SimCityNet: Quantifying Geo-Cultural Bias in AI-Generated Urban Videos through Interpretable Scene Embeddings

Cantay Caliskan, Abylay Iskakov, Yuanzhu Li, Brendan McCollum, and Yi Ren

Goergen Institute for Data Science and Artificial Intelligence, University of Rochester, Rochester NY, United States

Abstract What happens when artificial intelligence is asked to envision the world's streets—and, crucially, whose world does it reproduce? This study examines how two state-of-the-art text-to-video systems, OpenAI's *Sora* and Kuaishou's *Kling AI*, portray urban life in the Middle East and Europe—regions chosen both for their visual contrast and for what they reveal about the persistent East–West divide in AI representation. To support a systematic audit, we introduce **SimCityNet**, a 33-dimensional, interpretable, image-only embedding that encodes vegetation cover, color–illumination signatures, pedestrian and vehicle counts, and surface proportions (roads, sidewalks, buildings). Unlike prior approaches, **SimCityNet** dispenses with ancillary GIS or POI data, enabling large-scale, model-agnostic comparison between synthetic and ground-truth imagery. Our benchmark comprises more than 600,000 images: 2,130 AI-generated clips (297,000 frames) from 110 economically active districts in Europe and the Middle East, paired with 312,533 Google Street View (GSV) photographs. Across these paired samples, we find that fidelity is block-specific and region-conditional rather than uniformly biased. Using attribute-level Mahalanobis and Wasserstein distances, *Sora* is typically closer to GSV than *Kling* on pedestrian and vehicle counts in both Europe (median differences of -2.65 and -0.75) and the Middle East (-1.09 and -1.97), while *Kling* is consistently closer on the color block (Europe: $+29.56$; Middle East: $+25.72$; positive values indicate *Sora* is farther from GSV). Cluster-aware Mann–Whitney tests show that only the Middle Eastern color block yields a robust difference ($p_{\text{cluster}} = 0.0008$), with European pedestrian differences marginally significant ($p_{\text{cluster}} = 0.046$), and most other blocks statistically indistinguishable. Linear mixed models with a Model×Region interaction confirm cross-region bias magnitudes: *Sora*'s vegetation distortion increases in the Middle East ($\beta_3 = 0.01998$, $p < 3 \times 10^{-68}$), its vehicle distortion rises ($\beta_3 = 1.94$, $p < 6 \times 10^{-258}$), and its pedestrian understatement deepens ($\beta_3 = -1.58$, $p < 7 \times 10^{-162}$). When we test whether models overstate static features and understate dynamic ones, Wilcoxon signed-rank results indicate that *Sora* significantly overstates the static block (median $\Delta_{\text{static}} = 3.88$, FDR-adjusted $p < 0.001$), but neither model exhibits a statistically reliable dynamic understatement after multiple-test correction. Finally, variance analyses reveal selective compression: color variance in synthetic frames collapses relative to GSV (variance ratios of 0.45 for *Sora*, $p = 0.0003$, and 0.13 for *Kling*, $p = 0.0001$), whereas vegetation, built surfaces, pedestrians, and vehicles show equal or greater variance than the ground truth (permutation $p \approx 1$). These results suggest that generative systems do not homogenize urban form wholesale. Instead, bias emerges where models can economize visually (e.g., color palettes) and where dynamic content is harder to render consistently across cultural contexts. **SimCityNet** thus serves both as a scalable audit tool and as a cautionary signal: when generative AI imagines the city, it may subtly redraw its chromatic and cultural boundaries. Future text-to-video systems should explicitly calibrate dynamic object rendering and color distributions across regions if they are to avoid quiet yet consequential distortions of how cities look and feel.

Keywords: AI-generated videos, Urban embeddings, Street-level benchmarking, Algorithmic fairness, Cross-cultural bias

1 Whose City Does the Algorithm Remember?

When a user types *"evening rush on a narrow souq in Doha"* or *"canal-side morning in Amsterdam"* into a modern text-to-video engine, the model now fabricates an entire streetscape in seconds—every cobblestone, neon sign, and passing pedestrian rendered in photorealistic motion. Systems such as OpenAI's *Sora* OpenAI (2025) and Kuaishou's *Kling AI* Kuaishou Technology (2024) already underpin film pre-visualization, tourism promotion, and academic urban simulation. Yet the seductive fidelity of these clips masks a fundamental uncertainty: *whose city are we really seeing*? Because the training corpora that power contemporary

generative models lean heavily toward Western imagery, an AI may portray a Middle-Eastern boulevard as eerily vacant or a European plaza as improbably verdant. Left unexamined, such distortions risk hard-coding cultural hierarchies into the way machines "remember" urban space.

To interrogate this problem at scale, we introduce **SimCityNet**, an interpretable embedding framework paired with a dataset that aligns 2,130 AI-generated videos (297,000 frames) with 312,533 geo-matched Google Street View photographs from 110 economically active neighborhoods across Europe and the Middle East. Rather than treating images as opaque pixels, *SimCityNet* converts every frame—synthetic or real—into a five-dimensional signature that records (i) the intensity and mix of motorized and non-motorized vehicle traffic, (ii) pedestrian density, (iii) the proportional area of trees, grass, and other vegetation, (iv) the relative footprint of sidewalks, roads, and buildings, and (v) the joint color-illumination profile that captures chromatic palettes and prevailing lighting conditions. These continuous descriptors let us test, for example, whether Sora portrays Abu Dhabi's Yas Marina as busier than it is during the off-season, or whether Kling overstates the nighttime crowds and lights of Dublin's Temple Bar.

Our full-scale analysis reveals a subtler, more patterned asymmetry than the preliminary snapshot implied. Fidelity varies by attribute block and region rather than shifting uniformly in one direction. Broadly, *Sora* hews closer to ground-truth counts of people and vehicles (e.g., median differences of roughly -2.6 pedestrians and -0.7 vehicles in Europe, and -1.1 and -2.0 in the Middle East), while *Kling* more faithfully reproduces color–illumination palettes (e.g., *Sora* is about $+29.6$ and $+25.7$ units farther from GSV on the color block in Europe and the Middle East, respectively). Only the Middle Eastern color block shows a difference that remains clearly robust under cluster-aware testing ($p_{\text{cluster}} = 0.0008$), and the apparent European pedestrian advantage for *Sora* is merely marginal ($p_{\text{cluster}} = 0.046$). Mixed-effects models confirm that *Sora*'s distortions in vegetation, vehicles, and pedestrians intensify outside its "home" region, but these effects do not translate into a wholesale pattern of static overstatement and dynamic understatement across both systems. Indeed, only *Sora* exhibits a statistically credible overstatement of static features; neither model shows a reliable, systematic understatement of dynamic activity once multiple testing is accounted for. Finally, where the models do homogenize, they do so selectively: color variance collapses relative to GSV (variance ratios ≈ 0.45 for *Sora* and 0.13 for *Kling*, permutation $p < 0.001$), while variance in vegetation, built surfaces, and human or vehicular activity remains as large—or larger—than in the ground truth (permutation $p \approx 1$). These findings echo geo-cultural skews documented in recent text-to-image evaluations Hall, Bell, et al. (2024); Wan et al. (2024), but they also complicate simplistic narratives of "flattened" urban diversity. By pairing narrative critique with a rigorously quantified audit, this paper advances ongoing debates in computational social science about algorithmic fairness. *SimCityNet* illustrates how large-scale, feature-rich benchmarking can move discussions beyond impressionistic claims toward evidence-based guidance for data curation, prompt design, and evaluation practices in next-generation generative video systems.

2 Literature on Generative Urban Media and Cityscape Analysis

OpenAI's text-to-video model *Sora* has quickly become a touchstone in debates over the promises and risks of generative media. Early empirical studies document a mix of excitement about lowering barriers to high-quality video production and apprehension about misinformation, copyright disputes, and the environmental costs of large-scale inference Zhou et al. (2024); Mogavi et al. (2024). Technical audits show that state-of-the-art models still struggle with temporal coherence and faithful frame-to-text alignment, producing motion artefacts or omitting key objects that the prompt demands Zeng et al. (2024); Sun et al. (2024). Liu Liu et al. (2024) goes further, arguing that without robust provenance tracking and content filtering, text-to-video systems risk perpetuating—or even amplifying—systemic inequities. These analyses foreground governance questions but rarely quantify how convincingly synthetic footage represents the richness of real urban life, leaving open the empirical question this study takes up.

Generative image research offers clues about the kinds of distortions that may migrate to video. Large diffusion models have been shown to reproduce and amplify demographic and geographic stereotypes across gender, race, occupation, and region Bianchi et al. (2023); Nadeem et al. (2025). Even ostensibly neutral prompts default to North-American visual tropes: Bianchi et al. report that 100% of "backyard," 99% of "front door," and 99% of "armchair" images resemble U.S. or Canadian settings, with scant representation of the Global South. Location-explicit queries reveal parallel patterns. Hall (Hall, Ross, et al. (2024)) found that "a car in Africa" yields sparse, low-diversity outputs, whereas the same prompt tagged "in Europe"

produces polished, modern scenes. Naik (Naik & Nushi (2023)) reaches similar conclusions using both human judges and Azure Cognitive Services: prompts referencing Nigeria, Ethiopia, or Papua New Guinea are among the least faithfully rendered, while the United States and Germany dominate across several diffusion platforms. Although most of these audits focus on single objects or faces, they underline the importance of studying entire urban environments—a gap the present work addresses by comparing *Sora* street scenes in Middle-Eastern and European contexts.

Computer-vision scholarship on real-world street imagery supplies a methodological blueprint for judging synthetic cityscapes. Gebru et al. demonstrated that convolutional networks trained on fifty million Google Street View (GSV) images can recover neighbourhood-level demographics via detected car types Gebru, Krause, Wang, Chen, Deng, et al. (2017). Zhang R. Zhang et al. (2018) shows that pedestrian-density embeddings, extracted with Faster R-CNN, capture demographic skew in large image collections—an approach readily transposed to video frames for assessing whether generated footage under-represents human activity. Vegetation metrics have also matured: Li et al. introduced a multi-angle Green View Index that correlates strongly with canopy coverage (r = 0.616) in Manhattan Li et al. (2015a), and Seiferling et al. validated GSV-based tree-cover estimates against LiDAR across two U.S. cities Seiferling et al. (2017). Dynamic features receive attention as well: deep spatio-temporal networks model traffic flows city-wide, proving that neural nets can approximate realistic mobility patterns at scale Y. Zhang et al. (2019); J. Zhang, Zheng, Qi, Li, et al. (2017). Dubey et al. extend the agenda to subjective perception, training a Siamese network on 1.17 million crowd-sourced comparisons that infer perceived safety and wealth from street photos Dubey et al. (2016a). Building on these precedents, the present study adapts VGG-16 embeddings and YOLO detectors to benchmark *Sora* videos against GSV baselines, measuring vegetation cover, people and vehicle counts, façade characteristics, and colour–illumination distributions.

Efforts to embed entire neighbourhoods in joint visual-textual spaces show additional promise. Urban2Vec fuses Inception-v3 features from street-view imagery with point-of-interest data, producing an unsupervised representation that explains nearly half the variance in real-estate sale prices Wang et al. (2020). Yet its reliance on geolocated image–POI pairs and its focus on property valuation limit its utility for auditing synthetic content. Our evaluation pipeline therefore eschews POI conditioning, instead computing normalised visual embeddings that enable direct, frame-level comparison between generated and authentic scenes. In doing so, we extend multimodal urban representation learning to the forensic analysis of AI-generated videos, quantifying where synthetic cityscapes deviate from the empirical distributions of greenery, human presence, traffic, and architectural density observed in analogous real-world footage.

Taken together, the literature suggests three unresolved challenges: existing text-to-video evaluations provide limited metrics for spatial-temporal realism, most bias audits neglect complex urban contexts, and multimodal urban embeddings have yet to be adapted for systematic comparison of synthetic and real scenes. By applying established street-level analytic tools to *Sora* outputs and juxtaposing them with matched GSV imagery, this study offers a principled framework for assessing the fidelity and fairness of AI-generated urban videos.

3 Research Questions and Hypotheses

Generative text-to-video models are trained on corpora that are unevenly distributed across the globe. Such imbalances raise a key question: when an American-built model such as *Sora* or a Chinese-built model such as *Kling AI* is asked to *re-create* a city street, whose visual reality is preserved and whose is distorted? To investigate, we rely on the *SimCityNet* corpus, which pairs synthetic frames from both systems with Google Street View (GSV) photographs from 110 economically active districts in Europe and the Middle East. Five measurable attributes—pedestrian activity, vehicle activity, vegetation cover, built-surface mix, and color–illumination profile—provide a common basis for comparison. Guided by this setting, we pose four research questions and a set of falsifiable hypotheses.

1. **RQ1 — Model comparison (Sora vs. Kling):** How differently do the American-developed *Sora* and the Chinese-developed *Kling AI* portray the same urban scenes?
2. **RQ2 — Regional comparison (Europe vs. Middle East):** Do synthetic depictions diverge more in one world region than in the other when benchmarked against GSV?

3. **RQ3 — Element-level fidelity:** Which dimensions of the five-part urban embedding—pedestrian activity, vehicle activity, vegetation cover, built surfaces, and color–illumination palette—are most prone to distortion by either model?
4. **RQ4 — City-level distinctiveness:** When AI tools are used, do cities come to look more alike or more different from one another than they do in ground-truth imagery?

Each hypothesis expresses a directional, testable claim. Taken together they reflect two guiding expectations: (i) a model is most faithful in the region that dominates its training data, and (ii) both models economize by emphasizing static, visually inexpensive content (vegetation and built surfaces) while under-representing dynamic, visually expensive content (pedestrians and vehicles).

– **H1 — Origin-aligned fidelity.**
  - *H1a (Europe):* In European districts, *Sora* is closer to GSV than *Kling AI*.
  - *H1b (Middle East):* In Middle-Eastern districts, *Kling AI* is closer to GSV than *Sora*.
– **H2 — Cross-region bias magnitude.**
  - *H2a (Environmental & structural gap): Sora* overstates vegetation and built surfaces (roads, sidewalks, buildings) more in the Middle East, whereas *Kling AI* does so more in Europe.
  - *H2b (Activity gap): Sora* understates pedestrian and vehicle counts more in the Middle East, whereas *Kling AI* does so more in Europe.
– **H3 — Static-versus-dynamic representation.**
  - *H3a (Static overstatement):* Both models depict higher levels of static features—vegetation, building facades, roads, and sidewalks—than are visible in GSV.
  - *H3b (Dynamic understatement):* Both models depict fewer pedestrians and vehicles than are visible in GSV.
– **H4 — Variance compression.**
  - *H4a (Environmental & structural convergence):* Variance across cities in vegetation cover, built-surface mix, and color–illumination profiles is lower in synthetic frames than in GSV.
  - *H4b (Activity convergence):* Variance across cities in pedestrian and vehicle counts is lower in synthetic frames than in GSV.

These research questions and hypotheses provide an empirical roadmap for evaluating whether contemporary text-to-video systems faithfully reproduce—or systematically distort—the diverse visual realities of European and Middle-Eastern urban life, and whether the efficiencies of generative modeling come at the cost of homogenizing the world's cities into a handful of familiar visual tropes.

4 Mathematical Formulation of the *SimCityNet* Embedding

Building on the gaps identified in existing audits of generative urban media, we now introduce **SimCityNet**, an interpretable, image-only embedding framework designed to compare synthetic and real street-level footage. Rather than relying on external GIS or POI data, *SimCityNet* distills each frame into a compact, 33-dimensional "scene fingerprint" that captures vegetation cover, color distribution, dynamic object counts, and surface proportions. Below, we present the mathematical formulation of this embedding and describe the code-driven procedures by which each component is extracted, normalized, and concatenated to yield the final representation.

SimCityNet encodes each street-level photograph I into a fixed-length, interpretable feature vector $\Phi(I) \in R^{33}$ using only pixel-derived information. By dispensing with ancillary GIS layers or proprietary metadata, this purely image-driven framework applies uniformly to both real-world panoramas and synthetically generated frames. Concretely, *SimCityNet* converts each input image into a 33-dimensional vector composed of five interpretable blocks—vegetation cover, color signature, pedestrian count, remaining dynamic detections, and surface proportions—that jointly capture the static form, chromatic atmosphere, and human and vehicular activity of an urban scene.

We formalize *SimCityNet* as a mapping

$$\Phi : \mathcal{I} \longrightarrow R^D, \qquad \Phi(I) = \big[f_{\text{veg}}(I),\ f_{\text{color}}(I),\ c_{\text{ped}}(I),\ \mathbf{d}_{\neg\text{ped}}(I),\ \mathbf{u}(I)\big] \in R^{33},$$

where $\mathcal{I}$ is the space of all street-level images and in our implementation $D = 33$. For each image I, we compute five interpretable, pixel-derived measurement blocks and concatenate them into

$$\Phi(I) = \begin{bmatrix} f_{\text{veg}}(I) \\ f_{\text{color}}(I) \\ c_{\text{ped}}(I) \\ \mathbf{d}_{\neg\text{ped}}(I) \\ \mathbf{u}(I) \end{bmatrix} \in R^D.$$

Here:

- $f_{\text{veg}}(I) \in [0, 1]$ is the fractional vegetation cover.
- $f_{\text{color}}(I) \in R^{20}$ comprises five RGB centroids and their normalized weights.
- $c_{\text{ped}}(I) \in N$ is the pedestrian (human) count, isolated as a single scalar.
- $\mathbf{d}_{\neg\text{ped}}(I) \in N^8$ contains COCO detection counts for the remaining dynamic classes (bicycle, car, motorcycle, airplane, bus, train, truck, boat).
- $\mathbf{u}(I) \in [0, 1]^3$ records road, sidewalk, and building pixel proportions.

To render these heterogeneous measurements directly comparable, we apply a diagonal scaling matrix $N = \text{diag}(n_1, \ldots, n_D)$ with

$$n_k = \frac{1}{\max_{I \in \mathcal{I}} \Phi_k(I)} \quad \text{(or a chosen upper bound)},$$

so that no single coordinate dominates the embedding. The normalized vector

$$\widetilde{\Phi}(I) \;=\; N\,\Phi(I)$$

thereby lies in $[0, 1]^D$ for all fractional blocks, while integer counts (including $c_{\text{ped}}(I)$ and entries of $\mathbf{d}_{\neg\text{ped}}(I)$) are rescaled to a comparable range (with all coordinates equally weighted in cross-picture comparisons).

Within this D-dimensional feature space, standard similarity and distance measures become meaningful probes of visual affinity. The cosine similarity between two images I and J is

$$\text{sim}(I, J) = \frac{\widetilde{\Phi}(I) \cdot \widetilde{\Phi}(J)}{\|\widetilde{\Phi}(I)\|_2 \, \|\widetilde{\Phi}(J)\|_2},$$

capturing angular alignment in the normalized space. Alternatively, the (equally weighted) Euclidean distance

$$d(I, J) = \|\widetilde{\Phi}(I) - \widetilde{\Phi}(J)\|_2 = \sqrt{\sum_{k=1}^{D} \big(\widetilde{\Phi}_k(I) - \widetilde{\Phi}_k(J)\big)^2}$$

treats each coordinate as contributing identically to overall dissimilarity. If desired, a positive-definite weighting matrix W could be introduced to emphasize or de-emphasize blocks, but all reported cross-picture analyses in this study use the equal-weight default.

Because each coordinate of $\Phi(I)$ admits a direct semantic interpretation, *SimCityNet* serves both as an embedding for downstream tasks—clustering, classification, retrieval—and as a diagnostic lens: a change along the kth axis corresponds to a concrete visual property (e.g., an increase in f_{veg} signals more tree canopy, whereas a spike in $c_{\text{ped}}(I)$ indicates higher foot traffic). In the next section, we detail the code-based procedures for computing each component of $\Phi(I)$, as well as the normalization and fusion operations that yield the final *SimCityNet* representation.

4.1 Vegetation Cover (f_{veg}): A Depth-Aware Indexing Pipeline

The vegetation component of the *SimCityNet* embedding, $f_{\text{veg}}(I)$, is computed by a depth-aware vegetation indexing pipeline, which integrates image normalization, semantic segmentation, spectral refinement, and monocular depth estimation into a unified process that emphasizes perceptual salience.

Each panoramic input is first converted to grayscale and thresholded to identify the primary content window, removing extraneous black borders and warped regions. The result is resized to a fixed resolution (e.g. 4000×2000) via Lanczos interpolation to ensure consistency for downstream models.

A Mask2Former ((Cheng et al., 2022)) network then produces a per-pixel class map $\mathcal{S}(i, j)$. From this map we extract three vegetation masks M_{tree}, M_{grass}, and M_{plant} by selecting ADE20K indices $\{4, 9, 17\}$. To filter out false positives—such as green-painted surfaces—each mask is refined in HSV space: the binary mask $\widetilde{M}_c(i, j)$ retains only those pixels for which both the semantic label and category-specific HSV thresholds (e.g. hue–saturation range $[20, 40]$–$[95, 255]$ for trees) are satisfied.

To capture visual prominence, we estimate a dense depth map $D(i, j) \in [0, 1]$ using a pretrained Dense Prediction Transformer. The raw depth predictions are min–max normalized and spatially aligned with the segmentation masks, yielding a continuous weighting field that prioritizes foreground foliage.

Finally, for each vegetation category c, we compute the depth-weighted Green View Index

$$\text{GVI}_c^{\text{depth}} = \frac{\sum_{i,j} \widetilde{M}_c(i, j)\, D(i, j)}{\sum_{i,j} D(i, j)}.$$

A combined index over the union of all refined masks is defined analogously. These depth-weighted indices $\{\text{GVI}_c^{\text{depth}}\}$ are concatenated to form the single component $f_{\text{veg}}(I)$, which encodes both the extent and the perceptual salience of urban greenery.

4.2 Computing the Color Signature ($f_{\text{color}}(I)$)

The color signature, denoted $f_{\text{color}}(I)$, constitutes the component of the *SimCityNet* embedding responsible for encoding the chromatic properties of an urban scene. Its generation is operationalized through a statistical color distribution analysis, the foundation of which rests on partitional clustering. The underlying methodology employs a K-Means algorithm, specifically parameterized with $K = 5$ clusters, to partition the pixel data of an input image I into five distinct groups based on color similarity.

Subsequent to clustering, the algorithm extracts two sets of parameters. First, the centroid of each cluster, $\mu_k \in [0, 255]^3$ for $k = 1, \ldots, 5$, is identified as a dominant color. Concurrently, the relative prevalence of each cluster is quantified as a normalized weight, P_k, representing the fraction of pixels assigned to that cluster, such that $\sum_{k=1}^{5} P_k = 1$. The final feature vector is constructed through the direct concatenation of these two sets of parameters. The five 3-dimensional RGB centroids are flattened and combined with the five scalar percentage weights, yielding the final 20-dimensional vector formally expressed as:

$$f_{\text{color}}(I) = [\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, P_1, P_2, P_3, P_4, P_5] \in R^{20}$$

This vector provides a rich, quantitative fingerprint of the scene's color palette, capturing both the primary hues and their relative spatial importance.

4.3 Human and Vehicle Dynamics: c_{ped} and $\text{d}_{\neg\text{ped}}$

The human and vehicle components of the *SimCityNet* embedding quantify the presence of pedestrians and the composition of vehicular activity in each frame. After benchmarking several state-of-the-art detectors on a held-out validation set (Table 1), YOLOv9e exhibited the highest recall for persons (0.985) and cars (0.938), as well as the leading mAP scores for the person class. We therefore adopt YOLOv9e as the default detector, prioritizing accurate counting over marginal gains in runtime.

Formally, let an input image I produce a set of detections

$$\mathcal{B}(I) = \{(\ell_j, \alpha_j)\}_{j=1}^{J_I},$$

where ℓ_j is the predicted class label and $\alpha_j \in [0, 1]$ its associated confidence. With a fixed threshold τ, define $\mathbf{1}_\tau(\alpha_j) = I[\alpha_j \geq \tau]$. The pedestrian count is

$$c_{\text{ped}}(I) = \sum_{j=1}^{J_I} I[\ell_j = \texttt{person}]\, \mathbf{1}_\tau(\alpha_j),$$

isolated as a single scalar coordinate. The remaining dynamic detections—vehicles and related classes—form an 8-dimensional vector

$$\mathbf{d}_{\neg\text{ped}}(I) = \big[d_{\texttt{bicycle}}(I),\ d_{\texttt{car}}(I),\ d_{\texttt{motorcycle}}(I),\ d_{\texttt{airplane}}(I),\ d_{\texttt{bus}}(I),\ d_{\texttt{train}}(I),\ d_{\texttt{truck}}(I),\ d_{\texttt{boat}}(I)\big]^\top,$$

with

$$d_c(I) = \sum_{j=1}^{J_I} I[\ell_j = c]\,\mathbf{1}_\tau(\alpha_j), \quad c \in \{\texttt{bicycle}, \texttt{car}, \texttt{motorcycle}, \texttt{airplane}, \texttt{bus}, \texttt{train}, \texttt{truck}, \texttt{boat}\}.$$

Together, these counts populate the dynamic block

$$\mathbf{f}_{\text{dyn}}(I) = \begin{bmatrix} c_{\text{ped}}(I) \\ \mathbf{d}_{\neg\text{ped}}(I) \end{bmatrix} \in N^9,$$

which is concatenated with the other components to form $\Phi(I)$ and then normalized by the diagonal matrix N, as described in the main formulation of *SimCityNet*. After normalization, all coordinates are equally weighted in cross-picture comparisons, ensuring that a unit change in any dimension carries the same influence on distance calculations.

Model	YOLOv8x	YOLOv9e	YOLOv11x	Faster R-CNN	RetinaNet
Recall – person	0.912	0.985	0.857	0.6000	0.6238
Recall – car	0.907	0.938	0.887	0.6727	0.6682
mAP50 – person	0.950	0.973	0.952	0.8821	0.8740
mAP50 – car	0.941	0.958	0.965	0.8619	0.8830
mAP50–95 – person	0.628	0.648	0.626	0.5060	0.5367
mAP50–95 – car	0.653	0.673	0.667	0.5796	0.5885
Inference Time (ms)	44.8	53.2	39.1	128.4	121.0

Table 1: Performance comparison of object detectors on held-out validation frames. Highlighted cells mark the best score per row. YOLOv9e offers the strongest recall and mAP for the key human and car classes, motivating its use in *SimCityNet*. All inferences were performed on a laptop equipped with an NVIDIA RTX 4090 GPU.

4.4 Surface Proportions u(I): Roads, Sidewalks, and Buildings

The structural component of *SimCityNet* quantifies how much of a frame's visible area is occupied by roads, sidewalks, and buildings. Let Ω_I be the set of all pixels in image I, and let $L_I : \Omega_I \to \mathcal{C}$ denote a semantic label map produced by a segmentation model trained on urban scenes (e.g., a DeepLabV3+ ((Chen et al., 2017)) variant fine-tuned on Cityscapes). We focus on $\mathcal{C}_{\text{surf}} = \{\texttt{road}, \texttt{sidewalk}, \texttt{building}\}$. For each class $c \in \mathcal{C}_{\text{surf}}$, define

$$u_c(I) \;=\; \frac{1}{|\Omega_I|} \sum_{p \in \Omega_I} I[L_I(p) = c],$$

the fraction of pixels labeled as c. Stacking these entries yields

$$\mathbf{u}(I) = \big[u_{\texttt{road}}(I),\ u_{\texttt{sidewalk}}(I),\ u_{\texttt{building}}(I)\big]^\top \in [0,1]^3.$$

By construction, $\sum_{c \in \mathcal{C}_{\text{surf}}} u_c(I) \le 1$, since any remaining pixels belong to other classes (e.g., vegetation, sky, water). This vector $\mathbf{u}(I)$ is concatenated with the vegetation, color, pedestrian, and vehicle blocks to form the full embedding $\Phi(I)$ and subsequently normalized by the diagonal matrix N. After normalization, $\mathbf{u}(I)$ is treated on equal footing with all other coordinates in cross-picture comparisons (i.e., all dimensions receive equal weight unless an alternative scheme is explicitly specified), ensuring that surface composition contributes transparently and comparably to our similarity and distance metrics.

5 Data Collection

This study draws on a hybrid corpus that combines real-world street imagery with synthetic video frames generated by two advanced text-to-video systems, OpenAI's Sora and Kuaishou's Kling AI. The data pipeline unfolds in three coordinated stages: defining the sampling universe, harvesting ground-truth images, and producing matched synthetic content. The sampling universe was anchored in economic comparability. Using the 2024 World Bank GDP-per-capita rankings, we selected the seven wealthiest nations in Europe—Luxembourg, Switzerland, Norway, Denmark, Ireland, Iceland, and the Netherlands—and the seven wealthiest in the Middle East—Qatar, the United Arab Emirates, Israel, Saudi Arabia, Kuwait, Bahrain, and Oman. Within each country we targeted neighborhoods with dense commercial activity—highlighted in yellow on Google Maps because of their concentrations of restaurants, shops, and offices—and manually traced polygons around every qualifying cluster. This procedure yielded geofences for fourteen Middle-Eastern cities (Doha; Haifa; Jerusalem; Tel Aviv; Dubai; Abu Dhabi; Riyadh; Kuwait City; Manama; Muscat) and seven European cities (Luxembourg City; Zürich; Oslo; Copenhagen; Dublin; Reykjavik; Amsterdam). An exhaustive inventory of economic activity zones appears in Appendix A. Figures 1 and 2 show the boundaries of the selected cities, highlighting regions of economic activity in red color.

Following the conversion of geofences into polygons, each polygon was rasterized into a lattice of ten-meter squares. The four vertices of every grid cell served as candidate observation points, and where Google Street View (GSV) imagery was available, the corresponding panorama was downloaded. Between January and March 2025 this process yielded 209,117 European images and 103,416 Middle-Eastern images. Table 2 summarizes the resulting corpus, reporting—for each city and for each region in aggregate—the total number of images retrieved, land area, population, image density, and images per capita. Pronounced regional variation is evident. Several European cities, notably Copenhagen and Dublin, exceed 390 images per square kilometer, whereas many Middle-Eastern neighborhoods are markedly sparser, and some locations in UAE, Saudi Arabia and Kuwait lack any GSV coverage at all. Our analysis relies exclusively on panoramic images captured by Google's own Street View fleet; photographs contributed by local residents through Google Maps are intentionally excluded to ensure uniform camera hardware, capture protocols, and geospatial metadata. These disparities in official coverage constitute an exogenous geographic bias that must be considered when benchmarking the fidelity of AI-generated video content.

To generate the textual prompts for our synthetic dataset, we instructed ChatGPT-4o to produce five alternative descriptions for each neighborhood using the following query:

> *Can you create me a general description of a scene in [neighborhood, city, country]? Please create five alternative descriptions of the same location for me.*

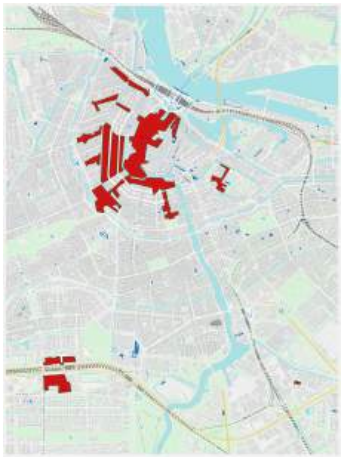
Each resulting description was then embedded within a standardized prompt structure designed for video generation:

> *"Show me a hyper-realistic street scene captured with a wide-angle lens, similar to Google Street View, based on the following description of a neighborhood: [DESCRIPTION]."*

This use of varied descriptions was designed to capture diverse facets of local culture, architecture, vegetation, traffic, and lighting. Such descriptive diversity broadens the representativeness of the AI-generated videos; by submitting multiple, stylistically distinct prompts to Sora and Kling AI, we mitigate sensitivity to any single phrasing and yield a broader spectrum of plausible visual outcomes. Each complete prompt was submitted independently to both AI platforms, and all frames were extracted from every resulting video to create two parallel synthetic corpora per neighborhood. Representative prompt structures are shown in Table 3.

To facilitate this large-scale generation, we utilized premium subscriptions to ChatGPT Pro for two months (200 USD per month) and Kling AI Premier for three months (64.99 USD per month). The increased quotas and faster processing times associated with these tiers were essential for producing the requisite volume of data. The total expenditure of 594.97 USD yielded a comprehensive video corpus spanning 110 commercially active areas across 21 cities in Europe and the Middle East. Table 4 provides a detailed distribution of the generated videos by city and platform.

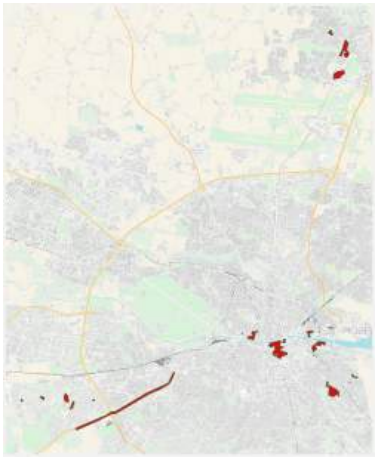
Finally, all real and synthetic frames are geotagged by neighborhood and timestamped, providing a unified spatial index that underpins downstream tasks such as generating *SimCityNet* embeddings, quantifying



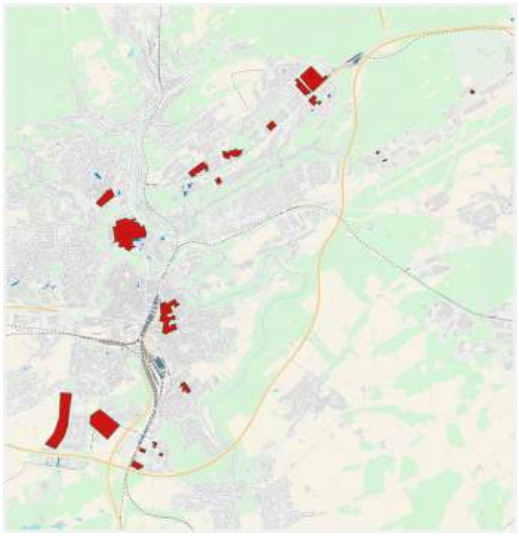
(a) Amsterdam, Netherlands



(b) Copenhagen, Denmark



(c) Dublin, Ireland



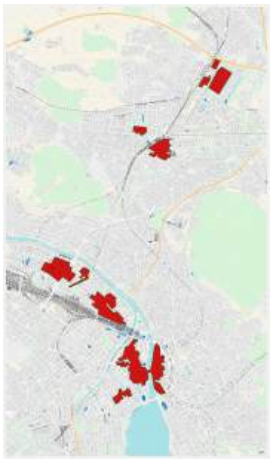
(d) Luxembourg, Luxembourg



(e) Oslo, Norway

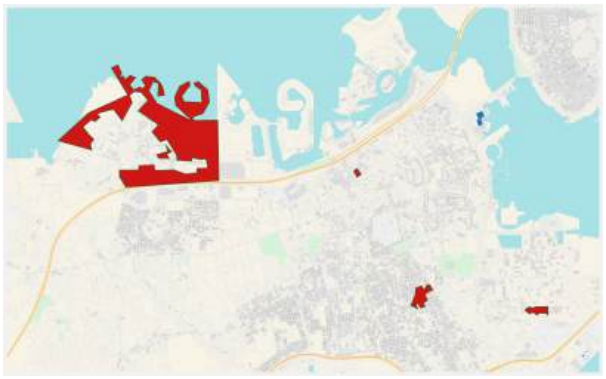


(f) Reykjavik, Iceland



(g) Zürich, Switzerland

Figure 1: Areas of economic activity (denoted by red shading) in seven selected *European* cities.



(a) Manama, Bahrain



(b) Muscat, Oman



(c) Abu Dhabi, UAE



(d) Haifa, Israel



(e) Kuwait City, Kuwait



(f) Riyadh, Saudi Arabia



(g) Dubai, UAE



(h) Jeddah, Saudi Arabia



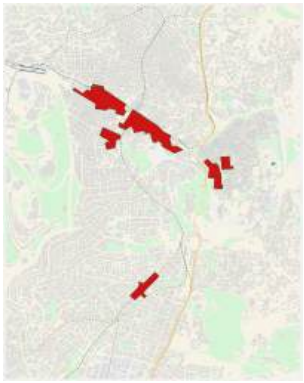
(i) Mecca, Saudi Arabia



(j) Tel Aviv, Israel



(k) Doha, Qatar



(l) Jerusalem, Israel



(m) Medina, Saudi Arabia



(n) Sharjah, UAE

Figure 2: Areas of economic activity (denoted by red shading) in fourteen selected *Middle Eastern* cities.

cross-domain visual similarity, and assessing representation fairness. Figures 3 and 4 illustrate the pairing of ground-truth panoramic imagery and corresponding AI-generated frames across the fourteen study cities, demonstrating the breadth of visual content captured by the pipeline.

Table 2: Summary of Ground-Truth Street View Coverage in European and Middle Eastern Cities

City	Country	Region	#Pictures	Area (km²)	Population	#Pics/km²	#Pics/Person
Amsterdam	Netherlands	Europe	53 988	165.00	933 680	327.20	0.0578
Copenhagen	Denmark	Europe	44 775	88.25	661 000	507.37	0.0677
Dublin	Ireland	Europe	45 166	114.99	593 000	392.78	0.0762
Luxembourg City	Luxembourg	Europe	9 762	51.73	136 208	188.71	0.0717
Oslo	Norway	Europe	34 554	454.00	717 710	76.11	0.0482
Reykjavik	Iceland	Europe	6 483	274.00	136 000	23.62	0.0477
Zürich	Switzerland	Europe	14 389	91.88	448 664	156.61	0.0321
Europe Total			**209,117**	**1,239.85**	**3,626,262**	**168.66**	**0.0577**
Abu Dhabi	UAE	Middle East	NA	972.00	1,590,000	NA	NA
Doha	Qatar	Middle East	3 107	132.00	666 000	23.54	0.0047
Dubai	UAE	Middle East	64 839	1 553.45	3,799,109	41.74	0.0171
Haifa	Israel	Middle East	14 746	63.70	291,000	231.49	0.0507
Jeddah	Saudi Arabia	Middle East	NA	1 765.00	4,943,000	NA	NA
Jerusalem	Israel	Middle East	3 293	126.00	983,000	26.13	0.0034
Kuwait City	Kuwait	Middle East	NA	200.00	600,000	NA	NA
Manama	Bahrain	Middle East	NA	30.00	297,502	NA	NA
Mecca	Saudi Arabia	Middle East	NA	760.00	2,185,000	NA	NA
Medina	Saudi Arabia	Middle East	NA	1 531.00	1,599,000	NA	NA
Muscat	Oman	Middle East	3 669	3 500.00	1,499,549	1.05	0.0024
Riyadh	Saudi Arabia	Middle East	NA	1 973.00	7,821,000	NA	NA
Sharjah	UAE	Middle East	7 457	506.33	1,872,200	14.73	0.0040
Tel Aviv	Israel	Middle East	6 305	52.00	498,915	121.25	0.0126
Middle East Total			**103,416**	**13,164.48**	**28,645,275**	**17.43**	**0.0108**
Grand Total			**312,533**	**14,404.33**	**32,271,537**	**43.57**	**0.0236**

6 Empirical Strategy

The purpose of this section is twofold: to spell out what we test and to explain *why* each methodological choice follows from the structure of *SimCityNet*, our hypotheses, and the characteristics of image-derived data. Our strategy unfolds along three interlocking steps. First, each interpretable block of the *SimCityNet* embedding is reduced to a single, comparable divergence metric—Mahalanobis distances for vector-valued blocks and the Wasserstein (Earth Mover's) distance for the color distribution—while raw signed differences are retained for directional hypotheses. Second, we employ statistical tools that align with the data's architecture: paired images and nested districts with likely non-normal distributions motivate Mann–Whitney tests with clustered p-values, linear mixed models, Wilcoxon signed-rank tests, and variance-equality tests supplemented by permutation procedures. Third, these tests are situated within multivariate and robustness frameworks to demonstrate that the conclusions are not artifacts of particular specifications or sample idiosyncrasies. Table 5 previews this empirical road map; the subsections that follow elaborate each step in detail.

6.1 Operationalizing the *SimCityNet* Blocks

For each AI–GSV pair we work with two normalized 33-dimensional vectors, $\widetilde{\Phi}^{\mathrm{AI}}(I)$ and $\widetilde{\Phi}^{\mathrm{GSV}}(I)$, whose coordinates are equally weighted in cross-picture comparisons. To obtain a single, comparable measure of divergence for each attribute block, we standardize the relevant coordinates using z-scores computed from the GSV distribution and then, for vector-valued blocks, measure discrepancy with a Mahalanobis distance. Concretely, for pedestrians (c_{ped}) and vehicles ($\mathbf{d}_{\neg\mathrm{ped}}$), as well as for vegetation (f_{veg}) and built surfaces

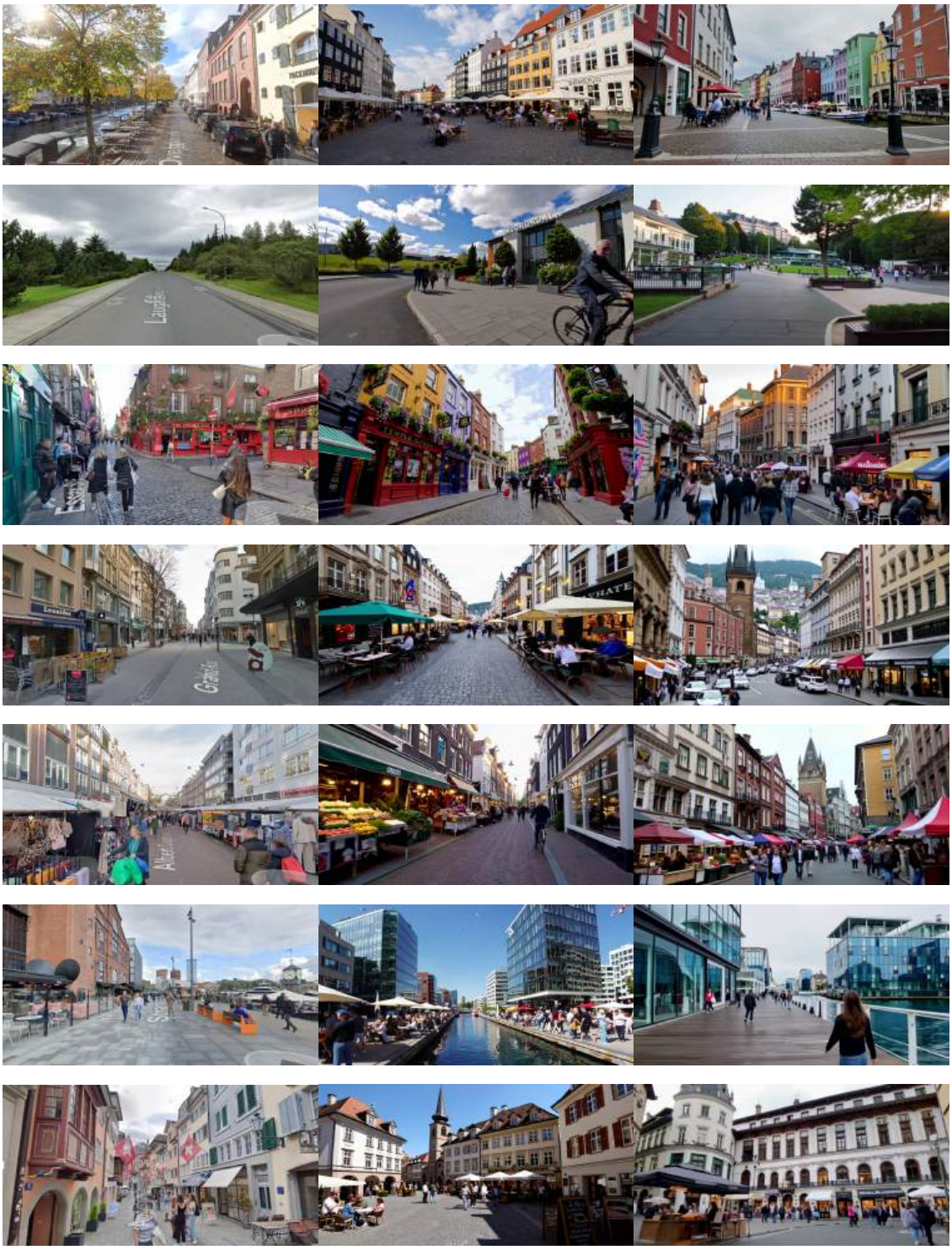


Figure 3: Comparison of (i) real Google Street View frames, (ii) SORA-generated frames, and (iii) Kling-generated frames for seven neighborhoods in *Europe*: (1) Christianshavn (Copenhagen, Denmark), (2) Laugardalur (Reykjavik, Iceland), (3) Temple Bar (Dublin, Ireland), (4) Ville Haute (Luxembourg City, Luxembourg), (5) De Pijp (Amsterdam, Netherlands), (6) Aker Brygge & Tjuvholmen (Oslo, Norway), and (7) Altstadt (Zurich, Switzerland).

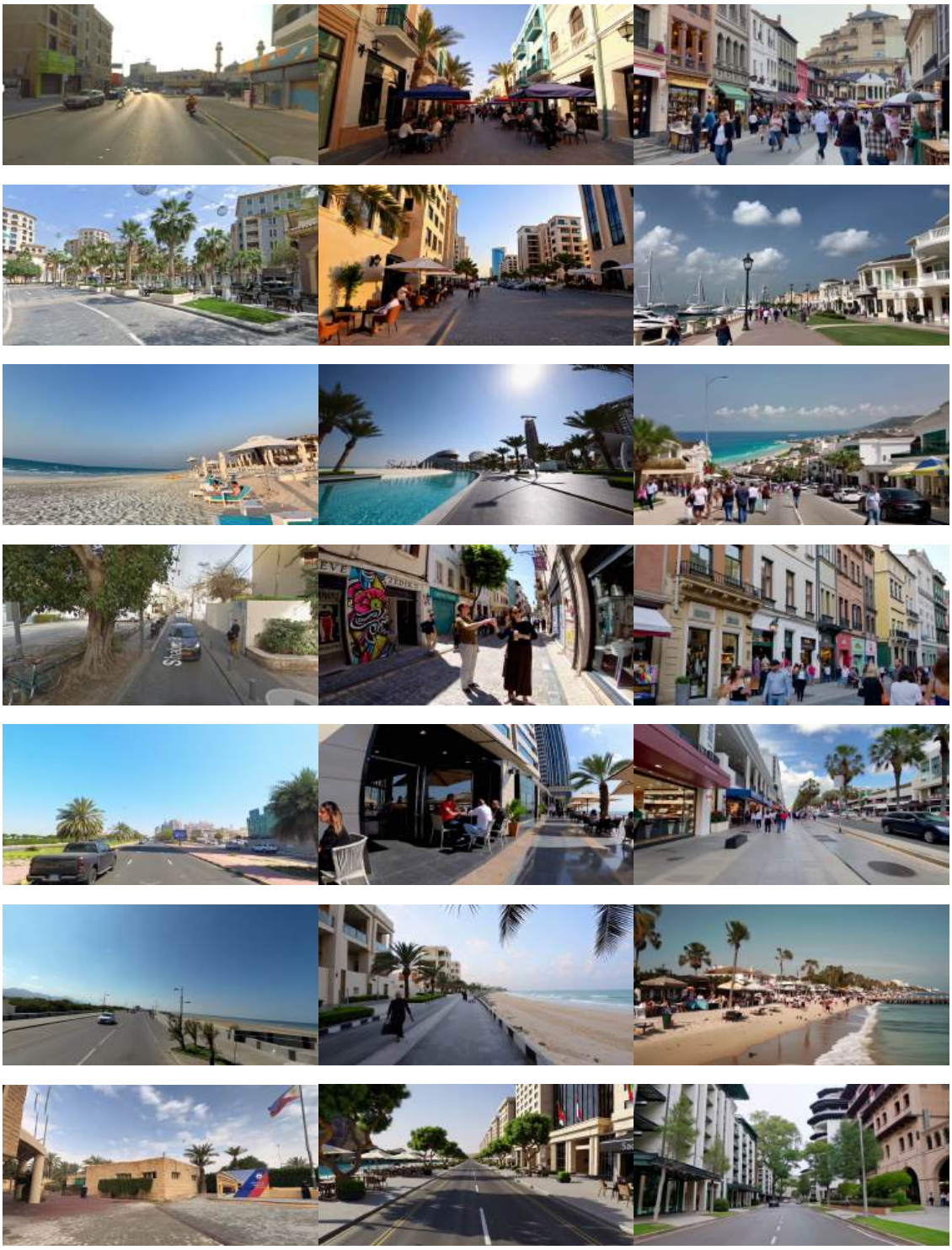


Figure 4: Comparison of (i) real Google Street View frames, (ii) SORA-generated frames, and (iii) Kling-generated frames for seven neighborhoods in the *Middle East*: (1) Adliya (Manama, Bahrain), (2) The Pearl (Doha, Qatar), (3) Saadiyat Island (Abu Dhabi, United Arab Emirates), (4) Neve Tzedek (Tel Aviv, Israel), (5) Salmiya (Kuwait City, Kuwait), (6) Qurum (Muscat, Oman), and (7) Diplomatic Quarter (Riyadh, Saudi Arabia).

Table 3: Sample Prompt Structures by Neighborhood

Region	**Neighborhood**	**Prompt**
Europe	Christianshavn, Copenhagen	Show me a hyper-realistic street scene captured with a wide-angle lens, similar to Google Street View, based on the following description of a neighborhood: Christianshavn is a vibrant and trendy district in Copenhagen, with colorful buildings lining the canal. The area is bustling with life, as locals and tourists enjoy outdoor cafes, modern restaurants, and the view of the water. The contrast of historic warehouses and new developments gives the district a unique, eclectic charm.
Europe	Temple Bar, Dublin	Show me a hyper-realistic street scene captured with a wide-angle lens, similar to Google Street View, based on the following description of a neighborhood: Temple Bar is Dublin's lively cultural quarter, where cobblestone streets are lined with historic pubs, galleries, and cafes. The area is a favorite for both locals and tourists, offering a mix of traditional Irish music, vibrant nightlife, and artistic vibes.
Europe	City Center, Zürich	Show me a hyper-realistic street scene captured with a wide-angle lens, similar to Google Street View, based on the following description of a neighborhood: The Altstadt, or Old Town, is the historic heart of Zürich, where narrow cobblestone streets wind between medieval buildings. The area is filled with charming boutiques, cafes, and ancient churches, giving it a timeless appeal. Locals and tourists alike enjoy wandering through this picturesque district, soaking in its rich history and vibrant atmosphere.
Middle East	Downtown Abu Dhabi, Abu Dhabi	Show me a hyper-realistic street scene captured with a wide-angle lens, similar to Google Street View, based on the following description of a neighborhood: Downtown Abu Dhabi, anchored by the picturesque Corniche, is a bustling district known for its stunning waterfront views, modern skyscrapers, and lively urban atmosphere. With family-friendly parks, pristine beaches, and cultural landmarks like Qasr Al Hosn, it serves as a dynamic hub for both residents and visitors.
Middle East	Downtown Dubai, Dubai	Show me a hyper-realistic street scene captured with a wide-angle lens, similar to Google Street View, based on the following description of a neighborhood: An awe-inspiring view of Downtown Dubai at dusk, with the towering Burj Khalifa piercing the sky, its glass façade reflecting golden sunlight, while the Dubai Fountain performs its choreographed dance below, surrounded by luxury hotels and modern architecture.
Middle East	Jaffa Road, Jerusalem	Show me a hyper-realistic street scene captured with a wide-angle lens, similar to Google Street View, based on the following description of a neighborhood: Jaffa Road buzzes with life as both locals and visitors move through the heart of the city. The street is flanked by charming cafés, diverse shops, and historic architecture, all glowing with a vibrant atmosphere. People pause to socialize, browse storefronts, or sip coffee at sidewalk tables, giving the area a warm and lively character.

$(\mathbf{u}(I))$, we compute

$$d_{\text{Mah}}(I) = \sqrt{\big(\mathbf{z}^{\text{AI}}(I) - \mathbf{z}^{\text{GSV}}(I)\big)^{\top} \boldsymbol{\Sigma}^{-1} \big(\mathbf{z}^{\text{AI}}(I) - \mathbf{z}^{\text{GSV}}(I)\big)},$$

where $\mathbf{z}^{\text{AI}}$ and $\mathbf{z}^{\text{GSV}}$ are the standardized vectors for that block and $\boldsymbol{\Sigma}$ is the corresponding covariance matrix estimated from GSV.

The color–illumination block (f_{color}) differs in nature: it is treated as a probability distribution over five RGB centroids, and we therefore quantify its dissimilarity using the first-order Wasserstein (Earth Mover's) distance in RGB space. Formally, let $P = \{(\mu_k, p_k)\}_{k=1}^{5}$ and $Q = \{(\nu_m, q_m)\}_{m=1}^{5}$ denote the two color histograms (centroids $\mu_k, \nu_m \in R^3$; weights $p_k, q_m \geq 0$, $\sum_k p_k = \sum_m q_m = 1$). Define the ground cost $c_{km} = \|\mu_k - \nu_m\|_2$. The EMD solves the optimal transport problem

$$\text{EMD}(P, Q) = \min_{\{f_{km}\}} \sum_{k=1}^{5} \sum_{m=1}^{5} f_{km}\, c_{km} \quad \text{s.t.} \quad \sum_{m=1}^{5} f_{km} = p_k,\ \sum_{k=1}^{5} f_{km} = q_m,\ f_{km} \geq 0.$$

Table 4: Breakdown of AI-generated videos by city and platform

City	Country	Region	#Neighborhoods	#Videos	Total Duration (s)	Platform	#Frames
Amsterdam	Netherlands	Europe	9	180	8,100	Sora	13,500
						Kling AI	10,800
Copenhagen	Denmark	Europe	8	90	3,600	Sora	12,000
						Kling AI	9,600
Dublin	Ireland	Europe	10	200	10,000	Sora	15,000
						Kling AI	12,000
Luxembourg City	Luxembourg	Europe	7	140	4,900	Sora	10,500
						Kling AI	8,400
Oslo	Norway	Europe	7	140	4,900	Sora	10,500
						Kling AI	8,400
Reykjavík	Iceland	Europe	8	160	6,400	Sora	12,000
						Kling AI	9,600
Zürich	Switzerland	Europe	5	100	2,500	Sora	7,500
						Kling AI	6,000
Europe Total			**54**	**1,010**	**40,400**		**145,800**
Abu Dhabi	UAE	Middle East	4	80	1,600	Sora	6,000
						Kling AI	4,800
Doha	Qatar	Middle East	5	100	2,500	Sora	7,500
						Kling AI	6,000
Dubai	UAE	Middle East	8	160	6,400	Sora	12,000
						Kling AI	9,600
Haifa	Israel	Middle East	4	80	1,600	Sora	6,000
						Kling AI	4,800
Jeddah	Saudi Arabia	Middle East	3	60	900	Sora	4,500
						Kling AI	3,600
Jerusalem	Israel	Middle East	4	80	1,600	Sora	6,000
						Kling AI	4,800
Kuwait City	Kuwait	Middle East	5	100	2,500	Sora	7,500
						Kling AI	6,000
Manama	Bahrain	Middle East	4	80	1,600	Sora	6,000
						Kling AI	4,800
Mecca	Saudi Arabia	Middle East	1	20	100	Sora	1,500
						Kling AI	1,200
Medina	Saudi Arabia	Middle East	1	20	100	Sora	1,500
						Kling AI	1,200
Muscat	Oman	Middle East	5	100	2,500	Sora	7,500
						Kling AI	6,000
Riyadh	Saudi Arabia	Middle East	4	80	1,600	Sora	6,000
						Kling AI	4,800
Sharjah	UAE	Middle East	2	40	400	Sora	3,000
						Kling AI	2,400
Tel Aviv	Israel	Middle East	6	120	3,600	Sora	9,000
						Kling AI	7,200
Middle East Total			**56**	**1,120**	**27,000**		**151,200**
Grand Total			**110**	**2,130**	**67,400**		**297,000**

Because the histograms are normalized, the denominator that often appears in the EMD definition is unity, and the objective directly yields the 1-Wasserstein distance. This formulation ensures that "moving" pixel mass between perceptually similar colors is penalized less than moving it between distant colors.

For hypotheses about direction (over- or understatement), we also compute raw signed differences (AI − GSV) and, where appropriate, aggregate across components within each attribute (e.g., summing vehicle types or static features) to maintain interpretability while preserving the sign of the distortion.

Table 5: Empirical Road Map: Hypotheses, Questions, Goals, Tests, and Nulls

Hypothesis	RQ Conceptual Goal	Primary Test	Key Null (H_0) / Alternative (H_a)
H1a (Europe)	RQ1 Origin-aligned fidelity (Sora)	Mann–Whitney U (1-tailed), clustered p-values	H_0: $d_{\text{Mah}}^{\text{Sora}} \geq d_{\text{Mah}}^{\text{Kling}}$; H_a: $d_{\text{Mah}}^{\text{Sora}} < d_{\text{Mah}}^{\text{Kling}}$
H1b (Middle East)	RQ1 Origin-aligned fidelity (Kling)	Mann–Whitney U (1-tailed), clustered p-values	H_0: $d_{\text{Mah}}^{\text{Kling}} \geq d_{\text{Mah}}^{\text{Sora}}$; H_a: $d_{\text{Mah}}^{\text{Kling}} < d_{\text{Mah}}^{\text{Sora}}$
H2a (Env./Struct.)	RQ2 Cross-origin bias magnitude (static)	Linear mixed model (Model × Region)	H_0: $\beta_3 = 0$; H_a: $\beta_3 \neq 0$
H2b (Activity)	RQ2 Cross-origin bias magnitude (dynamic)	Linear mixed model (Model × Region)	H_0: $\beta_3 = 0$; H_a: $\beta_3 \neq 0$
H3a (Static overstate)	RQ3 Static vs. dynamic fidelity	Wilcoxon signed-rank (paired, 1-tailed)	H_0: median(Δ_{static}) ≤ 0; H_a: median(Δ_{static}) > 0
H3b (Dynamic understate)	RQ3 Static vs. dynamic fidelity	Wilcoxon signed-rank (paired, 1-tailed)	H_0: median(Δ_{dynamic}) ≥ 0; H_a: median(Δ_{dynamic}) < 0
H4a (Env./Struct. var.)	RQ4 Variance compression (static)	Brown–Forsythe + permuted variance ratio	H_0: $\text{Var}_{\text{GSV}} = \text{Var}_{\text{AI}}$; H_a: $\text{Var}_{\text{GSV}} \neq \text{Var}_{\text{AI}}$
H4b (Activity var.)	RQ4 Variance compression (dynamic)	Brown–Forsythe + permuted variance ratio	H_0: $\text{Var}_{\text{GSV}} = \text{Var}_{\text{AI}}$; H_a: $\text{Var}_{\text{GSV}} \neq \text{Var}_{\text{AI}}$

6.2 Hypothesis Tests with Explicit Null and Alternative Statements

This section details how each hypothesis is operationalized and statistically evaluated. For every claim, we specify the relevant test, state the null and alternative hypotheses, and justify the choice of method in light of the data's paired structure, regional nesting, and non-normal distributions.

H1 (Origin-Aligned Fidelity). Within each region, one-tailed Mann–Whitney U tests compare the rank distributions of d_{Mah} for *Sora* and *Kling*. **Europe (H1a):** H_0: median($d_{\text{Mah}}^{\text{Sora}}$) $\geq$ median($d_{\text{Mah}}^{\text{Kling}}$); H_a: $<$ (more faithful *Sora*). **Middle East (H1b):** H_0: median($d_{\text{Mah}}^{\text{Kling}}$) $\geq$ median($d_{\text{Mah}}^{\text{Sora}}$); H_a: $<$ (more faithful *Kling*). P-values are cluster-robust at the district level.

H2 (Cross-Origin Bias Magnitude). For each attribute's signed distortion $\texttt{Distortion} = \text{AI} - \text{GSV}$, we estimate a linear mixed model:

$$\texttt{Distortion}_i = \beta_0 + \beta_1 \texttt{Model}_i + \beta_2 \texttt{Region}_i + \beta_3 (\texttt{Model} \times \texttt{Region})_i + u_{\text{district}(i)} + \varepsilon_i,$$

with random intercept $u_{\text{district}} \sim \mathcal{N}(0, \sigma_u^2)$ to capture intra-district dependence. H_0: $\beta_3 = 0$; H_a: $\beta_3 \neq 0$ (directional sign specified in advance).

H3 (Static vs. Dynamic Representation). Define

$$\Delta_{\text{static}} = (\text{AI} - \text{GSV})_{\text{vegetation}} + (\text{AI} - \text{GSV})_{\text{built}}, \quad \Delta_{\text{dynamic}} = (\text{AI} - \text{GSV})_{\text{pedestrian}} + (\text{AI} - \text{GSV})_{\text{vehicle}}.$$

Wilcoxon signed-rank tests (paired at the image level) evaluate:

$$H_0 : \text{median}(\Delta_{\text{static}}) \leq 0 \quad \text{vs.} \quad H_a : \text{median}(\Delta_{\text{static}}) > 0,$$
$$H_0 : \text{median}(\Delta_{\text{dynamic}}) \geq 0 \quad \text{vs.} \quad H_a : \text{median}(\Delta_{\text{dynamic}}) < 0.$$

False-discovery-rate correction (Benjamini–Hochberg, $\alpha = 0.05$) is applied across these tests.

H4 (Variance Compression). For each attribute, we compare group variances (GSV, *Sora*, *Kling*) using the Brown–Forsythe test. To increase robustness, we also compute variance ratios (e.g., $\text{Var}_{\text{AI}}/\text{Var}_{\text{GSV}}$) and obtain permutation-based p-values (10 000 label permutations). Significant reductions in AI variance relative to GSV support H4.

7 Results

7.1 Results for H1: Origin-Aligned Fidelity

Hypothesis 1 posits that each model is most faithful in the region that most closely matches its presumed training distribution: *Sora* in Europe (H1a) and *Kling AI* in the Middle East (H1b). We evaluate this claim by comparing model–GSV dissimilarities within each region using one-tailed Mann–Whitney U tests with cluster-robust p-values obtained via location-level label permutations (5,000 iterations). Two sets of results are reported. First, we examine a composite distance (the mean of all five embedding blocks) to test the hypothesis in aggregate. Second, we disaggregate by block (pedestrians, vehicles, vegetation, built surfaces, and color) to see whether origin-aligned fidelity emerges only for specific urban attributes. Table 6

summarizes the inferential statistics, while Table 7 reports medians (smaller is better) and differences to illustrate magnitude and direction.

The cluster-permutation p-values are 0.879 for Europe and 0.555 for the Middle East (Table 6, first panel), indicating no statistically significant difference between *Sora* and *Kling AI* when the five blocks are averaged. Taken together, these results suggest that H1a and H1b are not supported in the aggregate.

A different picture emerges when we examine individual attributes. In Europe, the pedestrian block yields a cluster-permutation p-value of 0.046, supporting H1a: *Sora*'s median pedestrian distance to GSV is 10.92 versus 13.57 for *Kling AI* (difference -2.65), and Cliff's $\delta = -0.185$ indicates a small-to-moderate effect size in the expected direction. No other European block is significant at $\alpha = 0.05$. In the Middle East, only the color block is significant ($p = 0.0008$), and here distances are smaller for *Kling AI*; *Sora*'s median color distance is 75.94 versus 50.22 for *Kling AI* (difference +25.72), supporting H1b for color. The remaining blocks show no reliable differences once clustering is accounted for.

Overall, the "origin-aligned fidelity" pattern does not hold uniformly but appears selectively: *Sora* is more faithful for pedestrians in Europe, and *Kling AI* is more faithful for color in the Middle East. All other blocks fail to reject the null after cluster correction. These results suggest that fidelity advantages may be attribute-specific rather than global.

Table 6: Cluster-Permutation Mann–Whitney Results for H1 (Origin-Aligned Fidelity)

Region	Block	U_{frame}	p_{naive}	U_{cluster}	p_{cluster}	n_{perm}	n_{clusters}	n_{frames}	Alt.	Direction	Sig.@0.05
Europe	composite	42.0	0.860684	42.0	0.878624	5000	8	16	Sora < Kling	No	
Middle East	composite	48.0	0.574947	48.0	0.554489	5000	10	20	Sora > Kling	No	
Europe	pedestrian	33,652,934.0	$< 10^{-15}$	13.0	0.046391	5000	8	20,478	Sora < Kling	**Yes**	
Europe	vehicle	39,915,413.0	2.94×10^{-188}	15.0	0.087183	5000	8	20,478	Sora < Kling	No	
Europe	vegetation	68,394,529.0	1.000000	43.0	0.924215	5000	8	20,478	Sora < Kling	No	
Europe	built	49,962,190.0	2.58×10^{-8}	25.0	0.278144	5000	8	20,478	Sora < Kling	No	
Europe	color	102,702,725.0	1.000000	64.0	1.000000	5000	8	20,478	Sora < Kling	No	
Middle East	pedestrian	78,672,644.0	1.000000	44.0	0.688462	5000	10	26,204	Sora > Kling	No	
Middle East	vehicle	77,614,637.5	1.000000	33.0	0.925215	5000	10	26,204	Sora > Kling	No	
Middle East	vegetation	NaN	NaN	17.0	0.979204	5000	10	26,204	Sora > Kling	No	
Middle East	built	89,255,539.0	6.31×10^{-11}	48.0	0.562088	5000	10	26,204	Sora > Kling	No	
Middle East	color	156,833,605.0	$< 10^{-15}$	98.0	0.000800	5000	10	26,204	Sora > Kling	**Yes**	

Notes: U_{frame} and p_{naive} are Mann–Whitney statistics computed at the frame level (often yielding spuriously small p-values due to large n). U_{cluster} and p_{cluster} are computed on location means with a location-level label permutation (5,000 permutations). "Alt. Direction" indicates the one-tailed alternative tested (*Sora* closer in Europe, *Kling* closer in the Middle East). The vegetation block in the Middle East shows NaNs at the frame level due to ties/zero variance; the cluster permutation remains valid. Boldface indicates $p_{\text{cluster}} < 0.05$ without multiple-testing correction.

7.2 Results for H2: Cross-Region Bias Magnitude

Hypothesis 2 asks whether each model's distortion grows larger in its "non-native" region—*Sora* in the Middle East and *Kling AI* in Europe—and does so in attribute-specific ways. Formally, we model the signed distortion (AI − GSV) for each block as a function of `Model` (*Sora* vs. *Kling*), `Region` (Europe vs. Middle East), and their interaction, estimating linear mixed models with random intercepts at the location level. The interaction coefficient β_3 captures how the *change* in *Sora*'s distortion across regions differs from the corresponding change in *Kling*'s distortion; that is,

$$\beta_3 \;=\; \big[\text{(Sora–Kling) in Middle East}\big] - \big[\text{(Sora–Kling) in Europe}\big].$$

Because distortions are defined as (AI − GSV), a positive β_3 means that *Sora*'s distortion becomes *more positive* (i.e., more overstatement or less understatement) in the Middle East than in Europe, relative to *Kling*; a negative β_3 indicates that *Sora*'s distortion becomes more negative (i.e., larger understatement) in the Middle East.

Table 7: Block-Level Median Distances (Smaller = More Faithful) and Differences

Region	Block	$\text{Median}_{\text{Sora}}$	$\text{Median}_{\text{Kling}}$	$\text{Median}_{\text{Sora}} - \text{Median}_{\text{Kling}}$
Europe	pedestrian	10.9185	13.5728	−2.6543
Europe	vehicle	6.2441	6.9899	−0.7458
Europe	vegetation	2.9641	2.4343	+0.5298
Europe	built	18.3883	18.6580	−0.2697
Europe	color	80.1129	50.5508	+29.5621
Middle East	pedestrian	10.3776	11.4709	−1.0933
Middle East	vehicle	5.5098	7.4754	−1.9656
Middle East	vegetation	2.2343	4.2847	−2.0504
Middle East	built	17.3817	17.8922	−0.5105
Middle East	color	75.9390	50.2233	+25.7157

Notes: Negative differences indicate smaller (better) distances for *Sora*; positive differences indicate *Kling* is closer. These medians support the two significant findings in Table 6: (i) Europe–pedestrians: *Sora* closer; (ii) Middle East–color: *Kling* closer. For Europe–pedestrians, Cliff's $\delta = -0.185$ suggests a small-to-moderate effect size in favor of *Sora*.

Table 8 reports β_3 estimates, standard errors, p-values, and 95% confidence intervals for each outcome. All models include 162,000+ frames from 71 distinct locations. The vast majority of effects are statistically significant at conventional levels, owing in part to the large sample size; interpretation should therefore foreground the direction and substantive magnitude of the coefficients.

For the environmental/structural dimensions (H2a), *Sora*'s vegetation distortion increases in the Middle East ($\hat{\beta}_3 = 0.020$, $p < 10^{-67}$), consistent with the hypothesis that *Sora* overstates environmental greenery more in its non-native region. By contrast, the built-surface block shows no significant interaction effect ($\hat{\beta}_3 \approx -0.031$, $p = 0.844$), and when vegetation and built are aggregated into a "static" index, the overall effect is null ($\hat{\beta}_3 \approx -0.008$, $p = 0.959$). In other words, the environmental component of H2a holds for vegetation but not for built surfaces, and the combination of the two washes out.

For the activity dimensions (H2b), *Sora* understates pedestrians more in the Middle East ($\hat{\beta}_3 = -1.576$, $p < 10^{-160}$), exactly as predicted—this is a large negative shift. However, the vehicle block moves in the opposite direction ($\hat{\beta}_3 = 1.939$, $p < 10^{-257}$): *Sora* overstates vehicles (or understates them less) in the Middle East relative to Europe. When pedestrians and vehicles are combined into a single "dynamic" index, the positive vehicle effect dominates, yielding a small but significant positive interaction overall ($\hat{\beta}_3 = 0.360$, $p \approx 3.0 \times 10^{-7}$). Thus, H2b is supported for pedestrians but contradicted for vehicles, producing a net positive shift when the two are aggregated.

Taken together, the evidence shows that "cross-region bias magnitude" is real but highly attribute-dependent. *Sora*'s distortions do not uniformly intensify in the Middle East; the direction and size of the interaction term vary by block. Vegetation and pedestrian counts behave as hypothesized, while built surfaces and vehicles do not. These mixed outcomes underline the importance of examining each interpretable block of *SimCityNet* separately rather than collapsing everything into a single scalar.

7.3 Results for H3: Static vs. Dynamic Representation

Hypothesis 3 evaluates whether generative models systematically *overstate* static elements (vegetation, roads, sidewalks, buildings; H3a) and *understate* dynamic elements (pedestrians and vehicles; H3b) relative to ground truth. We operationalize these claims with two signed difference indices per frame:

$$\Delta_{\text{static}} = (\text{AI} - \text{GSV})_{\text{vegetation}} + (\text{AI} - \text{GSV})_{\text{built}}, \qquad \Delta_{\text{dynamic}} = (\text{AI} - \text{GSV})_{\text{pedestrian}} + (\text{AI} - \text{GSV})_{\text{vehicle}}.$$

Table 8: Linear Mixed-Model Interaction Results for H2 (Model × Region) with Concise Labels

Outcome	Interaction	$\hat{\beta}_3$	SE	p-value	95% CI (low)	95% CI (high)	N_{frames}	N_{loc}
Vegetation	Sora×ME	0.019975	0.001144	3.00×10^{-68}	0.017732	0.022217	162,108	71
Built surfaces	Sora×ME	-0.030703	0.156107	0.844	-0.336667	0.275261	162,109	71
Static (Veg + Built)	Sora×ME	-0.007983	0.155643	0.959	-0.313037	0.297071	162,108	71
Pedestrians	Sora×ME	-1.576193	0.058130	6.56×10^{-162}	-1.690126	-1.462261	162,109	71
Vehicles	Sora×ME	1.938753	0.056507	5.58×10^{-258}	1.828001	2.049505	162,109	71
Dynamic (Ped + Veh)	Sora×ME	0.359922	0.070291	3.05×10^{-7}	0.222154	0.497691	162,109	71

Notes: Distortions are defined as (AI − GSV). The interaction term "Sora×ME" tests how the change in *Sora*'s distortion between Europe and the Middle East differs from the corresponding change for *Kling AI*. Positive values indicate that *Sora* becomes more positive (greater overstatement or lesser understatement) in the Middle East relative to Europe, compared with *Kling*; negative values indicate the opposite. All models include a random intercept for `location_id`. "Static" combines vegetation and built-surface distortions; "Dynamic" combines pedestrian and vehicle distortions.

One-tailed Wilcoxon signed-rank tests (paired at the image level) assess whether the median of each index is greater than zero (static overstatement) or less than zero (dynamic understatement). We adjust p-values across tests using Benjamini–Hochberg FDR at $\alpha = 0.05$.

Table 9 summarizes the results. For *Sora*, the static index is significantly positive (median $\Delta_{\text{static}} = 3.88$; $p_{\text{FDR}} < 0.001$), supporting H3a: *Sora* overstates static features. By contrast, *Sora*'s dynamic index is also positive (median $\Delta_{\text{dynamic}} = 4.20$), yielding a one-tailed $p_{\text{raw}} = 1.0$ for the "less than zero" alternative and thus failing to support H3b. The *Kling AI* results move in the opposite direction: its static index is negative (median $= -2.52$), contradicting H3a, and its dynamic index is strongly positive (median $= 9.30$), contradicting H3b. When pooling both models, neither hypothesis is supported; the pooled medians are small and positive for static (0.50) and clearly positive for dynamics (6.86), leading to non-rejections under the specified one-tailed tests.

The preponderance of p-values equal to 1.0 arises because the data move in the direction *opposite* to the one-tailed alternative (e.g., positive medians when testing for negative medians). These findings underscore the importance of directional specification: the Wilcoxon test is sensitive not only to magnitude but also to the hypothesized sign of the median difference. Overall, H3a is supported for *Sora* alone, while H3b is rejected for both models, indicating that neither *Sora* nor *Kling AI* understates dynamic activity; on the contrary, both tend to overstate it.

Table 9: Wilcoxon Signed-Rank Tests for H3 (one-tailed; FDR-corrected)

Model	Outcome	Alt.	N_{pairs}	Statistic	p_{raw}	Median Δ	p_{FDR}	Reject FDR?
Sora	Static overstatement	> 0	75,899	1.75×10^9	0.000	3.88	0.000	Yes
Sora	Dynamic understatement	< 0	75,899	2.35×10^9	1.000	4.20	1.000	No
Kling	Static overstatement	> 0	86,209	1.42×10^9	1.000	-2.52	1.000	No
Kling	Dynamic understatement	< 0	86,210	3.57×10^9	1.000	9.30	1.000	No
Combined	Static overstatement	> 0	162,108	6.37×10^9	1.000	0.50	1.000	No
Combined	Dynamic understatement	< 0	162,109	1.19×10^{10}	1.000	6.86	1.000	No

Notes: $\Delta_{\text{static}} = (\text{AI} - \text{GSV})_{\text{vegetation}} + (\text{AI} - \text{GSV})_{\text{built}}$. $\Delta_{\text{dynamic}} = (\text{AI} - \text{GSV})_{\text{pedestrian}} + (\text{AI} - \text{GSV})_{\text{vehicle}}$. One-tailed Wilcoxon tests use the signed differences directly; zeros are dropped. "Alt." indicates the direction of the alternative hypothesis (static: > 0; dynamic: < 0). p_{FDR} values are Benjamini–Hochberg adjusted across all six tests.

7.4 Results for H4: Do Synthetic Cityscapes Flatten Cross–City Diversity?

Hypothesis 4 asked whether generative outputs compress the cross–city variance of urban attributes relative to ground truth (GSV). Using city–level means for each attribute block, we compared variances across GSV, *Sora*, and *Kling*. Brown–Forsythe tests (median–based Levene) confirmed that variances differ among groups

for every attribute ($p < 10^{-8}$ in all cases). To evaluate the direction of the difference, we computed variance ratios $\mathrm{Var}_{\mathrm{AI}}/\mathrm{Var}_{\mathrm{GSV}}$ and obtained permutation p–values (10 000 label permutations) for the one–sided hypothesis of variance compression (AI variance < GSV variance). We then controlled the false discovery rate (Benjamini–Hochberg, $\alpha = 0.05$) across all pairwise comparisons.

Only the color–illumination block shows statistically significant compression: *Sora*'s variance is 0.45× that of GSV (permutation $p = 0.0003$, FDR–adjusted $p = 0.0015$), while *Kling*'s variance is 0.13× GSV's (permutation $p = 0.0001$, FDR–adjusted $p = 0.0010$). All other attributes exhibit the opposite pattern—variance is inflated, often dramatically. Vegetation variance is roughly 20× (Sora) and 45× (Kling) higher than GSV; pedestrian variance is 21× (Sora) and 32× (Kling); built and vehicle blocks also exceed unity. None of these inflated ratios yields a significant permutation p–value in the compressive direction (all FDR–adjusted $p = 1.0$).

In short, H4 receives partial support: generative models homogenize color palettes but not vegetation, built surfaces, pedestrians, or vehicles. Rather than universally flattening urban diversity, *Sora* and *Kling* narrow stylistic variation in color while amplifying dispersion in counts and proportions of visible objects and surfaces.

Table 10: Variance Compression Tests (City–Level Means): Ratios and p–values

Attribute	Comparison	Var(AI)	Var(GSV)	Ratio	Perm. p	BF F	BF p	N_{AI}	N_{GSV}	Perm. p_{FDR}	Reject FDR
vegetation	Sora vs. GSV	0.002342	0.000117	19.987	1.000000	22.5267	8.44×10^{-10}	91	83	1.0000	False
vegetation	Kling vs. GSV	0.005310	0.000117	45.315	1.000000	22.5267	8.44×10^{-10}	110	83	1.0000	False
built	Sora vs. GSV	59.402361	30.638910	1.939	0.999100	19.2434	1.47×10^{-8}	91	83	1.0000	False
built	Kling vs. GSV	117.377926	30.638910	3.831	1.000000	19.2434	1.47×10^{-8}	110	83	1.0000	False
color	Sora vs. GSV	70.741523	157.333582	0.450	0.000300	27.2235	1.56×10^{-11}	91	83	0.0015	True
color	Kling vs. GSV	20.871502	157.333582	0.133	0.000100	27.2235	1.56×10^{-11}	110	83	0.0010	True
pedestrian	Sora vs. GSV	11.812547	0.574675	20.555	1.000000	60.5779	1.34×10^{-22}	91	83	1.0000	False
pedestrian	Kling vs. GSV	18.654088	0.574675	32.460	1.000000	60.5779	1.34×10^{-22}	110	83	1.0000	False
vehicle	Sora vs. GSV	4.728900	3.515997	1.345	0.916808	24.6634	1.35×10^{-10}	91	83	1.0000	False
vehicle	Kling vs. GSV	13.684652	3.515997	3.892	1.000000	24.6634	1.35×10^{-10}	110	83	1.0000	False

8 Conclusion

This study introduced *SimCityNet* and used it to compare 297,000 frames from 2,130 AI-generated videos with 312,533 matched Google Street View photographs across 110 commercially active districts in Europe and the Middle East. The results show that fidelity is attribute- and region-dependent rather than uniform: Sora is typically closer to ground truth on pedestrian and vehicle counts, while Kling is closer on color–illumination; differences concentrate in a few blocks and are strongest for Middle Eastern color palettes. Variance analyses reveal no wholesale homogenization of urban form. Instead, models compress color variance while allowing equal or larger dispersion in vegetation, built surfaces, pedestrians and vehicles relative to the Google Street View baseline. These patterns complicate simple narratives about generative systems either "flattening" or "faithfully reproducing" cities; where models can economize visually, such as palettes and illumination, they do so, while dynamic content exhibits harder-to-predict shifts that vary by region.

For urban planning and allied public decision making, the implications are concrete. First, reliance on synthetic footage for scoping studies, visual impact assessments, or participatory design risks mis-characterizing activity levels and environmental amenities. A planner who infers comfortable walkability from synthetic videos that understate crowding or overstate greenery may over-allocate sidewalk width, street furniture, or cooling interventions in some corridors and under-allocate them in others, with distributional consequences for districts already sensitive to comfort, access, or shade. Conversely, exaggeration of vehicles could bias traffic calming or curb-management proposals toward more interventionist treatments than local conditions warrant. Because these distortions are selective and region-specific, the same modeling pipeline could pull European and Middle Eastern projects in opposite directions, quietly reinforcing geo-cultural asymmetries documented in broader audits of generative systems. (Baheti et al., 2024; Nair et al., 2024)

Second, synthetic media are increasingly discussed in regulatory and standards venues, and those debates should be read directly into urban analytics practice. The EU Artificial Intelligence Act imposes transparency duties for AI-generated or manipulated audiovisual content that could be mistaken for authentic, including

labeling and detection obligations for so-called deepfakes. (*Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)*, 2024) In parallel, NIST's AI Risk Management Framework (AI RMF 1.0) and its Generative AI Profile stress governance controls, data documentation, and evaluation procedures tailored to the specific risks of generative systems. (National Institute of Standards and Technology, 2023, 2025a,b) Provenance standards such as C2PA offer cryptographically verifiable content credentials, but their security notes and independent research on watermark removal caution that provenance alone is not a panacea: robust detection and auditing must accompany labeling, and designers should anticipate adaptive adversaries as well as benign post-processing that weakens marks. (Coalition for Content Provenance and Authenticity, 2024, 2023; Carlini et al., 2023; Amos et al., 2023; Peters, 2024) For planning agencies, the operational lesson is that synthetic visuals used in public processes should carry explicit provenance, qualified uncertainty statements, and, where feasible, third-party audits.

Third, the findings argue for a co-design posture in which urban researchers, planners, and AI experts jointly specify performance targets and audit routines. Street-level imagery has a record of supporting social-scientific inference, from vehicle-fleet demographics and urban perception to vegetation measurement and citywide mobility modeling. (Gebru, Krause, Wang, Chen, Aiden, & Fei-Fei, 2017; Dubey et al., 2016b; Li et al., 2015b; J. Zhang, Zheng, & Qi, 2017; J. Zhang et al., 2019) Building on that tradition, generative-video tools intended for professional planning should publish attribute-level error budgets calibrated against local benchmarks rather than relying on generic aesthetic fidelity. Concretely, agencies could require pre-deployment evaluations that report, by district, the signed and absolute errors for pedestrians, vehicles, vegetation, principal built surfaces, and color–illumination, along with variance ratios relative to matched ground truth. When the intended use is distributional—e.g., prioritizing shade in heat-vulnerable neighborhoods—error tolerances should be more stringent, the audit sample larger, and sensitivity analyses mandatory. Accountability frameworks from the broader AI literature likewise recommend independent review, traceable documentation, and clear lines of responsibility for mitigation when model behavior drifts or fails. (Raji et al., 2020; National Institute of Standards and Technology, 2023)

Finally, the people who will design the next generation of video generators will shape whether these systems become reliable instruments for public reasoning or merely persuasive renderers. The agenda is necessarily interdisciplinary. Model developers can target three immediate goals: region-aware calibration that treats attribute blocks as first-class metrics; variance-preserving regularization that protects cross-city diversity in addition to mean fidelity; and integrated provenance that couples durable watermarking or credentials with robust detectors and red-teaming. Standards bodies can turn these goals into testable requirements for domain deployments. Planning scholars and practitioners, in turn, can articulate task-level thresholds—how accurate a synthetic depiction must be for crowd estimation, traffic operations, heat-risk audits, or public engagement—and build shared benchmarks. When synthetic media enter the public record, what matters is not only whether a frame looks convincing, but whether its claim about the city is empirically warranted, reproducible, and transparent about its limits. The SimCityNet evidence suggests that such a standard is achievable if urban experts and AI experts design toward it together.

A Geographic Distribution of Study Neighborhoods

Table 12 provides a breakdown of the neighborhood studied in this analysis.

Region	Country	City	Neighborhood
Europe	Denmark	Copenhagen	Amager
Europe	Denmark	Copenhagen	Christianshavn
Europe	Denmark	Copenhagen	Frederiksberg
Europe	Denmark	Copenhagen	Hellerup
Europe	Denmark	Copenhagen	Indre By
Europe	Denmark	Copenhagen	Nørrebro
Europe	Denmark	Copenhagen	Nytorv
Europe	Denmark	Copenhagen	Ørestad
Europe	Iceland	Reykjavík	Álftanes
Europe	Iceland	Reykjavík	Ármúli
Europe	Iceland	Reykjavík	Grandi
Europe	Iceland	Reykjavík	Austurvöllur
Europe	Iceland	Reykjavík	Laugardalur
Europe	Iceland	Reykjavík	Miðborg
Europe	Iceland	Reykjavík	Norðurmýri
Europe	Iceland	Reykjavík	Reykjavík South Harbor
Europe	Ireland	Dublin	Ballsbridge
Europe	Ireland	Dublin	Clondalkin
Europe	Ireland	Dublin	Dublin Docklands
Europe	Ireland	Dublin	IFSC (International Financial Services Centre)
Europe	Ireland	Dublin	Naas Road
Europe	Ireland	Dublin	Sandymount
Europe	Ireland	Dublin	Smithfield
Europe	Ireland	Dublin	St. Stephen's Green and Grafton Street
Europe	Ireland	Dublin	Swords
Europe	Ireland	Dublin	Temple Bar
Europe	Luxembourg	Luxembourg City	Ban de Gasperich
Europe	Luxembourg	Luxembourg City	Bonnevoie
Europe	Luxembourg	Luxembourg City	City Center (Ville Haute)
Europe	Luxembourg	Luxembourg City	Howald
Europe	Luxembourg	Luxembourg City	Kirchberg
Europe	Luxembourg	Luxembourg City	Findel
Europe	Luxembourg	Luxembourg City	Plateau de Kirchberg
Europe	Netherlands	Amsterdam	Amstel Business Park
Europe	Netherlands	Amsterdam	Amsterdam City Center (Centrum)
Europe	Netherlands	Amsterdam	Amsterdam South (Zuid)
Europe	Netherlands	Amsterdam	Amsterdam West
Europe	Netherlands	Amsterdam	De Pijp
Europe	Netherlands	Amsterdam	Houthavens
Europe	Netherlands	Amsterdam	NDSM Werf (Amsterdam North)
Europe	Netherlands	Amsterdam	Schiphol
Europe	Netherlands	Amsterdam	Sloterdijk
Europe	Norway	Oslo	Aker Brygge and Tjuvholmen
Europe	Norway	Oslo	Bjørvika
Europe	Norway	Oslo	Central Business District
Europe	Norway	Oslo	Majorstuen
Europe	Norway	Oslo	Nydalen
Europe	Norway	Oslo	Økern
Europe	Norway	Oslo	Skøyen

Region	**Country**	**City**	**Neighborhood**
Europe	Switzerland	Zürich	City Center (Altstadt)
Europe	Switzerland	Zürich	Enge (District 2)
Europe	Switzerland	Zürich	Glattpark (Opfikon)
Europe	Switzerland	Zürich	Oerlikon (District 11)
Europe	Switzerland	Zürich	Zürich West (District 5)
Middle East	Bahrain	Manama	Adliya
Middle East	Bahrain	Manama	Juffair
Middle East	Bahrain	Manama	Manama Souq / Bab Al Bahrain
Middle East	Bahrain	Manama	Seef District
Middle East	Israel	Haifa	Carmel Center
Middle East	Israel	Haifa	Downtown (Lower City)
Middle East	Israel	Haifa	German Colony
Middle East	Israel	Haifa	Hadar
Middle East	Israel	Jerusalem	City Center (Jaffa Road)
Middle East	Israel	Jerusalem	Mahane Yehuda
Middle East	Israel	Jerusalem	Old City
Middle East	Israel	Jerusalem	Rehavia / German Colony / Talbiya
Middle East	Israel	Tel Aviv	Florentin
Middle East	Israel	Tel Aviv	Jaffa (Old City of Jaffa)
Middle East	Israel	Tel Aviv	Neve Tzedek
Middle East	Israel	Tel Aviv	Rothschild Boulevard
Middle East	Israel	Tel Aviv	Sarona
Middle East	Israel	Tel Aviv	Tel Aviv City Center
Middle East	Kuwait	Kuwait City	Al Salhiya
Middle East	Kuwait	Kuwait City	Hawally
Middle East	Kuwait	Kuwait City	Mirqab
Middle East	Kuwait	Kuwait City	Salmiya
Middle East	Kuwait	Kuwait City	Sharq District
Middle East	Oman	Muscat	Al Khuwair
Middle East	Oman	Muscat	Al Mouj (The Wave)
Middle East	Oman	Muscat	Muttrah
Middle East	Oman	Muscat	Qurum
Middle East	Oman	Muscat	Ruwi (CBD)
Middle East	Qatar	Doha	Al Sadd
Middle East	Qatar	Doha	Doha Corniche
Middle East	Qatar	Doha	Msheireb Downtown Doha
Middle East	Qatar	Doha	The Pearl-Qatar
Middle East	Qatar	Doha	West Bay
Middle East	Saudi Arabia	Jeddah	Al-Balad
Middle East	Saudi Arabia	Jeddah	Al-Hamra
Middle East	Saudi Arabia	Jeddah	Tahlia Street
Middle East	Saudi Arabia	Mecca	Central Mecca
Middle East	Saudi Arabia	Medina	Central Medina
Middle East	Saudi Arabia	Riyadh	Al Olaya
Middle East	Saudi Arabia	Riyadh	Al Sulaymaniyah
Middle East	Saudi Arabia	Riyadh	Diplomatic Quarter (Al Safarat)
Middle East	Saudi Arabia	Riyadh	King Abdullah Financial District
Middle East	United Arab Emirates	Abu Dhabi	Al Maryah Island
Middle East	United Arab Emirates	Abu Dhabi	Al Reem Island
Middle East	United Arab Emirates	Abu Dhabi	Downtown Abu Dhabi (Corniche)
Middle East	United Arab Emirates	Abu Dhabi	Saadiyat Island
Middle East	United Arab Emirates	Dubai	Bur Dubai
Middle East	United Arab Emirates	Dubai	Business Bay

Region	**Country**	**City**	**Neighborhood**
Middle East	United Arab Emirates	Dubai	Deira
Middle East	United Arab Emirates	Dubai	Downtown Dubai
Middle East	United Arab Emirates	Dubai	Dubai International Financial Centre (DIFC)
Middle East	United Arab Emirates	Dubai	Dubai Marina
Middle East	United Arab Emirates	Dubai	Jumeirah Lakes Towers (JLT)
Middle East	United Arab Emirates	Dubai	Palm Jumeirah
Middle East	United Arab Emirates	Sharjah	Al Majaz
Middle East	United Arab Emirates	Sharjah	Rolla

Table 12: Breakdown of Neighborhoods by Region, Country, City, and Neighborhood

Bibliography

Amos, J., Wal, D. J. V., et al. (2023). Synthid: Imperceptible watermarks for generative media. *arXiv*. Retrieved from https://arxiv.org/abs/2308.04222

Baheti, A., Gupta, S., Kayser, M., Ramasubramanian, K., Prabhu, V. U., & Batra, D. (2024). Beyond aesthetics: Measuring cultural competence in text-to-image generative models. In *Neurips 2024 datasets and benchmarks track*. Retrieved from https://arxiv.org/abs/2406.06655

Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., ... Caliskan, A. (2023). Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 acm conference on fairness, accountability, and transparency* (pp. 1493–1504).

Carlini, N., Tramèr, F., Jagielski, M., Passananti, J., Nasr, M., Wallace, E., ... Zhang, C. (2023). Invisible watermarks are provably removable. *arXiv*. Retrieved from https://arxiv.org/abs/2301.10226

Chen, L., Papandreou, G., Schroff, F., & Adam, H. (2017, June). Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*. Retrieved from https://arxiv.org/abs/1706.05587 doi: https://doi.org/10.48550/arXiv.1706.05587

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., & Girdhar, R. (2022, June). Masked-attention mask transformer for universal image segmentation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)* (pp. 1280–1289). New Orleans, LA, USA: IEEE. doi: https://doi.org/10.1109/CVPR52688.2022.00135

Coalition for Content Provenance and Authenticity. (2023). *Security considerations for c2pa implementers.* https://c2pa.org/learn/security-considerations-for-c2pa-implementers/. (Accessed 2025-07-28)

Coalition for Content Provenance and Authenticity. (2024). *C2pa specification, version 2.0.* https://spec.c2pa.org/. (Accessed 2025-07-28)

Dubey, A., Naik, N., Parikh, D., Raskar, R., & Hidalgo, C. A. (2016a). *Deep learning the city : Quantifying urban perception at a global scale.* Retrieved from https://arxiv.org/abs/1608.01769

Dubey, A., Naik, N., Parikh, D., Raskar, R., & Hidalgo, C. A. (2016b). Deep learning the city: Quantifying urban perception at a global scale. In *Eccv 2016*. Retrieved from https://arxiv.org/abs/1608.01769

Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L., & Fei-Fei, L. (2017, November). 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*(50), 13108–13113. Retrieved from http://dx.doi.org/10.1073/pnas.1700035114 doi: https://doi.org/10.1073/pnas.1700035114

Gebru, T., Krause, J., Wang, Y., Chen, J. D., Aiden, E., & Fei-Fei, L. (2017). 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*(50), 13108–13113. Retrieved from https://www.pnas.org/doi/10.1073/pnas.1700035114

Hall, M., Bell, S., Ross, C., Williams, A., Drozdzal, M., & Romero Soriano, A. (2024). Towards geographic inclusion in the evaluation of text-to-image models. In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*. doi: https://doi.org/10.1145/3630106.3658927

Hall, M., Ross, C., Williams, A., Carion, N., Drozdzal, M., & Soriano, A. R. (2024). *Dig in: Evaluating disparities in image generations with indicators for geographic diversity.* Retrieved from https://arxiv.org/abs/2308.06198

Kuaishou Technology. (2024). *Kuaishou launches full beta testing for "Kling AI" to global users.* Retrieved from https://ir.kuaishou.com/news-releases/news-release-details/kuaishou-launches-full-beta-testing-kling-ai-global-users-0 (Accessed 18 July 2025)

Li, X., Zhang, C., Li, W., Ricard, R., Meng, Q., & Zhang, W. (2015a). Assessing street-level urban greenery using google street view and a modified green view index. *Urban Forestry & Urban Greening*, *14*(3), 675–685.

Li, X., Zhang, C., Li, W., Ricard, R., Meng, Q., & Zhang, W. (2015b). Assessing street-level urban greenery using google street view and a modified green view index. *Urban Forestry & Urban Greening*, *14*(3), 675–685. Retrieved from https://www.sciencedirect.com/science/article/pii/S1618866715000348

Liu, Y., Zhang, K., Li, Y., Yan, Z., Gao, C., Chen, R., ... Sun, L. (2024). *Sora: A review on background, technology, limitations, and opportunities of large vision models.* Retrieved from https://arxiv.org/abs/2402.17177

Mogavi, R. H., Wang, D., Tu, J., Hadan, H., Sgandurra, S. A., Hui, P., & Nacke, L. E. (2024). *Sora openai's prelude: Social media perspectives on sora openai and the future of ai video generation.* Retrieved from https://arxiv.org/abs/2403.14665

Nadeem, M., Sohail, S. S., Cambria, E., Schuller, B. W., & Hussain, A. (2025). *Gender bias in text-to-video generation models: A case study of sora.* Retrieved from https://arxiv.org/abs/2501.01987

Naik, R., & Nushi, B. (2023). Social biases through the text-to-image generation lens. In *Proceedings of the 2023 aaai/acm conference on ai, ethics, and society* (pp. 786–808).

Nair, P., Odena, A., Guu, K., et al. (2024). Measuring geographic diversity of foundation model generations. *arXiv*. Retrieved from https://arxiv.org/abs/2407.02947

National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (ai rmf 1.0).* https://airc.nist.gov/AI_RMF. (Accessed 2025-07-28)

National Institute of Standards and Technology. (2025a). *Generative ai profile (nist ai 600-1), initial public draft.* https://airc.nist.gov/AI_RMF/GenAI. (Referenced in NIST SP 800-218A; accessed 2025-07-28)

National Institute of Standards and Technology. (2025b). *Mitigating risk from the malicious use of generative ai systems: Considerations for developers, providers, and deployers* (Tech. Rep. No. SP 800-218A). NIST. Retrieved from https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-218A.pdf

OpenAI. (2025). *Sora: Creating video from text.* Retrieved from https://openai.com/index/sora/ (Accessed 18 July 2025)

Peters, J. (2024). *Google open-sourced its watermarking tool for ai-generated text.* https://www.theverge.com/2024/10/23/24277873/google-artificial-intelligence-synthid-watermarking-open-source. (The Verge, 23 Oct 2024.)

Raji, I. D., Smart, A., White, R., Mitchell, M., Gebru, T., Hutchinson, B., ... Barnes, P. (2020). Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Acm facct* (pp. 33–44). Retrieved from https://dl.acm.org/doi/10.1145/3351095.3372873

Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations (ec) no 300/2008, (eu) no 167/2013, (eu) no 168/2013, (eu) 2018/858, (eu) 2018/1139 and (eu) 2019/2144 and directives 2014/90/eu, (eu) 2016/797 and (eu) 2020/1828 (artificial intelligence act). (2024). https://eur-lex.europa.eu/eli/reg/2024/1689/oj. (Official Journal of the European Union, 12 July 2024.)

Seiferling, I., Naik, N., Ratti, C., & Proulx, R. (2017). Green streets- quantifying and mapping urban trees with street-level imagery and computer vision. *Landscape and Urban Planning*, *165*, 93–101.

Sun, R., Zhang, Y., Shah, T., Sun, J., Zhang, S., Li, W., ... Ranjan, R. (2024). *From sora what we can see: A survey of text-to-video generation.* Retrieved from https://arxiv.org/abs/2405.10674

Wan, Y., Subramonian, A., Ovalle, A., Lin, Z., Suvarna, A., Chance, C., ... Chang, K. (2024). Survey of bias in text-to-image generation: Definition, evaluation, and mitigation. *arXiv preprint arXiv:2404.01030*.

Wang, Z., Li, H., & Rajagopal, R. (2020). *Urban2vec: Incorporating street view imagery and pois for multi-modal urban neighborhood embedding.* Retrieved from https://arxiv.org/abs/2001.11101

Zeng, A., Yang, Y., Chen, W., & Liu, W. (2024). The dawn of video generation: Preliminary explorations with sora-like models. *arXiv preprint arXiv:2410.05227*.

Zhang, J., Zheng, Y., & Qi, D. (2017). Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Aaai* (pp. 1655–1661). Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/10735

Zhang, J., Zheng, Y., & Qi, D. (2019). Deep spatio-temporal residual networks for citywide crowd flows prediction. *IEEE Transactions on Knowledge and Data Engineering*, *31*(9), 1544–1557. Retrieved from https://ieeexplore.ieee.org/document/8765623

Zhang, J., Zheng, Y., Qi, D., Li, R., Yi, X., & Li, T. (2017). *Predicting citywide crowd flows using deep spatio-temporal residual networks.* Retrieved from https://arxiv.org/abs/1701.02543

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). *The unreasonable effectiveness of deep features as a perceptual metric.* Retrieved from https://arxiv.org/abs/1801.03924

Zhang, Y., Wang, S., Chen, B., Cao, J., & Huang, Z. (2019). Trafficgan: Network-scale deep traffic prediction with generative adversarial nets. *IEEE Transactions on Intelligent Transportation Systems*, *22*(1), 219–230.

Zhou, K. Z., Choudhry, A., Gumusel, E., & Sanfilippo, M. R. (2024). *"sora is incredible and scary": Emerging governance challenges of text-to-video generative ai models.* Retrieved from `https://arxiv.org/abs/2406.11859`