Feature set | $R^2$ | $\Delta R^2$ vs. M2_hand |
|---|---|---|---|
| M2_hand | 8 handcrafted morphology | 0.7801 | 0 |
| M2_all | + Prithvi spectral + embeddings | 0.7832 | +0.0031 |
| M2_tc | + 26 temporal descriptors (13 slopes, 13 diffs) | **0.7829** | +0.0028 |

### Prices capitalize the visible present, not the anticipated future

We formalize the world model's predictive step with an expected-visual-change index. For grid $g$ and deep feature $k$, the estimated slope $\hat{s}_{gk}$ describes how the neighborhood's visual state is evolving; weighting these slopes by the cross-sectional implicit prices of the deep features ($\hat{\gamma}_k$, from M_all) yields $X_{gi} = \sum_k \hat{s}_{gk}\hat{\gamma}_k$, which measures whether the neighborhood's *anticipated* visual change is one the market values positively. If prices capitalized expectations, $X$ should enter the hedonic model with a positive coefficient. It does not: in the full-period specification the coefficient is positive but

statistically insignificant ($\hat{\beta}_X = +0.97$, $p = 0.118$). In a no-lookahead specification, slopes are estimated only from 2012–2019 imagery, and only transactions from 2020 onward are retained, so that no information from the transaction period enters the forecast. Here the coefficient is significantly negative ($\hat{\beta}_X = -0.39$, $p = 0.012$; Table 13 and Fig. 6). Consistent with the discussion below, housing markets in the Taiwan region thus appear to price the *current, visible* morphology of a neighborhood rather than its anticipated future appearance. If anything, neighborhoods on improving visual trajectories were priced lower in the recent period, a pattern more consistent with early-stage appreciation discounting than with anticipation.

**Table 13** Expectation channel: coefficient on the expected-visual-change index $X$ (log unit price)

| Specification | Coef. | SE | $p$ | $N$ |
|---|---|---|---|---|
| Full-period slopes | +0.973 | 0.621 | 0.118 | 421,205 |
| No-lookahead (slopes 2012–2019; sample ≥2020) | −0.388 | 0.154 | 0.012 | 249,027 |

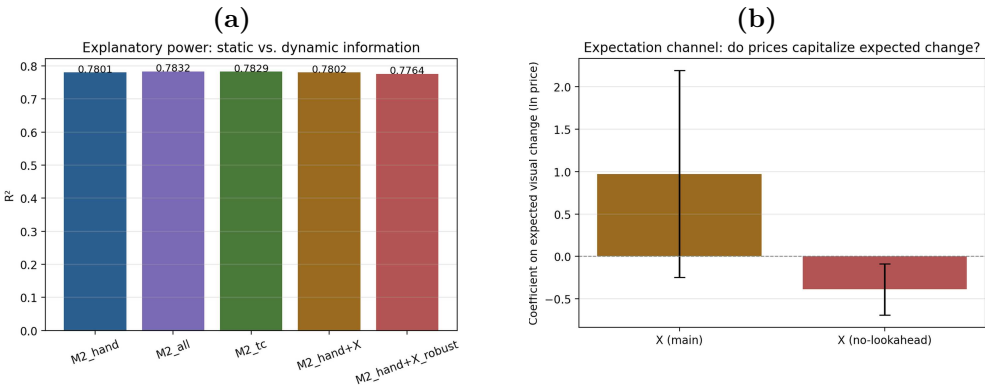



**Fig. 6** World-model evidence. (a) $R^2$ of the static handcrafted (M2_hand), static deep (M2_all) and spatiotemporal-context (M2_tc) models. (b) Coefficient on the expected-visual-change index in the full-period and no-lookahead specifications, with 95% confidence intervals; the significantly negative no-lookahead coefficient indicates that prices capitalize current visible morphology rather than anticipated future change.

## Out-of-sample validation of the dynamics module

A world model should, by definition, anticipate the future states of the system it represents. We therefore ask whether the dynamics module, the per-grid linear trajectories, can predict a neighborhood's future visual state out of sample. Table 14 and Fig. 7 validate the trajectories on 36,069 neighborhoods with at least eight annual observations: each trajectory is estimated on the first 60% of the sample years and used to forecast the visual state one and three years ahead.

**Table 14** Out-of-sample forecast validation of the world-model dynamics module (36,069 neighborhoods with ≥8 annual observations; RMSE in $z$-units; directional accuracy is the share of correctly signed one-year visual changes; the last column is the share of grid cells for which the trajectory beats persistence)

| Horizon | RMSE | | | Directional | Dynamics beat |
|---|---|---|---|---|---|
| | World-model | Persistence | Mean | accuracy (%) | persistence (%) |
| 1 year ($h$=1) | 0.547 | 0.178 | 0.592 | 55.1 | 16.9 |
| 3 years ($h$=3) | 0.970 | 0.204 | 0.604 | 51.5 | 11.3 |

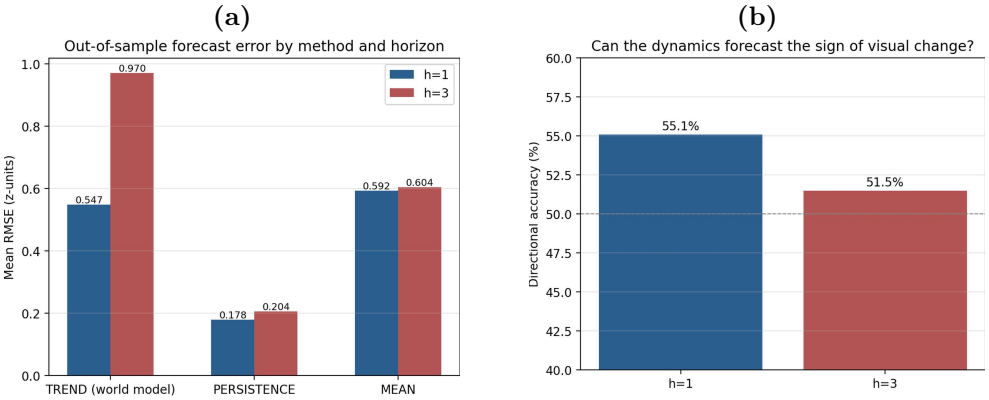



**Fig. 7** Out-of-sample validation of the world-model dynamics module. (a) RMSE of the trajectory forecast, the persistence (random-walk) benchmark and the historical-mean benchmark at one- and three-year horizons, in $z$-units. (b) Directional accuracy of the predicted one-year change in the visual state; values above the 50% chance line indicate that the trajectories carry genuine forward-looking information.

Two conclusions follow. The first is that neighborhood visual states in the Taiwan region are highly persistent. Forecasting "no change" (the persistence benchmark) achieves an RMSE of only 0.18 in $z$-units, roughly one-fifth of the cross-sectional standard deviation of a visual-state dimension, so the best one-year-ahead forecast of a neighborhood's appearance is essentially its current appearance.

Even so, the trajectories carry usable short-horizon directional information: the predicted sign of the one-year visual change is correct in 55.1% of cases ($p < 0.001$, far above the 50% chance level), and the trajectory forecast outperforms the historical-mean benchmark.

This directional signal decays with horizon (51.5% at three years), and extrapolated point forecasts of levels are dominated by persistence. The pattern is consistent with a mean-reverting visual environment in which most per-grid trend estimates are small relative to annual noise. Because the visual future is almost the visual present, there is little anticipatable variation for markets to price, which is precisely why prices capitalize the visible present rather than the anticipated future. At the same time, the dimensions on which genuine trends exist (several embedding components, especially PC5, PC9 and PC10) are exactly those on which the trajectory beats persistence for a

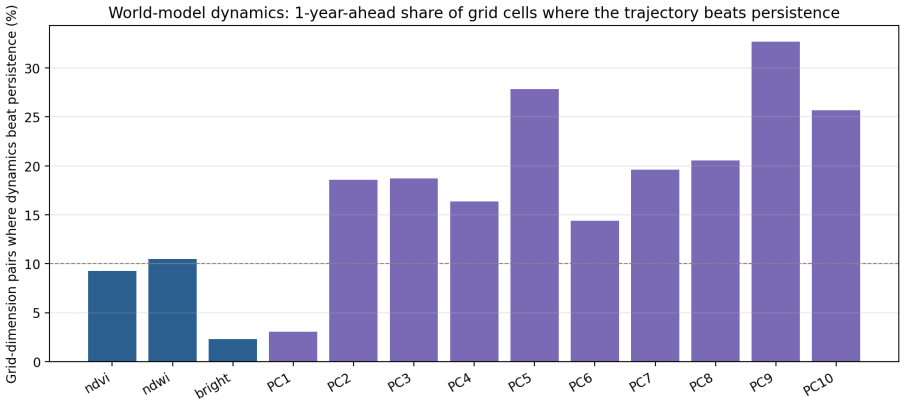



**Fig. 8** Share of grid cells (by visual-state dimension) for which the world-model trajectory beats the persistence benchmark in one-year-ahead level forecasts. Spectral dimensions are shown in blue and embedding principal components in purple; the trajectory is most informative for embedding components with systematic trends (PC5, PC9, PC10).

substantial share of grid cells (26–33%; Fig. 8), indicating that the dynamics module is informative where visual change is systematic rather than noise-driven.

## Model validation with a geospatial foundation model

A natural concern is whether our eight handcrafted features omit visual information that is salient to housing markets. To test this, we validated the interpretable measures against **Prithvi-EO-2.0** [21, 22], a vision-transformer foundation model [18–20] pre-trained on 4.2 M Harmonized Landsat–Sentinel (HLS) scenes. For every transaction's 500 m neighborhood we extracted two complementary sets of deep features from a six-band (blue, green, red, narrow-NIR, SWIR1, SWIR2) cloud-free annual composite: (i) three *spectral-semantic* metrics (NDVI, NDWI and mean brightness), and (ii) ten principal components of the model's 1,024-dimensional patch embeddings, which encode unsupervised, low-level-to-semantic urban appearance. Table 15 and Fig. 9 compare model fit on the Prithvi-matched subsample ($N = 492{,}691$).

**Table 15** Model fit with handcrafted versus Prithvi-EO-2.0 features (Prithvi-matched subsample, $N = 492{,}691$)

| Model | Feature set | $R^2$ | $\Delta R^2$ vs. M2_hand |
|---|---|---|---|
| M1 | Building attributes | 0.7701 | −0.0069 |
| M2_hand | + 8 handcrafted morphology | 0.7770 | 0 |
| M2_spec | + 3 spectral-semantic metrics | 0.7792 | +0.0022 |
| M2_emb | + 10 embedding principal components | 0.7718 | −0.0052 |
| M_all | + handcrafted + spectral + embedding | **0.7807** | +0.0037 |

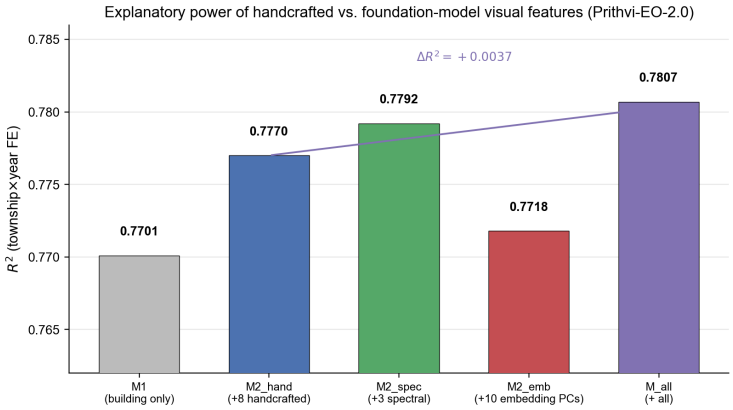



**Fig. 9** Explanatory power of handcrafted versus Prithvi-EO-2.0 visual features for log housing price (township×year fixed effects). M_all, which augments the eight handcrafted features with spectral-semantic metrics and deep embedding principal components, raises $R^2$ to 0.7807, a gain of +0.0037 over the handcrafted-only model.

Comparing the feature sets directly shows why the interpretable approach has the edge. On the Prithvi-matched subsample, the deep spectral-semantic metrics are jointly informative yet fully consistent with the handcrafted findings: NDVI and NDWI enter with significantly negative coefficients ($-2.50$, $p < 0.001$ and $-1.98$, $p < 0.001$), mirroring the negative green- and water-ratio premia we document, while mean brightness adds nothing. What stands out more is that the ten embedding principal components, each condensing rich low-level-to-semantic urban appearance learned from 4.2 M HLS scenes, pick up essentially no residual signal beyond the eight interpretable features: only three of ten are significant at the 5% level, and jointly they lift $R^2$ by a mere +0.0017 over the building-only baseline, actually *underperforming* the handcrafted set on its own ($R^2 = 0.7718$ vs. 0.7770). Even the most flexible specification (M_all) reaches $R^2 = 0.7807$, just +0.0037 above the handcrafted-only model. A 300-million-parameter multispectral foundation model thus adds less than four one-thousandths of $R^2$ on top of eight transparent, policy-relevant measures.

## Feature-space counterfactual policy simulations

Because the eight features map directly onto regulable elements of the built environment, the estimated implicit prices support transparent, feature-space counterfactual simulations of visual-environment interventions. Holding all other attributes fixed, we revalue affected homes under four policy scenarios built from the M2 implicit prices (Table 16; Fig. 10 and Fig. 11). Reducing the dark-pixel share in the darkest neighborhoods raises affected values by +2.79% (S2), reducing the bright-pixel share by +2.50% (S3), and the comprehensive package, which greens low-green neighborhoods while reducing dark and bright built form, raises values by +1.14% on average, corresponding to roughly NT$350 billion of aggregate value uplift across the 82.9% of transactions affected.

**Table 16** Feature-space counterfactual policy scenarios (mean predicted price effect with 95% confidence interval based on coefficient uncertainty; aggregate uplift in NT$ billions)

| Scenario | Affected (share) | Mean effect [95% CI] | Uplift (NT$B) |
|---|---|---|---|
| S1 Green+ (low-green, +1 SD) | 1,265,720 (40.0%) | −4.01% [−5.45, −2.54] | −557 |
| S2 Dark− (darkest, −1 SD) | 1,276,727 (40.3%) | +2.79% [+0.13, +5.53] | +469 |
| S3 Bright− (brightest, −1 SD) | 1,006,133 (31.8%) | +2.50% [+0.89, +4.14] | +210 |
| S4 Comprehensive (S1+S2+S3) | 2,625,532 (82.9%) | +1.14% [−2.26, +4.66] | +350 |

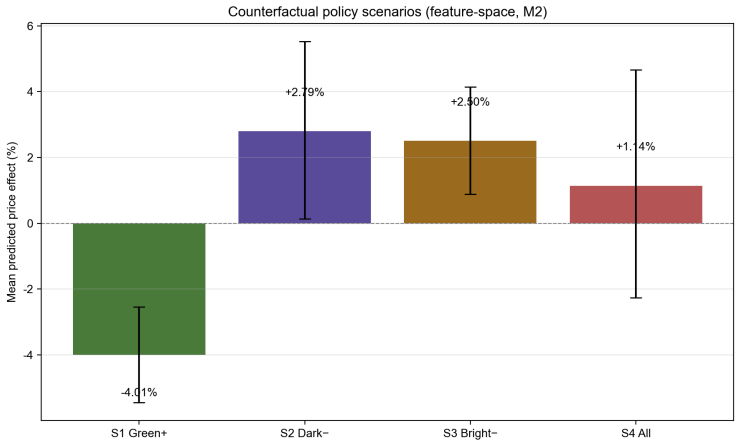



**Fig. 10** Mean predicted price effect of the four feature-space counterfactual policy scenarios, with 95% confidence intervals based on coefficient uncertainty.

Any policy reading of these simulations should be tempered. They are *feature-space* counterfactuals that inherit the partial, multicollinear morphological coefficients: raising the green ratio is predicted to *lower* values (−4.01%, S1) because, within the visual-morphology space, greenness marks older, lower-density neighborhoods rather than park or tree amenity, the same composition effect already discussed for the cross-sectional green premium. They also hold all other attributes fixed and ignore general-equilibrium and supply responses. Their value is therefore diagnostic: they translate the estimated implicit prices into policy-relevant, spatially targeted magnitudes and make the model's normative content explicit for urban planning.

# Discussion

Using transaction-level microdata from the Taiwan region, this study documents which neighborhood visual-morphology features are capitalized into housing prices, how these valuations have changed, and who captures them. Five findings stand out. Built-texture complexity is positively priced while visual disorder is discounted. The structure of capitalization has evolved over time, with negative premia on green, water

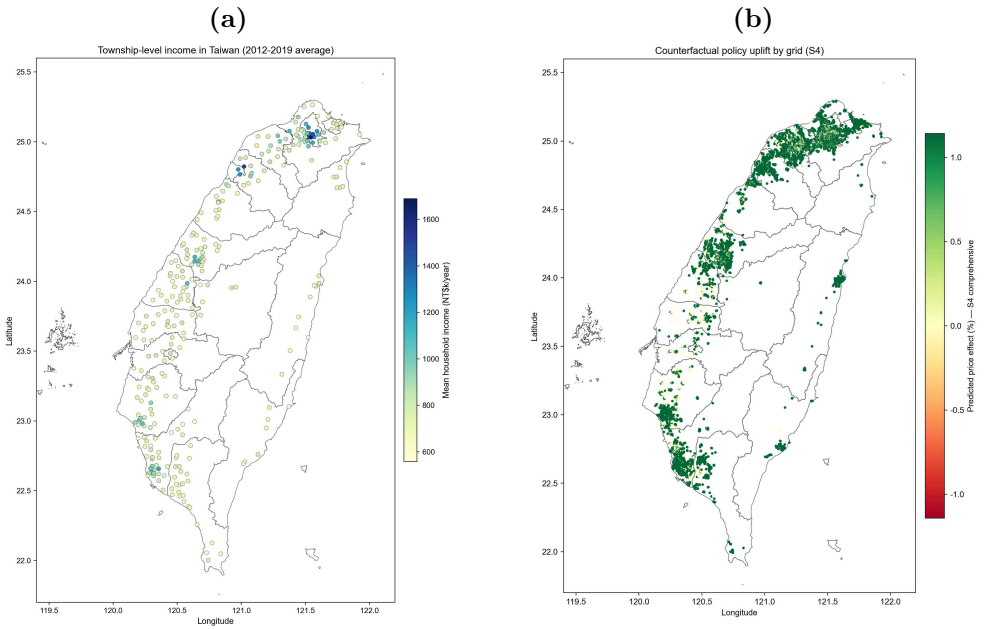



**Fig. 11** Spatial patterns of visual premiums and counterfactual price effects. (a) Township-level income terciles and the spatial distribution of visual premiums. (b) Spatial distribution of the predicted price effect of the comprehensive counterfactual scenario (S4) across the grid cells of the Taiwan region.

and dark built form deepening. Visual premiums are systematically concentrated in high-price, high-income communities, with a pronounced income-gradient reversal in the valuation of bright built form. Temporal visual trajectories carry nearly as much incremental information as the full deep-embedding set from a 300-million-parameter geospatial foundation model. And out-of-sample validation reveals that neighborhood visual morphology evolves as an almost perfect random walk, so the visible present is effectively the best available forecast of the visible future. The world-model analysis confirms that prices capitalize the current, observable morphology of a neighborhood rather than anticipated visual change, ruling out expectations-driven capitalization as a competing explanation of the cross-sectional premia.

These results extend the growing literature on computer vision and satellite imagery in urban economics [4, 5] and hedonic valuation of neighborhood amenities [1, 2, 6, 7]. The finding that visual capitalization is skewed toward already-advantaged communities implies that the visual environment, as an unpriced local public good, may exacerbate rather than mitigate spatial inequality, an important consideration for urban planning and equitable development policy. The handcrafted features are not a crude simplification but a near-sufficient statistic for the visual information that housing markets in the Taiwan region capitalize. The interpretable set carries practical advantages. Each coefficient maps to a concrete, regulable element of the built environment, so the results can be acted on. The measures are cheap to reproduce, resting on eight open-access satellite statistics computed in minutes rather than a

300-million-parameter foundation model. The finding that the foundation model adds almost nothing beyond these features supports their external validity. And the comparison leaves a forward-looking benchmark: future studies can use Prithvi-EO-2.0 embeddings to test whether any candidate feature adds information beyond the eight proposed here.

The analysis faces several limitations. Income data are available only through 2019, and income terciles are constructed at the township level. The price-based proxy is imperfectly correlated with income ($r = 0.708$). Geographic matching concentrates the sample in urban counties. The core conclusions nonetheless survive formal Wald tests of cross-group coefficient differences, the exclusion of the Taipei metropolitan area, explicit controls for rail-transit accessibility, and validation against deep features from a large geospatial foundation model (Prithvi-EO-2.0), which adds only marginal explanatory power over the interpretable features. The world-model analysis is subject to further caveats: per-grid visual trajectories are fitted with simple linear dynamics, the expected-visual-change index inherits the coefficients of the deep-feature model, and the counterfactual simulations are feature-space exercises that hold all else fixed and ignore supply and general-equilibrium responses. Three extensions seem most promising. Replicating the analysis in other cities and countries would test the external validity of visual capitalization. The mechanisms through which the visual environment enters prices, direct consumption amenities versus informational signals, deserve closer study. And the linear per-grid dynamics could be relaxed toward nonlinear neighborhood evolution and genuine out-of-sample forecasts of visual change.

# Methods

## Data

We combine three data sources: (i) the Taiwan region's real-estate actual-registration transaction records for 2012–2026; (ii) Sentinel-2 satellite imagery [23] (10 m/30 m) matched to each transaction's year and location; and (iii) township-level comprehensive income tax statistics from the Taiwan region's Ministry of Finance (2012–2019). For the foundation-model validation we additionally use six-band (blue, green, red, narrow-NIR, SWIR1, SWIR2) cloud-free annual composites from Sentinel-2 and Landsat, sourced via Google Earth Engine and processed through the Prithvi-EO-2.0 vision transformer [21, 22], which is evaluated on the standardized GEO-Bench Earth-monitoring benchmark [24]. For the accessibility robustness check we use the coordinates of 545 rail-transit stations (TRA rail, HSR and metro) from the Railways Administration of the Taiwan region and OpenStreetMap [25].

The complete analytical pipeline is implemented in the open-source scripts `research/step1~step22_*.py` and `step30~step32_*.py`; all result tables are stored in `data/research/model_results.csv`, `step6_results.csv`, `temporal_zyear_results.csv`, `step9_results.csv`, `step11~step14_*.csv`, `prithvi/worldmodel_results.csv` and `prithvi/counterfactual_scenarios.csv`.

Fig. 12 shows example Sentinel-2 true-color composites used in the study. The Taiwan region is tessellated into $0.2^\circ \times 0.2^\circ$ tiles per county and year; each tile is a cloud-masked median composite (at most 20 scenes, cloud fraction below 60%)

spanning the calendar year, providing a single consistent "annual image" for every neighborhood in every year. Transactions are geocoded to a 100 m grid; for each grid cell we crop a 500 m-radius neighborhood from the composite of the corresponding year, and compute the morphological features described below.

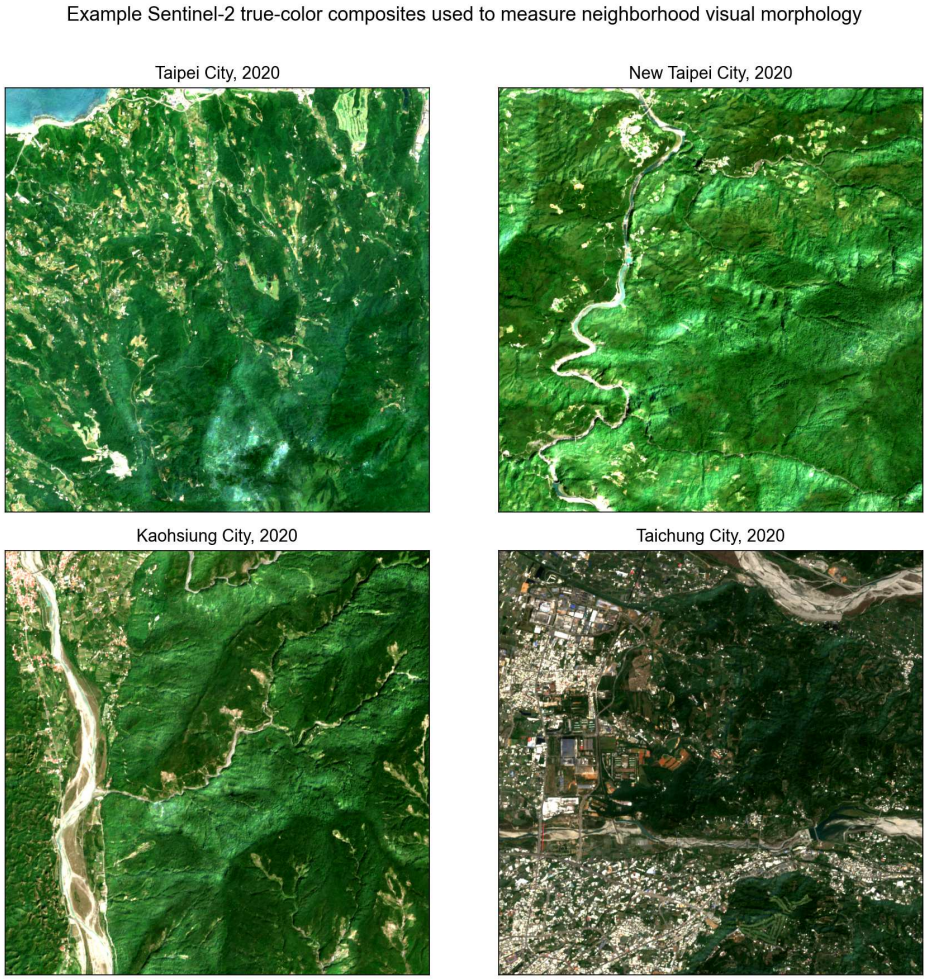



**Fig. 12** Example Sentinel-2 true-color composites (2020) used to measure neighborhood visual morphology. Central urban detail is shown for Taipei City, New Taipei City, Kaohsiung City and Taichung City.

## Sample construction

Residential transactions are geocoded and matched to a 100 m satellite grid. Neighborhood features are extracted within a 500 m radius using the image of the corresponding

year, with cloud-cover quality control (neighborhoods with more than 30% cloud cover are dropped). From the matched population we sample 600,000 transactions and drop observations with missing values, yielding a final estimation sample of $N = 544{,}232$.

## Feature extraction

For each neighborhood window, pixels are classified as cloud (all RGB channels $> 235$); remaining pixels are used to compute the eight morphological features. Let $L = 0.299R + 0.587G + 0.114B$ denote pixel luminance. The land-cover ratios are computed as

$$\begin{aligned} \text{green_ratio} &= \frac{|\{p : G_p > R_p \text{ and } G_p > B_p\}|}{N}, \\ \text{water_ratio} &= \frac{|\{p : B_p > R_p + 15,\ B_p > G_p + 15,\ L_p < 160\}|}{N}, \end{aligned} \tag{1}$$

where $N$ is the number of non-cloud pixels in the neighborhood. Bright and dark pixel ratios, mean brightness and brightness standard deviation are defined analogously,

$$\begin{aligned} \text{bright_ratio} &= \frac{|\{p : L_p > 170\}|}{N}, \qquad \text{dark_ratio} = \frac{|\{p : L_p < 50\}|}{N}, \\ \text{mean_brightness} &= \frac{1}{N}\sum_p L_p, \qquad \text{brightness_std} = \sqrt{\frac{1}{N}\sum_p \big(L_p - \overline{L}\big)^2}, \end{aligned} \tag{2}$$

where $\overline{L}$ denotes the sample mean of luminance. Edge density proxies built-texture complexity and is computed as the share of pixels with a luminance gradient exceeding a threshold,

$$\text{edge_density} = \frac{|\{p : \|\nabla L_p\| > \tau\}|}{N}, \tag{3}$$

with $\|\nabla L_p\|$ estimated from $3\times3$ Sobel operators and $\tau$ the 75th percentile of the gradient distribution. Texture entropy is the Shannon entropy of the 32-bin luminance histogram,

$$\text{texture_entropy} = -\sum_{k=1}^{32} h_k \log h_k, \tag{4}$$

where $h_k$ is the fraction of pixels in bin $k$. All features are sample-centered before estimation.

For the Prithvi-EO-2.0 validation, six-band cloud-free annual composites are built per township and year (median composite of Sentinel-2 or Landsat surface reflectance with cloud masking, at 10 m–30 m resolution). Three spectral-semantic metrics are computed at the pixel level as

$$\text{NDVI} = \frac{\rho_{\text{NIR}} - \rho_{\text{Red}}}{\rho_{\text{NIR}} + \rho_{\text{Red}}}, \qquad \text{NDWI} = \frac{\rho_{\text{Green}} - \rho_{\text{NIR}}}{\rho_{\text{Green}} + \rho_{\text{NIR}}}, \tag{5}$$

where $\rho_{\cdot}$ denotes surface reflectance, and are averaged within the same 500 m neighborhoods. The 1,024-dimensional patch embeddings of the pre-trained Prithvi-EO-2.0-300M encoder are computed on $224\times224$ windows covering each tile, bilinearly

sampled at every transaction's grid cell, and standardized. Let $E \in \mathbb{R}^{n\times 1024}$ be the row-standardized embedding matrix over the $n = 685{,}858$ grid cells; a memory-efficient incremental principal component analysis then solves the eigenvalue problem

$$E^\top E\, v_k = \lambda_k v_k, \qquad C = E\, V_{10}, \tag{6}$$

where $V_{10} = [v_1, \ldots, v_{10}]$ collects the ten leading eigenvectors and $C$ is the $n\times 10$ matrix of embedding principal components entering the regression alongside the spectral metrics (see Results).

For the temporal specifications (M3, M5), the morphological features are additionally standardized to zero mean and unit variance within each year,

$$\widetilde{M}^k_{iy} = \frac{M^k_{iy} - \overline{M}^k_y}{s^k_y}, \tag{7}$$

where $\overline{M}^k_y$ and $s^k_y$ are the mean and standard deviation of feature $k$ among year-$y$ transactions (years without within-year variation are set to zero). This purges year-to-year measurement drift in the annual composites, so that cross-period coefficient contrasts are measured in within-year-SD units.

## Analytical pipeline

The analytical pipeline implements this study's world-model perspective on visual-morphology capitalization in five stages: (1) Transaction records are cleaned and geocoded to a 100 m grid; (2) for each year and location, a cloud-masked annual satellite composite is assembled and the eight morphological features are extracted within a 500 m buffer; (3) features are merged with transaction, income and accessibility data to build the estimation sample; (4) hedonic models M1–M3 and the heterogeneity specifications are estimated with township-by-year fixed effects; and (5) the results are validated through scale, metropolitan, accessibility and Prithvi-EO-2.0 foundation-model checks.

## Hedonic model

The dependent variable is ln(price per unit area). Baseline model M1 includes building attributes; M2 adds the eight centered morphological features; M3 adds their interactions with (year − 2019); M_all additionally includes the three spectral-semantic metrics and the ten embedding principal components of the Prithvi-EO-2.0 validation (see Results). Heterogeneity models (M4, M6) interact morphology with tercile indicators for the price and income proxies, respectively. In the temporal models (M3, M5) the morphology features are within-year standardized as described in the Results, so that cross-period coefficient contrasts are measured in within-year-SD units. All models absorb township-by-year fixed effects, and standard errors are clustered at the township level. Grouped subsample regressions (M5) split the sample at 2018/2019. Cross-group coefficient differences in M4 and M6 are evaluated with single-constraint

Wald tests of the equality of the low- and high-group coefficients. In the urban-heterogeneity check, M2 is re-estimated on samples that exclude Taipei City or the entire Taipei metropolitan area. In the accessibility check, M2 is re-estimated with $\ln(1+d)$, where $d$ is the great-circle distance from the transaction's grid cell to the nearest of the 545 rail-transit stations.

The baseline and main specifications are, for transaction $i$ in township $t$ and year $y$,

$$\ln P_{ity} = \alpha' X_{ity} + \mu_{ty} + \varepsilon_{ity}, \tag{8}$$

$$\ln P_{ity} = \alpha' X_{ity} + \beta' M_{ity} + \mu_{ty} + \varepsilon_{ity}, \tag{9}$$

where $P$ is the price per unit area, $X$ is the vector of building attributes, $M$ is the vector of the eight centered (or, in the temporal models M3 and M5, within-year-standardized) morphological features, $\mu_{ty}$ is the township-by-year fixed effect, and $\varepsilon$ is the error term. M1 (8) identifies $\alpha$ from within-township-year variation in building attributes; M2 (9) identifies the implicit prices $\beta$ of the visual-morphology features.

The trend model M3 augments M2 with linear interactions between morphology and time,

$$\ln P_{ity} = \alpha' X_{ity} + \beta' M_{ity} + \gamma' \big(M_{ity} \times (\text{year}_{iy} - 2019)\big) + \mu_{ty} + \varepsilon_{ity}, \tag{10}$$

so that the time-varying marginal effect of feature $k$ is $\beta_k + \gamma_k(\text{year}_{iy} - 2019)$. The foundation-model specification M_all augments M2 with the three spectral-semantic metrics $S$ and the ten embedding principal components $C$,

$$\ln P_{ity} = \alpha' X_{ity} + \beta' M_{ity} + \theta' S_{ity} + \lambda' C_{ity} + \mu_{ty} + \varepsilon_{ity}, \tag{11}$$

which identifies the incremental explanatory power of the deep features over the interpretable set (see Results). The heterogeneity models M4 (price terciles) and M6 (income terciles) let the implicit prices vary across groups,

$$\ln P_{ity} = \alpha' X_{ity} + \beta' M_{ity} + \delta_L' \left(M_{ity} \times \mathbb{1}\{g_t = \text{low}\}\right) + \delta_H' \left(M_{ity} \times \mathbb{1}\{g_t = \text{high}\}\right) + \mu_{ty} + \varepsilon_{ity}, \tag{12}$$

where $g_t$ is the tercile of the price (M4) or income (M6) proxy at the township level and the middle tercile is the reference group, so that $\delta_L$ and $\delta_H$ capture the differential premia of low- and high-group neighborhoods. The early/recent model M5 instead interacts morphology with a recent-period indicator,

$$\ln P_{ity} = \alpha' X_{ity} + \beta' M_{ity} + \delta' \left(M_{ity} \times \mathbb{1}\{\text{year}_{iy} \geq 2019\}\right) + \mu_{ty} + \varepsilon_{ity}, \tag{13}$$

with grouped subsample estimates splitting the sample at 2018/2019. Cross-group coefficient differences in M4 and M6 are evaluated with single-constraint Wald tests of $H_0: \beta_k^{\text{high}} = \beta_k^{\text{low}}$, whose statistic

$$W_k = \frac{\left(\hat{\beta}_k^{\text{high}} - \hat{\beta}_k^{\text{low}}\right)^2}{\widehat{\text{Var}}\left(\hat{\beta}_k^{\text{high}} - \hat{\beta}_k^{\text{low}}\right)} \sim \chi_1^2 \tag{14}$$

is computed from the township-clustered covariance matrix of the pooled interaction model. For the price/income analysis, the morphological premium of each transaction is

$$\hat{p}_i = \sum_{k=1}^{8} \hat{\beta}_k M_{ik}, \tag{15}$$

and is aggregated to township and grid levels for the income-gradient and spatial-concentration analyses in the Results,

$$\hat{p}_j = \frac{1}{n_j} \sum_{i \in j} \hat{p}_i, \tag{16}$$

where $n_j$ is the number of sampled transactions in township (or grid) $j$.

## Scale robustness

For the same set of grids, morphological features are re-extracted within 300 m and 1,000 m neighborhoods and M2 is re-estimated to verify that the core conclusions are stable across scales.

## World-model analysis

The world-model analysis (Results and the Discussion) treats each 500 m neighborhood as a small dynamical system built on the 13 Prithvi dimensions (NDVI, NDWI, brightness and ten embedding principal components). Figure 13 presents the architecture of this world model: it encodes each neighborhood's annual observation into a low-dimensional visual state (perception module), learns how that state evolves over time (dynamics module), and decodes the resulting trajectories into implicit prices and expectation indices (decision module), which are finally evaluated with no-lookahead and counterfactual checks. For each grid cell $g$ and dimension $k$, we estimate a linear trajectory of the feature on calendar year by ordinary least squares,

$$\hat{s}_{gk} = \frac{\sum_y (y - \bar{y}_g)\big(F_{gky} - \bar{F}_{gk}\big)}{\sum_y (y - \bar{y}_g)^2}, \tag{17}$$

requiring at least three annual observations; $\hat{s}_{gk}$ is the per-year change in feature $k$, and a first-difference of $F_{gky}$ is used for the recent-change descriptors. The spatiotemporal-context specification M2_tc augments the handcrafted model with the thirteen slopes and thirteen first-differences, and is compared with the static M2_hand and M2_all on a common world-model subsample of $N$ = 421,205 transactions whose grids satisfy the repeated-annual-coverage requirement (Table 12).

An unusual design choice here is to treat Prithvi-EO-2.0 as the world model's perception module rather than as a competing black-box predictor. As a geospatial foundation model, Prithvi-EO-2.0 is a vision transformer pretrained in a self-supervised, masked-autoencoder fashion on large-scale satellite imagery [21, 22, 26–31], so its 1,024-dimensional patch embeddings encode generalizable spectral, spatial and

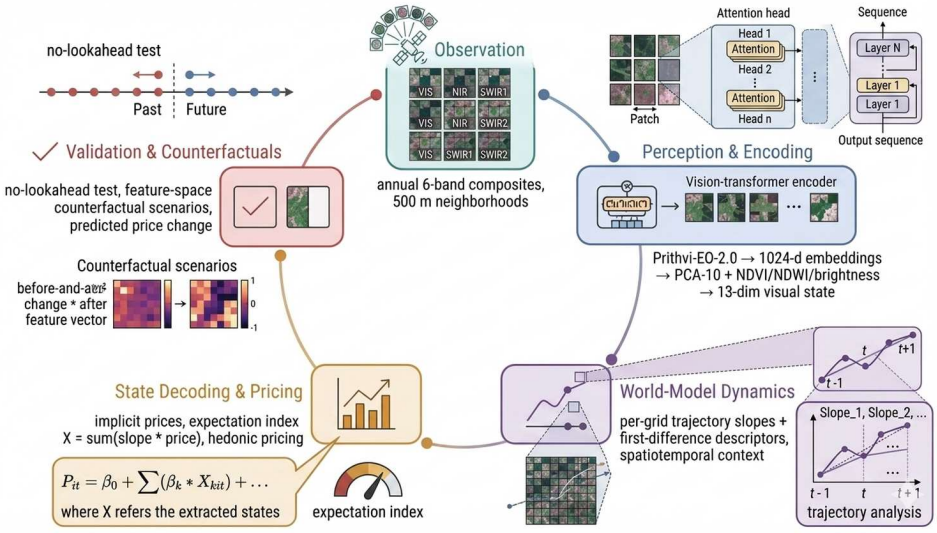



**Fig. 13** Research framework: a world-model view of visual-morphology capitalization. (1) Observation: annual six-band composites are assembled per township and year for 500 m neighborhoods. (2) Perception and encoding: the Prithvi-EO-2.0-300M encoder produces 1,024-dimensional patch embeddings that are projected by PCA onto ten components, which together with NDVI, NDWI and brightness form a 13-dimensional visual state. (3) Dynamics: each grid cell is treated as a small dynamical system whose per-dimension linear trajectories and first-difference descriptors constitute the spatiotemporal context. (4) State decoding and pricing: implicit prices and the expected-visual-change index are derived and enter the hedonic estimation. (5) Validation and counterfactuals: no-lookahead expectation tests and feature-space counterfactual policy simulations complete the framework.

seasonal-semantic structure of the land surface without ever observing housing transactions in the Taiwan region. Three properties of the encoder matter here. First, it transfers without fine-tuning: the encoder is kept frozen during estimation, so the visual state is produced by a pre-trained representation rather than by a task-specific network that could inadvertently absorb price information, guarding against the circularity that would arise if the encoder were optimized on the outcome variable. Second, it compresses information while retaining semantics: projecting the 1,024-dimensional embedding onto ten principal components (Eq. 6) yields a 13-dimensional state that keeps the deep features' incremental explanatory power (see Results) while remaining compact enough to be combined with interpretable handcrafted morphology and to be estimated across the 685,858 grid cells. Third, it temporalizes a static representation: the foundation model alone provides only a snapshot of the built environment, and the contribution of our world-model framing is to embed this deep state into a dynamical system, estimating per-grid trajectories (Eq. 17) and first-difference descriptors so that Prithvi's perceptual output becomes the state variable of an evolving neighborhood system. This design closes a perception–dynamics–decision loop that links the physical environment to economic outcomes and, by keeping the intermediate state explicit, retains the interpretability that pure deep-regression approaches sacrifice.

To test the expectation channel, we weight each grid's slope by the cross-sectional implicit price of the corresponding deep feature ($\hat{\gamma}_k$ from M_all) to form the expected-visual-change index,

$$X_{gi} = \sum_{k=1}^{13} \hat{s}_{g(i)k}\, \hat{\gamma}_k, \tag{18}$$

and include $X$ in the handcrafted hedonic model; a positive coefficient would indicate that prices capitalize anticipated visual change. The main specification uses full-period slopes, whereas a no-lookahead robustness specification estimates the slopes only from 2012–2019 imagery and retains only transactions from 2020 onward (Table 13).

To validate the predictive content of the dynamics module, we also evaluate it out of sample. For each grid cell with at least eight annual observations we estimate the per-dimension trajectory on the first 60% of the sample years and forecast the visual state at one- and three-year horizons from the last training year, using no test-period information (no lookahead). The world-model forecast $\hat{F}^{\mathrm{wm}}_{gk,t+h}$ extrapolates the training trajectory and is compared with two naive benchmarks: persistence (a random-walk forecast equal to the last training value) and the training-period mean. Features are $z$-standardized so that errors are comparable across the thirteen dimensions. Forecast skill relative to benchmark $b$ is

$$S_b = 1 - \frac{\sum_{gk,t}\left(\hat{F}^{\mathrm{wm}}_{gk,t+h} - F_{gk,t+h}\right)^2}{\sum_{gk,t}\left(\hat{F}^{b}_{gk,t+h} - F_{gk,t+h}\right)^2}, \tag{19}$$

together with the directional accuracy of the predicted one-year change in the visual state (Table 14 and Fig. 7).

Finally, feature-space counterfactual policy simulations (see Results) intervene on the handcrafted morphology features. For a scenario that changes feature $k$ by $\pm 1$ standard deviations among affected grids $T$, the predicted log-price change of an affected transaction is

$$\Delta\hat{p}_i = \sum_{k \in S} \hat{\beta}_k\, \sigma_k\, s_k\, \mathbb{1}\{i \in T\}, \tag{20}$$

where $\hat{\beta}_k$ and $\sigma_k$ are the M2 coefficient and the full-sample standard deviation of feature $k$, $s_k \in \{+1, -1\}$ is the intervention sign, and the percentage effect is $\exp(\Delta\hat{p}_i) - 1$. Confidence intervals are computed from the coefficient standard errors, ignoring feature co-variation (Table 16).

**Acknowledgements.** The authors thank the Ministry of the Interior of the Taiwan region for making the real-estate "actual registration" transaction records and township-level income statistics publicly available, and the European Space Agency's Copernicus programme for the Sentinel-2 imagery used in this study.

## Declarations

- **Funding:** The authors declare that no specific grant from any public, commercial, or not-for-profit funding agency was received for this research.
- **Competing interests:** The authors declare no competing interests.

- **Ethics approval and consent to participate:** Not applicable. This study uses publicly available, de-identified real-estate transaction records, satellite imagery and official township-level statistics; no human participants, personal data or animal subjects were involved.
- **Consent for publication:** Not applicable.
- **Data availability:** The real-estate "actual registration" transaction records of the Taiwan region are publicly available from the Ministry of the Interior (https://lvr.land.moi.gov.tw/). Sentinel-2 imagery is publicly available from the Copernicus Open Access Hub. Township-level income statistics are published by the Ministry of the Interior. Derived grid-level features and regression results are available from the corresponding author upon reasonable request.
- **Materials availability:** Not applicable.
- **Code availability:** Analysis code for data processing, feature extraction and estimation is available from the corresponding author upon reasonable request.
- **Author contributions:** J.P.-J.T. curated the data, processed the satellite imagery, extracted the visual-morphology features, ran the hedonic and world-model analyses, produced the figures and tables, and wrote the original draft. K.-K.T. conceptualized the study, designed the world-model framework, supervised the computational implementation, interpreted the results, and reviewed and edited the manuscript. R.F.-Y.L. designed the hedonic econometric specifications, led the heterogeneity and policy analysis, contributed the economic interpretation, and reviewed and edited the manuscript. All authors approved the final manuscript.

# References


[1] Rosen, S. Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy* **82**, 34–55 (1974).

[2] Malpezzi, S. Hedonic price models: A selective review of theory and application. In Housing Economics and Public Policy (eds O'Sullivan, T. & Gibb, K.), 43–70 (Blackwell Science, 2003).

[3] Sirmans, G. S., Macpherson, D. A. & Zietz, E. N. The composition of hedonic pricing models. *Journal of Real Estate Literature* **13**, 1–44 (2005).

[4] Naik, N., Kominers, S. D., Raskar, R., Glaeser, E. L. & Haggerty, C. A. Computer vision uncovers predictors of physical urban change. *Proceedings of the National Academy of Sciences* **114**, 7571–7576 (2017).

[5] Glaeser, E. L., Kominers, S. D., Luca, M. & Naik, N. Big data and big cities: The promises and limitations of improved measures of urban life. *Economic Inquiry* **56**, 114–137 (2018).

[6] Kong, F., Yin, H. & Nakagoshi, N. Using GIS and landscape metrics in the hedonic price modeling of the amenity value of urban green space: A case study in Jinan city, China. *Landscape and Urban Planning* **79**, 240–252 (2007).

[7] Dubé, J. & Legros, D. Spatial econometrics and the hedonic pricing model: What about the temporal dimension? *Journal of Property Research* **31**, 24–49 (2014).

[8] Law, S., Seresinhe, C. I., Shen, Y. & Gutmann, M. Street-frontage-net: urban image classification using deep convolutional neural networks. *International Journal of Geographical Information Science* **34**, 681–707 (2020).

[9] Ye, Y. *et al.* Measuring daily accessed street greenery: A human-scale approach for informing better urban planning practices. *Landscape and Urban Planning* **191**, 103434 (2019).

[10] Gebru, T. *et al.* Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. *Proceedings of the National Academy of Sciences* **114**, 13108–13113 (2017).

[11] Zhang, F. *et al.* Measuring human perceptions of a large-scale urban region using machine learning. *Landscape and Urban Planning* **180**, 148–160 (2018).

[12] Henderson, J. V., Storeygard, A. & Weil, D. N. Measuring economic growth from outer space. *American Economic Review* **102**, 994–1028 (2012).

[13] Donaldson, D. & Storeygard, A. The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives* **30**, 171–198 (2016).

[14] Jean, N. *et al.* Combining satellite imagery and machine learning to predict poverty. *Science* **353**, 790–794 (2016).

[15] Ha, D. & Schmidhuber, J. *Recurrent world models facilitate policy evolution.* Advances in Neural Information Processing Systems (NeurIPS), Vol. 31, 2455–2467 (Curran Associates, Inc., 2018).

[16] LeCun, Y. A path towards autonomous machine intelligence. https://openreview.net/forum?id=BZ5a1r-kVsf (2022). Position paper, version 0.9.2.

[17] Hafner, D., Pasukonis, J., Ba, J. & Lillicrap, T. Mastering diverse control tasks through world models. *Nature* **640**, 647–653 (2025).

[18] Dosovitskiy, A. *et al. An image is worth 16x16 words: Transformers for image recognition at scale.* International Conference on Learning Representations (ICLR) (OpenReview.net, 2021).

[19] Oquab, M. *et al.* Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research* (2024).

[20] Radford, A. *et al. Learning transferable visual models from natural language supervision.* Proceedings of the 38th International Conference on Machine Learning (ICML), Vol. 139, 8748–8763 (PMLR, 2021).

[21] Jakubik, J. *et al.* Foundation models for generalist geospatial artificial intelligence. *IEEE Transactions on Big Data* **10**, 613–630 (2024).

[22] Szwarcman, D. *et al.* Prithvi-EO-2.0: A versatile multi-temporal foundation model for Earth observation applications. https://arxiv.org/abs/2412.02732 (2025). ArXiv:2412.02732.

[23] Drusch, M. *et al.* Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sensing of Environment* **120**, 25–36 (2012).

[24] Lacoste, A. *et al. Geo-bench: Toward foundation models for earth monitoring.* Advances in Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks (Curran Associates, Inc., 2023).

[25] OpenStreetMap contributors. Openstreetmap. https://www.openstreetmap.org (2026). Accessed August 2026.

[26] Cong, Y. *et al. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery.* Advances in Neural Information Processing Systems (NeurIPS), Vol. 35, 197–211 (Curran Associates, Inc., 2022).

[27] Reed, C. J. *et al. Scale-mae: A scale-aware masked autoencoder for multi-scale geospatial representation learning.* IEEE/CVF International Conference on Computer Vision (ICCV) (IEEE/CVF, 2023).

[28] Sun, X. *et al.* Ringmo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing* **61**, 5620821 (2023).

[29] Guo, X. *et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery.* IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 27662–27673 (IEEE, 2024).

[30] Hong, D. *et al.* Spectralgpt: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[31] Fuller, A., Millard, K. & Green, J. R. *Croma: Remote sensing representations with contrastive radar-optical masked autoencoders.* Advances in Neural Information Processing Systems (NeurIPS), Vol. 36 (Curran Associates, Inc., 2023).