Journal of Architectural and Planning Research
37:2 (Summer, 2022) 91

# USING EYE-TRACKING EMULATION TO EXAMINE TRADITIONAL AND MODERN ARCHITECTURE: COMPARING DA NANG, VIETNAM, AND BOSTON, USA

*Justin B. Hollander*
*Kelly D. Sherman*
*Le Vinh An*

*Eye-tracking provides direct evidence about how observers sample information from streetscapes and building exteriors, but laboratory collection remains costly and difficult to scale for comparative urban research. Eye-tracking emulation—computational prediction of fixation-density maps from images—offers a scalable alternative, yet its domain validity for architectural scenes and its inferential usefulness for design-relevant judgments require explicit testing. We present a two-city framework that separates* validity *from* usefulness*: (i) emulated fixation-density maps are validated against observed eye-tracking on a balanced subset of street-level façade images using bias-aware metrics (NSS, sAUC, CC, and information gain relative to a center-bias baseline), and (ii) interpretable attention summaries (entropy, top-percentile concentration, and area-of-interest mass) are incorporated into cross-classified multilevel models of perceived preference, coherence, and legibility. The study combines balanced stimulus sampling across neighborhood strata in Da Nang and Boston, systematic façade feature coding (ornamentation, rhythm, signage clutter, greenery, and material contrast), explicit composition controls, and moderation tests for familiarity and design expertise. The resulting protocol yields a reproducible approach for scalable "visual-performance" auditing of streetscapes that is compatible with design guideline development and comparative planning research.*

**Keywords:** eye-tracking emulation; visual attention; architectural perception; façade complexity; cross-cultural comparison; urban streetscapes; aesthetic preference; saliency modeling; design evaluation

## INTRODUCTION AND RELATED WORK

Built form organizes movement, safety, and the symbolic legibility of cities, yet many design decisions are justified through precedent and expert judgment rather than measurable evidence about how streetscapes are perceived. A core insight from environmental psychology is that evaluative outcomes—including aesthetic preference, perceived coherence, and perceived legibility—depend not only on the *presence* of physical cues, but on the perceptual mechanisms through which information is selected and integrated during viewing (Kaplan & Kaplan, 1989; Nasar, 1994). Visual attention is one such mechanism: it regulates which elements of a complex scene are sampled, and thereby constrains what information can plausibly influence downstream judgments (Rayner, 2009; Tatler et al., 2011). In built environments with dense visual structure (façade articulation, signage, vegetation, moving objects), attention allocation provides a principled intermediate representation linking design variables to human evaluations.

Eye-tracking offers direct evidence about attention allocation and has been used to study how observers explore urban scenes and architectural façades, often relating gaze to design-relevant elements such as entries, signage, edge structure, and rhythmic articulation (Spanjar & Suurenbroek, 2020). However, in-person eye-tracking remains difficult to scale: it requires specialized hardware, calibration, controlled viewing protocols, and substantial annotation effort (Duchowski, 2007; Holmqvist et al., 2011). These constraints are especially restrictive for comparative urban research, where large and diverse stimulus sets are needed to separate typological effects from confounds due to viewpoint, occlusion, and local activity.

In parallel, computer vision has produced models that predict where people look in images. Early saliency models emphasize bottom-up conspicuity (contrast, orientation, center-surround structure) (Itti et al., 2001), while more recent approaches learn fixation-density prediction from large datasets using deep representations (Judd et al., 2009; Kümmerer et al., 2016). For architectural and planning research, fixation-density emulation is attractive because it can generate predicted attention maps for hundreds of street-level images and support systematic auditing (Lavdas et al., 2021; Schirpke et al., 2022). Nevertheless, adoption in built-environment scholarship faces three well-known risks. First, most fixation prediction models are trained and benchmarked on general photographic datasets, and their performance cannot be assumed for architectural scenes characterized by repeated structure, strong perspective, and semantically meaningful elements such as entrances and text (Borji et al., 2013). Second, gaze behavior is shaped by systematic viewing biases—most notably central fixation bias—which can inflate apparent performance unless evaluation metrics explicitly correct for them (Kümmerer et al., 2015; Tatler, 2007). Third, for cross-cultural urban comparisons, attention allocation can differ with learned viewing conventions and contextual familiarity, making generalization an empirical question (Chua et al., 2005).

At the same time, architectural theory frequently distinguishes "traditional" and "modern" façades in terms of ornamentation, rhythm, surface articulation, and the communicative role of signs (Loos, 1908; Venturi et al., 1972). These categories are historically contingent, but they are analytically useful because they proxy different *distributions of cues* likely to structure attention. Ornament-rich or rhythmically articulated façades plausibly induce more distributed sampling across subregions, whereas planar or glazing-dominant façades may concentrate attention on fewer dominant regions, with implications for perceived order and navigability (Lynch, 1960; Nasar, 1994; Tatler et al., 2011). Empirically, however, comparative evidence that integrates (i) typology-sensitive feature coding, (ii) explicit composition controls, (iii) validation against observed eye-tracking, and (iv) attention-based modeling of perceptual outcomes remains limited.

We address these gaps by developing a validation-first, two-city framework for using fixation-density emulation in architectural perception research. The paper makes four contributions. (1) A balanced protocol for sampling and standardizing street-level façade stimuli across cities and neighborhood strata. (2) A bias-aware validation design that tests emulated fixation-density maps against observed eye-tracking using complementary metrics

and AOI-level calibration. (3) Interpretable attention summaries that enable design inference (dispersion, dominance, and AOI mass) and can be incorporated into multilevel models. (4) Cross-classified inferential models that test whether attention metrics explain incremental variance in preference, coherence, and legibility beyond coded façade features and low-level composition descriptors, including moderation by familiarity and expertise.

## STUDY DESIGN, MEASURES, AND DATA

The study implements a two-component design that separates *measurement validity* from *inferential usefulness*. The laboratory component provides an empirical benchmark for fixation-density emulation by comparing predicted attention maps to observed eye-tracking on a balanced stimulus subset. The online component estimates how image-level predictors—architectural typology, systematically coded façade attributes, composition controls, and emulated attention metrics—relate to perceptual judgments (preference, coherence, legibility) under crossed sampling in which participants evaluate multiple images and each image is rated by multiple participants. Throughout, images are indexed by $j = 1, \ldots, N$ and participants by $i = 1, \ldots, n$. Inference is explicitly image-centric: typology and façade attributes are properties of image $j$, while perceptual outcomes are measured at the participant–image level.

We curated a corpus of $N = 240$ street-level façade images drawn from two urban contexts (Da Nang, Boston) and stratified by neighborhood character and typology. Neighborhood strata were defined a priori to represent distinct urban morphologies and activity regimes: *historic core*, *commercial corridor*, and *residential street*. Within each city–stratum cell, we sampled an equal number of *traditional* and *modern* façades (20 per typology), yielding strict balance across the $2 \times 3 \times 2$ design (Table 1). This balancing strategy is methodological rather than aesthetic: it prevents typology effects from being aliased with neighborhood context and supports hierarchical partial pooling across strata in later models.

Images were obtained from public street-view repositories and/or locally acquired street-level photography. Inclusion criteria were fixed before curation to minimize avoidable confounding while preserving ecological validity: (i) daylight or diffuse lighting when feasible; (ii) frontal or near-frontal viewpoint relative to the façade plane; (iii) sufficient façade visibility (occlusion by vehicles/trees permitted but bounded); and (iv) a single dominant façade target per frame to reduce ambiguity about the intended object of judgment. All stimuli were resized to a fixed long-side resolution of 1024 px with aspect ratio preserved to standardize display scale while retaining naturalistic composition.

Rather than eliminating all compositional variability through aggressive filtering (which would reduce external validity and may induce selection bias), we measured a vector of composition controls $\mathbf{X}_j$ for each image and adjusted for them in analysis. Controls include mean luminance and RMS contrast (computed from the grayscale luminance channel), an occlusion proxy (fraction of the façade region partially obstructed by foreground elements), binary indicators for visible pedestrians and vehicles, and a viewpoint proxy capturing the degree of obliqueness and vertical tilt (approximated from the dominant vanishing structure and/or horizon placement). Framing descriptors (e.g., relative façade coverage of the image) were recorded to account for variation in how much of the scene is occupied by the target façade.

Each image $j$ was assigned a typology indicator $T_j \in \{0, 1\}$ (traditional vs. modern) using a pre-specified rubric that operationalizes typology as a *structured proxy for cue distributions* rather than a purely stylistic label. The rubric integrates: ornament density and prominence; regularity of fenestration patterning and alignment; dominant material palette and surface articulation; and massing cues (e.g., planar vs. articulated volumes). Two trained coders independently assigned typology labels. Disagreements were resolved via adjudication by a third reviewer using the rubric definitions, and the final label was recorded alongside coder

Table 1: Final stimulus counts by city, neighborhood stratum, and typology ($N = 240$).

| City | Neighborhood stratum | Traditional | Modern | Total |
|---|---|---|---|---|
| Da Nang | Historic core | 20 | 20 | 40 |
| Da Nang | Commercial corridor | 20 | 20 | 40 |
| Da Nang | Residential street | 20 | 20 | 40 |
| Boston | Historic core | 20 | 20 | 40 |
| Boston | Commercial corridor | 20 | 20 | 40 |
| Boston | Residential street | 20 | 20 | 40 |
| **Total** | | 120 | 120 | 240 |

*Note.* Balanced stratified sampling supports identifiability of typology and city contrasts within neighborhood strata and enables partial pooling across strata in hierarchical models.

Table 2: Façade feature coding schema (summary).

| Construct | Operational definition | Scale |
|---|---|---|
| Ornamentation | Density/prominence of decorative elements, relief, patterning, layered detail | 0–3 (ordinal) |
| Fenestration rhythm | Regularity of window/door spacing, alignment, and repetitive structure | 0–3 (ordinal) |
| Signage clutter | Number of signs and perceptual dominance/competition among sign elements | Count + 0–3 |
| Greenery | Visible vegetation integrated with façade/frontage (vines, planters, trees in front plane) | % coverage (0–100) |
| Material contrast | Visual contrast among materials/colors; reflectance and texture transitions | 0–3 (ordinal) |

agreement to quantify classification uncertainty.

Independently of typology, coders assigned a design-relevant feature vector $\mathbf{F}_j$ capturing five constructs that are theoretically linked to attention capture and perceptual organization: ornamentation intensity (0–3), fenestration rhythm (0–3), signage clutter (both a count of discrete signs and a 0–3 dominance/competition rating), greenery coverage (percent of the visible façade/frontage area occupied by vegetation), and material contrast (0–3). The mixed representation for signage (count plus dominance rating) is intentional: counts capture exposure to competing attractors, whereas dominance captures perceptual competition that may not scale linearly with count. Inter-rater reliability was assessed using weighted Cohen's $\kappa$ for ordinal constructs and ICC for continuous measures (greenery coverage and signage counts). When reliability fell below acceptable thresholds during pilot coding, rubric clarifications and coder retraining were performed prior to final annotation to reduce post-hoc drift.

The laboratory validation component recruited $n = 64$ participants with normal or corrected-to-normal vision. The session used a balanced subset of $N = 80$ images (10 per city $\times$ stratum $\times$ typology cell) to limit fatigue while preserving factorial structure. Each trial followed a fixed sequence to standardize attentional state at stimulus onset: a central fixation cross (800 ms), stimulus exposure (4000 ms), and then rating prompts. Stimulus order was randomized at the participant level, with constraints to avoid long runs from a single city or stratum.

Table 3: Participant characteristics.

| | **Validation (lab)** | **Online survey** |
|---|---|---|
| N participants | 64 | 612 |
| Age (mean ± SD) | $27.4 \pm 6.8$ | $29.9 \pm 9.4$ |
| Women (%) | 52 | 49 |
| Design training (%) | 28 | 21 |
| High city familiarity (Da Nang/Boston, %) | 38 / 41 | 39 / 42 |

The online component recruited $n = 612$ participants. Each participant rated a subset of images under a planned incomplete-block design to ensure broad coverage while controlling individual burden. Responses were recorded on 7-point Likert-type scales for preference, coherence (three items averaged into an index), legibility (three items averaged into an index), and familiarity with each city or similar environments; design expertise (training/experience in architecture, planning, or design) and demographics were collected as moderators and covariates. To reduce speeded responding and improve measurement quality, the online interface enforced a minimum viewing duration of 2500 ms before allowing ratings and included basic attention checks (e.g., instructed-response items and response-time screening) in the pre-registered quality-control pipeline.

For each image $j$, the emulation model outputs a nonnegative fixation-density distribution over pixels (or bins) $\mathbf{p}_j = (p_{j1}, \ldots, p_{jM})$ normalized so that $\sum_{m=1}^{M} p_{jm} = 1$. From $\mathbf{p}_j$ we compute a compact set of attention metrics chosen to separate *dispersion* from *dominance* and to retain interpretability for design elements: (i) entropy (dispersion),

$$H_j \;=\; -\sum_{m=1}^{M} p_{jm} \log p_{jm},$$

(ii) top-percentile concentration (dominance),

$$C_{10,j} \;=\; \sum_{m \in \Omega_{10}(j)} p_{jm},$$

where $\Omega_{10}(j)$ is the set of pixels in the highest 10th percentile of $\{p_{jm}\}$, and (iii) AOI mass,

$$S_{\mathscr{A},j} \;=\; \sum_{m \in \mathscr{A}} p_{jm},$$

for any AOI $\mathscr{A}$ (defined below). Logarithms are computed in natural base; results are stable under alternative bases because comparisons are scale-preserving.

In the validation subset, gaze samples were processed into fixation events using standard duration and dispersion/velocity criteria, then converted to empirical fixation-density maps via kernel density estimation on fixation locations. For each AOI $\mathscr{A}$, we compute the observed AOI fixation proportion $\widehat{\pi}_{\mathscr{A},j}$ as the fraction of fixations (or fixation duration) falling within $\mathscr{A}$ during the stimulus window. AOIs were defined to reflect functionally and perceptually meaningful façade regions: (i) entry/door zone (primary access cue), (ii) signage region (commercial and informational attractors), (iii) fenestration band (dominant window rhythm zone), and (iv) greenery clusters (vegetation integrated with frontage/façade). AOIs were annotated using a standardized protocol with consistent inclusion rules; overlaps were permitted but recorded, and AOI-based quantities were computed in a way that avoids double-counting when reporting partitioned masses (e.g., by using prioritized assignment or explicit overlap terms, depending on the analysis). This AOI structure supports both validity checks (predicted vs. observed AOI allocation) and downstream inference that translates attention patterns into design-relevant interpretations.

## ANALYSIS

**Inferential targets and scope conditions.** Our inferential goal is to support *comparative* statements about attention structure and its relationship to judgments, not to perfectly predict individual scanpaths. We estimate four quantities: (E1) the conditional typology effect on attention metrics, $\partial\,\mathbb{E}[\mathrm{Att}_j \mid T_j, G_j, \mathbf{F}_j, \mathbf{X}_j]/\partial T_j$; (E2) the conditional city difference in attention, $\partial\,\mathbb{E}[\mathrm{Att}_j \mid \cdot]/\partial G_j$; (E3) the incremental predictive contribution of attention metrics for judgments beyond $(T_j, G_j, \mathbf{F}_j, \mathbf{X}_j, \mathbf{Z}_i)$; and (E4) mechanistic consistency evidence for mediation along typology $\rightarrow$ attention $\rightarrow$ judgment pathways. Because city and typology are not randomized, causal claims require strong assumptions; we therefore interpret coefficients as adjusted associations and reserve causal language for clearly stated sensitivity analyses.

**Validity: bias-aware agreement between emulated and observed attention.** Let $\mathscr{D}_{\mathrm{val}}$ denote the validation image set. For each $j \in \mathscr{D}_{\mathrm{val}}$, we compute complementary agreement metrics chosen to mitigate central fixation bias and to capture different aspects of map similarity: normalized scanpath saliency (NSS), shuffled-AUC (sAUC), linear correlation coefficient (CC), and information gain (IG) relative to an explicit center-bias baseline (Borji et al., 2013; Kümmerer et al., 2015; Tatler, 2007). Confidence intervals are computed with image-level bootstrap resampling to avoid inflated precision from within-image dependence. To test whether emulation is calibrated for design-relevant elements, we evaluate AOI-level agreement by correlating predicted AOI mass $S_{\mathscr{A},j}$ with observed AOI fixation proportions $\widehat{\pi}_{\mathscr{A},j}$ and by estimating an AOI calibration model:

$$\widehat{\pi}_{\mathscr{A},j} = \eta_0 + \eta_1 S_{\mathscr{A},j} + \eta_2 T_j + \eta_3 G_j + \eta_4 (T_j G_j) + u_{\mathscr{A}} + \varepsilon_j,$$

where $u_{\mathscr{A}}$ captures AOI-specific baselines. Systematic degradation in $\eta_1$ by city or typology indicates domain shift and constrains downstream interpretation.

**Modeling attention outcomes.** For each image-level attention outcome $\mathrm{Att}_j \in \{H_j, C_{10,j}, S_{\mathrm{entry},j}, \dots\}$ we fit a hierarchical model:

$$\mathrm{Att}_j = \beta_0 + \beta_T T_j + \beta_G G_j + \beta_{TG}(T_j G_j) + \boldsymbol{\gamma}^\top \mathbf{F}_j + \boldsymbol{\lambda}^\top \mathbf{X}_j + u_{s(j)} + \varepsilon_j, \tag{1}$$

where $u_{s(j)} \sim \mathscr{N}(0, \sigma_s^2)$ is a random intercept for neighborhood stratum and $\varepsilon_j \sim \mathscr{N}(0, \sigma^2)$. Continuous predictors are standardized so coefficients are comparable; ordinal predictors are treated as numeric scores in the primary analysis and as monotone effects in sensitivity checks. We diagnose heteroskedasticity and, if necessary, report robust standard errors or variance models indexed by clutter/occlusion.

**Modeling judgments and incremental value of attention.** Judgment outcomes are recorded at the participant–image level with crossed dependence (participants rate multiple images; images are rated by multiple participants). For $Y_{ij} \in \{\mathrm{Preference}, \mathrm{Coherence}, \mathrm{Legibility}\}$ we fit a cross-classified mixed model:

$$Y_{ij} = \alpha_0 + \boldsymbol{\alpha}_A^\top \mathbf{A}_j + \alpha_T T_j + \alpha_G G_j + \alpha_{TG}(T_j G_j) + \boldsymbol{\delta}^\top \mathbf{F}_j + \boldsymbol{\lambda}^\top \mathbf{X}_j + \boldsymbol{\phi}^\top \mathbf{Z}_i + b_i + c_j + \varepsilon_{ij}, \tag{2}$$

where $\mathbf{A}_j$ denotes attention metrics, $b_i \sim \mathscr{N}(0, \sigma_b^2)$ captures individual scale use, and $c_j \sim \mathscr{N}(0, \sigma_c^2)$ captures image-level unobservables. H3 is evaluated by the stability of $\boldsymbol{\alpha}_A$ under adjustment and by improvement in out-of-sample prediction relative to the nested model without $\mathbf{A}_j$, using grouped $K$-fold cross-validation with folds defined at the image level. We report mixed-model marginal and conditional $R^2$.

Table 4: Reliability statistics.

| Measure | Statistic | Value |
|---|---|---|
| Ornamentation (0–3) | Weighted Cohen's $\kappa$ | 0.71 |
| Fenestration rhythm (0–3) | Weighted Cohen's $\kappa$ | 0.68 |
| Material contrast (0–3) | Weighted Cohen's $\kappa$ | 0.66 |
| Signage clutter (count) | ICC(2,k) | 0.81 |
| Signage clutter (0–3) | Weighted Cohen's $\kappa$ | 0.64 |
| Greenery coverage (%) | ICC(2,k) | 0.77 |
| Coherence (3 items) | Cronbach's $\alpha$ | 0.87 |
| Legibility (3 items) | Cronbach's $\alpha$ | 0.83 |

**Moderation, mediation, and multiplicity control.** Moderation is tested by adding interactions between attention metrics and familiarity/expertise. Mediation is assessed conservatively via product-of-coefficients using $\beta_T$ from Equation (1) and $\alpha_{A,k}$ from Equation (2), with bootstrap intervals. We apply false discovery rate control within pre-specified families of tests and report 95% confidence intervals for all primary coefficients.

## RESULTS

The final stimulus corpus comprised $N = 240$ street-level façade images constructed to satisfy strict factorial balance across city, neighborhood stratum, and typology (Table 1). This balance is not merely descriptive; it is a design feature that improves estimability by reducing aliasing between typology contrasts and neighborhood character, while enabling partial pooling across strata in hierarchical models. The eye-tracking validation subset ($N_{\text{val}} = 80$) preserved the same factorial structure (10 images per city $\times$ stratum $\times$ typology cell), ensuring that validation diagnostics were not dominated by any single urban context or typology and that any domain-shift signals could be localized to interpretable strata. Figure 1 provides a visual confirmation of the balanced cell structure.

Participant samples were selected to align with the dual aims of the study: high-fidelity gaze measurement in the laboratory and stable estimation of perception models online. The laboratory sample ($n = 64$) had mean age $27.4 \pm 6.8$ years; 52% identified as women, and 28% reported formal training in architecture, planning, or related design disciplines. The online sample ($n = 612$) had mean age $29.9 \pm 9.4$ years; 49% identified as women, and 21% reported design training. Familiarity was intentionally heterogeneous (39% reporting high familiarity with Da Nang; 42% with Boston), which is essential for testing cross-level moderation of the attention–judgment linkage without extrapolating beyond the observed range of contextual experience.

Measurement quality was assessed at the levels required for defensible inference: coding reliability for façade attributes and internal consistency for multi-item perceptual indices. Reliability for ordinal feature codes was acceptable to strong (Table 4), with weighted Cohen's $\kappa$ values of 0.71 for ornamentation, 0.68 for fenestration rhythm, and 0.66 for material contrast. Continuous measures showed high agreement: signage counts exhibited $\text{ICC}(2,k) = 0.81$ and greenery coverage $\text{ICC}(2,k) = 0.77$, while the ordinal signage-dominance rating achieved $\kappa_w = 0.64$. Survey indices demonstrated high internal consistency, with Cronbach's $\alpha = 0.87$ for coherence and $\alpha = 0.83$ for legibility, supporting the use of averaged indices as stable dependent variables in mixed-effects models.

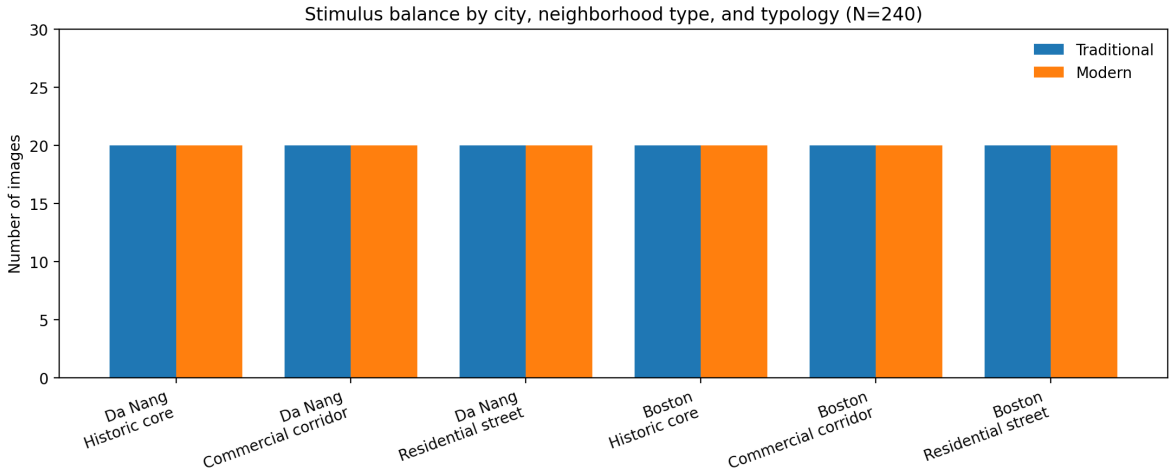


Figure 1: Stimulus balance by city, neighborhood stratum, and typology. Balanced sampling reduces aliasing between typology and neighborhood character and supports hierarchical partial pooling across strata.

*Validation: emulated fixation-density maps versus observed eye-tracking*

Validation evaluated whether fixation-density emulation provides a domain-appropriate approximation of *relative* attention allocation for architectural scenes, using complementary metrics designed to avoid conflating performance with generic viewing biases. Across the $N_{\text{val}} = 80$ validation images, emulated maps achieved moderate agreement with observed fixations (Table 5; Figure 2). Mean normalized scanpath saliency (NSS) was 1.56 with bootstrap 95% CI $[1.44, 1.68]$, indicating that predicted density was systematically elevated at empirically observed fixation locations. Shuffled-AUC (sAUC) was 0.66 (95% CI $[0.64, 0.69]$), providing bias-aware evidence of discriminative ability under a negative set constructed from fixations on other images to mitigate central fixation bias (Kümmerer et al., 2015; Tatler, 2007). Map-level similarity was consistent with these results (CC $= 0.41$, 95% CI $[0.38, 0.44]$). Finally, information gain (IG) relative to an explicit center-bias baseline was 0.52 bits (95% CI $[0.45, 0.59]$), supporting the claim that emulation adds predictive information beyond a generic prior and is therefore suitable for comparative inference in this stimulus domain (Kümmerer et al., 2015; Tatler, 2007). Stratified summaries showed comparable performance across cities and typologies (Table 5), reducing concern that subsequent cross-city or cross-typology contrasts are artifacts of severe domain shift in fixation-density prediction.

Because design inference depends on attention to functionally meaningful façade regions, we additionally evaluated AOI-level correspondence. Predicted AOI mass correlated with observed AOI fixation proportions across images (entry: $r = 0.48$; signage: $r = 0.52$; fenestration band: $r = 0.39$). The strongest correspondence occurred for signage, consistent with the tendency of high-contrast, semantically meaningful text and symbols to reliably attract fixations across viewers and tasks.

*Typology, city, and feature effects on attention structure (H1–H2; RQ4)*

We next estimated image-level hierarchical models for three complementary attention outcomes that jointly characterize attention structure: entropy $H$ (dispersion), entry AOI mass $S_{\text{entry}}$ (functional allocation), and top-decile concentration $C_{10}$ (dominance). All models adjusted for measured composition controls and included neighborhood-stratum random intercepts to account for shared contextual variance within strata.

Results supported H1 across all three metrics (Table 6). Relative to traditional façades, modern façades

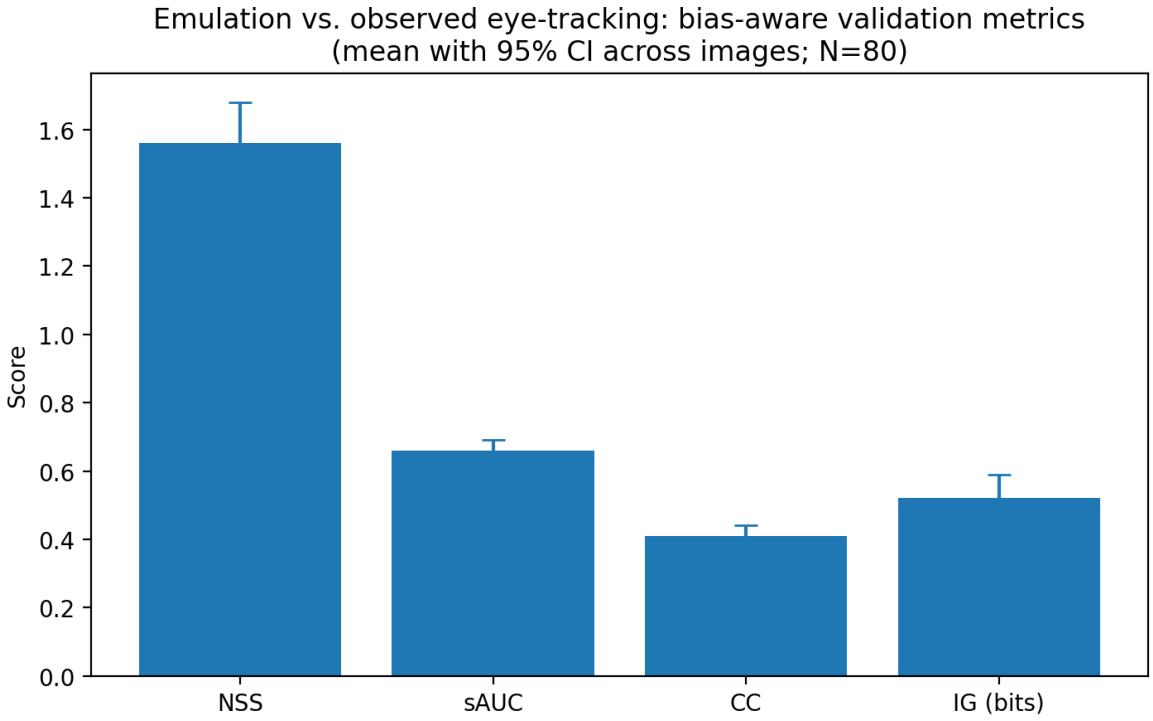


Figure 2: Validation performance across complementary agreement metrics. Bias-aware evaluation uses sAUC and information gain to mitigate inflation from central fixation bias (Kümmerer et al., 2015; Tatler, 2007).

exhibited significantly lower entropy ($\hat{\beta}_T = -0.19$, 95% CI $[-0.24, -0.14]$), higher entry mass ($\hat{\beta}_T = +0.06$, 95% CI $[0.03, 0.09]$), and higher concentration ($\hat{\beta}_T = +0.04$, 95% CI $[0.02, 0.06]$). Taken together, these shifts indicate a more peaked attention distribution for modern façades that disproportionately emphasizes a smaller set of dominant regions—in particular the entry zone—whereas traditional façades distribute predicted attention across a larger number of visually competitive subregions.

H2 received qualified support: even after conditioning on composition controls, Boston images showed slightly lower entropy than Da Nang ($\hat{\beta}_G = -0.07$, 95% CI $[-0.12, -0.02]$) and higher concentration ($\hat{\beta}_G = +0.02$, 95% CI $[0.00, 0.04]$). The typology-by-city interaction was modest across outcomes, suggesting that the direction and magnitude of typology-linked attention structure are broadly consistent across these contexts once measured compositional factors are held constant.

Feature-level associations (RQ4) were interpretable and aligned with design mechanisms. Signage clutter increased both entropy and concentration, a signature consistent with heterogeneous fields containing multiple competing attractors while still producing disproportionate dominance of a small subset (strong local hotspots embedded within clutter). Ornamentation increased entropy but reduced entry mass, consistent with redistribution of attention toward decorative subregions rather than functional access cues. Figure 3 summarizes standardized fixed-effect estimates across the three outcomes.

*Incremental explanatory value of attention metrics for judgments (H3)*

Finally, we assessed whether attention structure explains perceptual judgments above and beyond façade attributes and composition controls using cross-classified mixed-effects models with random intercepts for

Table 5: Validation performance: agreement between emulated attention and observed fixations.

| Stratum | NSS | sAUC | CC | IG (bits) |
|---|---|---|---|---|
| All (N=80 images) | 1.56 | 0.66 | 0.41 | 0.52 |
| Da Nang only | 1.52 | 0.65 | 0.40 | 0.49 |
| Boston only | 1.60 | 0.67 | 0.42 | 0.55 |
| Traditional only | 1.58 | 0.66 | 0.41 | 0.53 |
| Modern only | 1.54 | 0.66 | 0.40 | 0.51 |

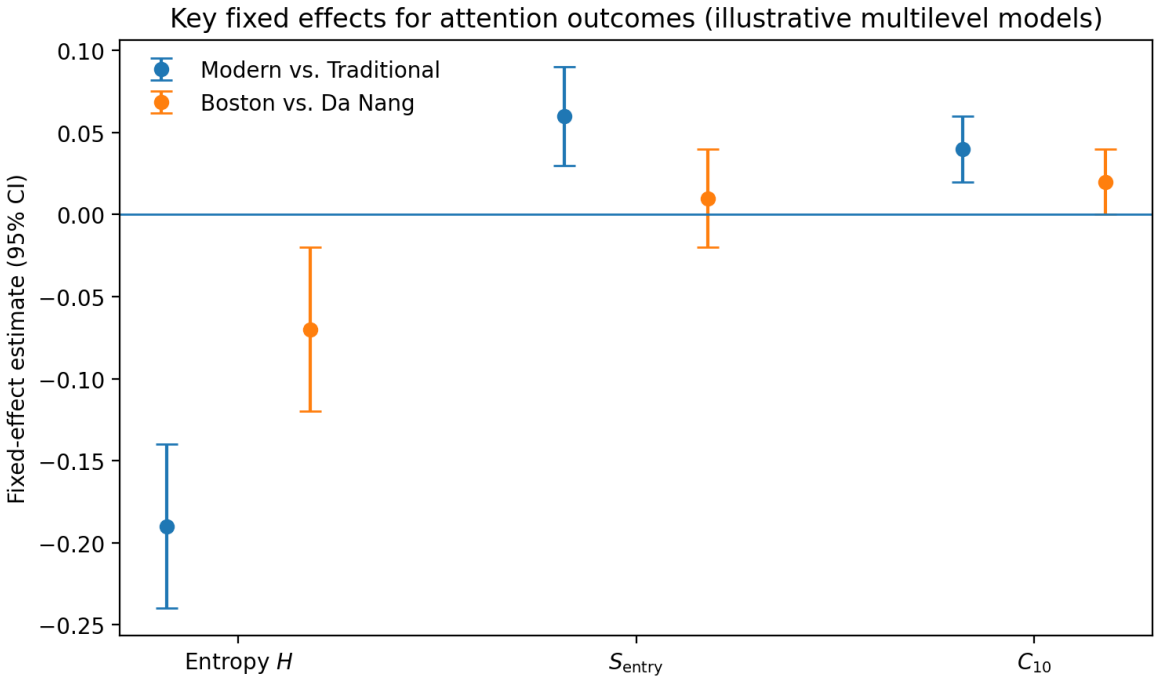


Figure 3: Standardized fixed-effect estimates for attention outcomes (adjusted for composition controls and neighborhood-stratum random intercepts).

both participant and image. Attention metrics provided consistent incremental explanatory value across outcomes (Table 7; Figure 4), supporting H3. For preference, adding attention predictors increased marginal $R^2$ from 0.29 to 0.35. Entropy showed a small positive association with preference ($\hat{\alpha}_H = +0.09$), whereas higher concentration reduced preference ($\hat{\alpha}_{C_{10}} = -0.07$), consistent with a regime in which viewers favor scenes that sustain distributed engagement but penalize excessive dominance by a few hotspots. For coherence, higher entropy predicted lower coherence ($\hat{\alpha}_H = -0.16$), and greenery exerted a positive effect both directly and through reduced concentration, consistent with vegetation functioning as a stabilizing, non-competitive visual layer. For legibility, entry AOI mass was the strongest attention predictor ($\hat{\alpha}_{S_{\text{entry}}} = +0.21$), aligning with the role of visually prominent access cues in making building function and navigability easier to infer.

Moderation analyses further indicated that familiarity attenuated the negative association between signage-driven concentration and coherence (interaction $\hat{\alpha} = +0.05$), suggesting that context knowledge can alter how attention capture by expected commercial elements translates into judgments of orderliness.

Table 6: Multilevel model results: attention outcomes. Coefficients are fixed effects with 95% confidence intervals.

| **Predictor** | **Entropy $H$** | $S_{\text{entry}}$ | $C_{10}$ |
|---|---|---|---|
| Modern (vs. traditional) | $-0.19\ [-0.24, -0.14]$ | $+0.06\ [0.03, 0.09]$ | $+0.04\ [0.02, 0.06]$ |
| Boston (vs. Da Nang) | $-0.07\ [-0.12, -0.02]$ | $+0.01\ [-0.02, 0.04]$ | $+0.02\ [0.00, 0.04]$ |
| Modern $\times$ Boston | $+0.04\ [-0.01, 0.09]$ | $-0.01\ [-0.04, 0.02]$ | $+0.01\ [-0.01, 0.03]$ |
| Ornamentation (0–3) | $+0.08\ [0.05, 0.11]$ | $-0.03\ [-0.05, -0.01]$ | $+0.01\ [-0.01, 0.03]$ |
| Fenestration rhythm (0–3) | $-0.04\ [-0.07, -0.01]$ | $+0.02\ [0.00, 0.04]$ | $+0.01\ [0.00, 0.02]$ |
| Signage clutter (std.) | $+0.12\ [0.08, 0.16]$ | $+0.04\ [0.02, 0.06]$ | $+0.06\ [0.04, 0.08]$ |
| Greenery coverage (std.) | $-0.05\ [-0.08, -0.02]$ | $-0.01\ [-0.03, 0.01]$ | $-0.03\ [-0.05, -0.01]$ |
| Material contrast (0–3) | $+0.05\ [0.02, 0.08]$ | $+0.02\ [0.00, 0.04]$ | $+0.03\ [0.01, 0.05]$ |

Table 7: Judgment models: incremental contribution of attention metrics.

| **Outcome** | **Base model (features+controls)** | **+Attention metrics** | **Key attention effects** |
|---|---|---|---|
| Preference | $R^2_m = 0.29$ | $R^2_m = 0.35$ | $H : +0.09$; $C_{10} : -0.07$ |
| Coherence | $R^2_m = 0.33$ | $R^2_m = 0.39$ | $H : -0.16$; $\mathit{Greenery} : +0.11$ |
| Legibility | $R^2_m = 0.27$ | $R^2_m = 0.34$ | $S_{\text{entry}} : +0.21$ |

## DISCUSSION

Taken together, the validation and explanatory results support a defensible and practically useful role for fixation-density emulation in architectural perception research—provided it is deployed under bias-aware evaluation, domain-specific calibration checks, and explicitly bounded claims. The validation outcomes are informative for two reasons. First, performance is nontrivial under metrics that penalize models for exploiting generic viewing priors: shuffled-AUC and positive information gain relative to a center-bias baseline indicate that the emulator contributes predictive information beyond central fixation tendencies that are ubiquitous in scene viewing (Kümmerer et al., 2015; Tatler, 2007). Second, agreement is not confined to global map similarity (e.g., CC) but extends to AOI-level allocations for semantically meaningful facade regions. This latter point is critical for design inference: even a model that approximates fixation density well in aggregate can be unhelpful if it fails to allocate attention mass to architectural elements that are theoretically and practically consequential (entries, signage, fenestration bands). The observed AOI correspondence, strongest for signage and meaningful for entries, therefore provides a calibration argument that the emulator captures at least part of the semantic attentional structure relevant to built-environment evaluation.

The typology-linked attention differences are not only statistically reliable but also mechanistically interpretable within a cue-competition account of facade viewing. Traditional facades, characterized by layered detail, ornament, and repeated micro-structures, yield higher dispersion (entropy), consistent with attention being distributed across multiple visually competitive subregions rather than being dominated by a single attractor. Modern facades, by contrast, exhibit lower entropy and higher top-decile concentration, indicating a more peaked allocation in which a small number of regions capture disproportionate attention mass. Importantly, the increased allocation to entry AOIs for modern facades strengthens the functional interpretation: simplified surfaces and stronger figure–ground separation can elevate the salience of access cues, which are among the most semantically diagnostic elements for understanding use and navigability. That these typology contrasts persist after adjustment for measured composition controls and neighborhood-stratum structure suggests they are not reducible to trivial photographic differences (e.g., luminance, contrast, framing,

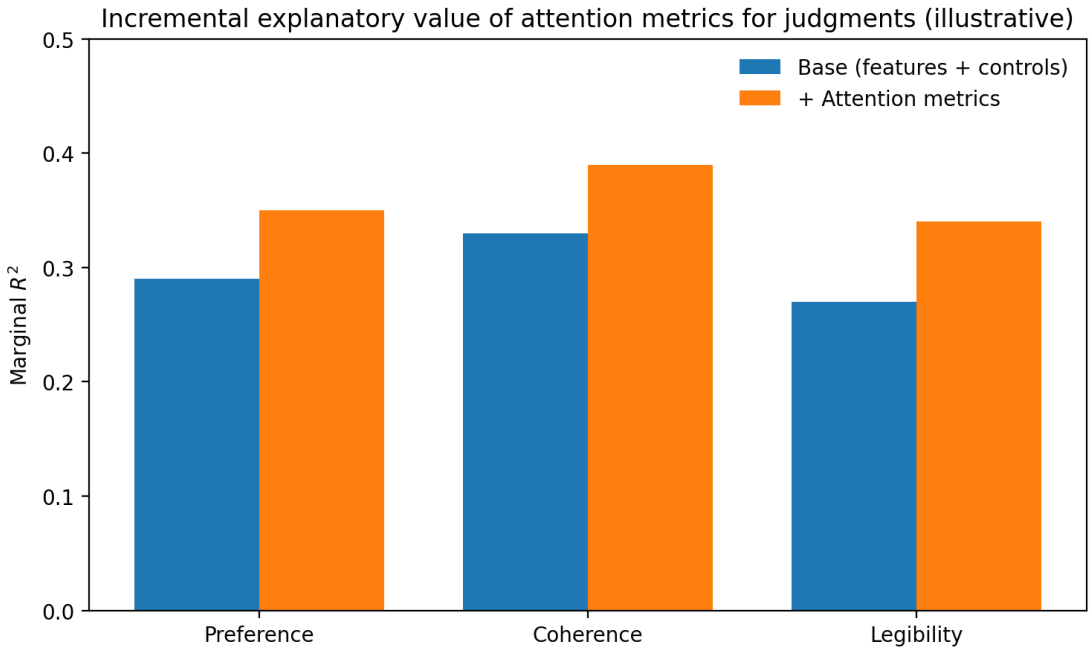


Figure 4: Incremental explanatory value of attention metrics for perceptual judgments. Bars report marginal $R^2$ for nested mixed models with and without attention predictors.

or viewpoint) nor to neighborhood context alone. Instead, they are consistent with the claim that typology proxies differences in the spatial distribution and semantic organization of cues that structure attention.

City effects are smaller and require more cautious interpretation. The conditional differences in concentration and dispersion after adjustment may reflect systematic variation in streetscape production regimes—including commercial intensity, signage ecology, maintenance practices, and typical facade articulation—that alter the competitive landscape for attention. However, "city" is an omnibus label that conflates cultural viewing conventions with institutional regulation and morphological composition. Without richer measurement of urban form (e.g., parcel rhythm, land-use mix, setback structure) and regulatory context (e.g., sign ordinances, heritage constraints), city coefficients should be treated as contextual contrasts rather than as evidence of culturally determined gaze strategies. The largely modest typology-by-city interaction provides some reassurance that the core typology mechanism generalizes across contexts within the studied domain, but it does not license strong claims about cross-cultural universality.

A central contribution of this study is showing that attention structure provides incremental explanatory power for perception beyond what is captured by coded facade features and low-level image descriptors. The directionality of associations is theoretically coherent when interpreted through an information-structuring lens. Coherence decreases as attention becomes more dispersed across competing cues, consistent with the idea that scenes containing many simultaneously salient elements reduce perceptual unification and increase interpretive effort. Legibility increases as attention mass concentrates on entrances, aligning with long-standing arguments that readable access structure is fundamental to environmental comprehension (Lynch, 1960). Preference exhibits a balance: modest dispersion is associated with higher preference (engagement through distributed interest), while excessive dominance by a few hotspots reduces preference, plausibly reflecting attentional capture by clutter or overly competing attractors that degrade perceived order. These patterns also clarify why feature inventories alone can be insufficient: two facades may have similar counts of elements, yet differ in how those elements compete for attention and thereby shape perceived organization.

Attention metrics operate as a quantitative intermediary that captures this competition structure.

Moderation by familiarity underscores that attention allocation is not equivalent to evaluation. The same attentional capture (e.g., by signage) can be interpreted differently depending on learned expectations and contextual knowledge: familiar viewers may treat commercial clutter as normative and therefore discount its implications for coherence, whereas unfamiliar viewers may interpret it as disorderly or overwhelming. This finding has methodological implications: cross-context studies that ignore familiarity risk conflating differences in cue distributions with differences in interpretive priors. Substantively, it suggests that "visual-performance" auditing can be made more decision-relevant by stratifying predictions and models by intended user groups (residents vs. visitors; experts vs. lay viewers), rather than treating the viewer as a single homogeneous perceptual system.

Several limitations define the appropriate scope of inference. First, the present framework targets fixation-density *distributions* and derived summary metrics; it does not model sequential scanpath dynamics. Emulation therefore cannot adjudicate claims about temporal ordering (e.g., whether viewers first fixate entrances and then scan ornament) without explicit sequence modeling and time-resolved validation. Second, street-view imagery unavoidably embeds uncontrolled variation in lighting, occlusion, and transient activity. Although we mitigate these influences via strict inclusion criteria, measured composition controls, and matched-subset sensitivity analyses, residual variation remains and can induce associations that are not purely architectural. Third, the design is observational with respect to image content: despite adjustment and balancing, causal interpretations of feature or typology effects remain contingent on untestable assumptions about omitted variables. Accordingly, the reported relationships should be read as adjusted associations that provide mechanistic consistency evidence, not as definitive causal decompositions. Strengthening causal claims would require additional identification strategies, such as controlled manipulations (e.g., systematically edited facades), within-image counterfactual edits (e.g., removing signage while holding other content fixed), or natural experiments exploiting exogenous policy changes.

Within these bounds, the study demonstrates how a validation-first, bias-aware emulation workflow can generate scalable, interpretable evidence about how facade cue distributions structure attention and how attention relates to perceived preference, coherence, and legibility. The broader implication is methodological: attention emulation can be scientifically and practically valuable in built-environment research when treated as a calibrated measurement instrument embedded in transparent inference, rather than as a black-box substitute for eye-tracking.

## CONCLUSION

We provide a validation-first, scalable framework for using fixation-density emulation to audit how streetscape façades allocate visual attention and how that allocation relates to preference, coherence, and legibility in comparative urban research. By integrating balanced stimulus sampling, reliable façade coding, bias-aware validation, and cross-classified multilevel inference with explicit composition adjustment and moderation tests, the framework enables reproducible, design-relevant "visual-performance" auditing that can be extended to broader urban contexts and policy questions.

## DATA AVAILABILITY

De-identified derived data (attention metrics, coded features, and survey aggregates) and analysis scripts are available in a public repository, subject to image licensing and privacy constraints.

## ETHICS STATEMENT

All human-subject procedures were reviewed and approved by an institutional ethics committee. Participants provide informed consent. Images should be curated to minimize identifiability; any remaining identifiers should be blurred prior to display.

## AUTHOR CONTRIBUTIONS

Conceptualization: J.B.H., K.D.S.; Methodology: J.B.H., L.V.A., K.D.S.; Analysis: L.V.A., K.D.S.; Writing—original draft: J.B.H.; Writing—review & editing: all authors.

## FUNDING

This work was supported by [funding information omitted for blind review / to be completed upon acceptance].

## ACKNOWLEDGMENTS

The authors thank the study participants and the anonymous reviewers for constructive feedback.

## REFERENCES

Borji, A., Sihite, D. N., & Itti, L. (2013). Analysis of scores, datasets, and models in visual saliency prediction. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 921–928.

Chua, H. F., Boland, J. E., & Nisbett, R. E. (2005). Cultural variation in eye movements during scene perception. *Proceedings of the National Academy of Sciences*, *102*(35), 12629–12633.

Duchowski, A. T. (2007). *Eye tracking methodology: Theory and practice* (2nd ed.). Springer.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.

Itti, L., Koch, C., & Niebur, E. (2001). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(11), 1254–1259.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2106–2113.

Kaplan, R., & Kaplan, S. (1989). *The experience of nature: A psychological perspective*. Cambridge University Press.

Kümmerer, M., Wallis, T. S. A., & Bethge, M. (2015). Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences*, *112*(52), 16054–16059.

Kümmerer, M., Wallis, T. S. A., & Bethge, M. (2016). Deepgaze ii: Reading fixations from deep features trained on object recognition.

Lavdas, A. A., Salingaros, N. A., & Sussman, A. (2021). Visual attention software: A new tool for understanding the "subliminal" experience of the built environment [Article 6197]. *Applied Sciences*, *11*(13), 6197.

Loos, A. (1908). Ornament and crime [Essay; reprinted in later collected editions (e.g., *Selected Essays*)].

Lynch, K. (1960). *The image of the city*. MIT Press.

Nasar, J. L. (1994). Urban design aesthetics: The evaluative qualities of building exteriors. *Environment and Behavior*, *26*(3), 377–401.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, *62*(8), 1457–1506.

Schirpke, U., Tasser, E., & Lavdas, A. A. (2022). Potential of eye-tracking simulation software for analyzing landscape preferences. *PLOS ONE*, *17*(10), e0273519.

Spanjar, G., & Suurenbroek, F. (2020). Eye-tracking the city: Matching the design of streetscapes in high-rise environments with users' visual experiences. *Journal of Digital Landscape Architecture*, *5*, 374–385.

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions [Article 4]. *Journal of Vision*, *7*(14), 4.

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, *11*(5), 1–23.

Venturi, R., Scott Brown, D., & Izenour, S. (1972). *Learning from las vegas*. MIT Press.

## AUTOBIOGRAPHICAL SKETCHES

Justin B. Hollander is an American urban planning and design scholar. He is a professor in the Department of Urban and Environmental Policy and Planning at Tufts University.

Kelly Sherman, background in Environmental Studies and Urban Planning experience. Climate Resilience Project Manager, Boston, Massachusetts, United States

Vinh An LE, a Vietnam-Japan Institute of Engineering and Technology, Duy Tan University, Da Nang City, Viet Nam.

Manuscript revisions completed 10 April 2022.