Digital Twin-Based Urban Perception Modeling Through Real-Time Social Sensing and Multimodal Artificial Intelligence

Sathvik Gaba
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
contactsathvik@unr.edu

Abstract

Urban perception, the affective and evaluative character of streetscapes and public places, is central to planning, mobility, and human-centered urban analytics, yet it remains difficult to model because social media signals are sparse and biased, visual evidence is incomplete, and perceptions change over time. This manuscript develops a digital twin-based methodological framework for real-time urban perception modeling that integrates geotagged social sensing with multimodal artificial intelligence. The digital twin is formulated as a probabilistic spatiotemporal state layer that maintains estimates of perceived safety, comfort, liveliness, and related dimensions, together with their uncertainty, over street segments and places. Social posts are treated as biased observations of this latent state rather than direct measurements. Visual, textual, geospatial, and temporal encoders are fused through reliability-weighted attention, while a graph regularizer encodes spatial autocorrelation and street-network dependence. A calibration-aware learning objective separates model uncertainty from social sampling bias, and a sequential assimilation rule updates the twin as new streams arrive. The framework yields explicit decision rules for identifying perception deficits under uncertainty and provides diagnostic criteria for modality failure, temporal lag, and feedback contamination. The main contribution is not a single predictive architecture but an integrated representation, inference, and update logic that makes urban perception estimable under realistic data constraints. The analysis clarifies how multimodal fusion, spatial regularization, and uncertainty calibration interact, and it identifies implementation trade-offs for municipal digital twins, participatory sensing platforms, and human-centered urban analytics.

Keywords

digital twins; urban perception; social sensing; multimodal learning; uncertainty quantification; spatiotemporal modeling

1. Introduction

Digital twins are increasingly used to represent the state and behavior of complex urban systems in a form that can be queried, updated, and used for scenario analysis [1]. The real-time city agenda has similarly emphasized the value of continuous data streams for understanding urban dynamics, although it has also exposed the limits of treating cities as fully instrumented systems [2]. Volunteered geographic information and citizen sensing provide complementary observations of places that formal sensors rarely capture [3]. Social sensing extends this logic by using geotagged posts, text, images, and interactions as proxies for socioeconomic and experiential conditions [4]. A recent deep neural network study

combined geotagged social media posts and street view imagery to model urban streetscapes, illustrating both the promise and the representational complexity of linking social data to visual urban form [5].

Urban perception research has largely focused on visual assessment of streetscapes. Pairwise comparison studies have shown that perceived safety, beauty, and liveliness can be mapped at scale from street-level imagery [6]. Deep learning models have extended such judgments to global image collections [7], and Streetscore demonstrated that perceived safety can be predicted for large street networks [8]. Other work has examined high-level judgments of urban perception using learned visual representations [9]. These studies establish that streetscape appearance carries information about experiential qualities of places, but they usually produce static maps, rely on image-centric evidence, and treat perception as a property of visual form rather than a dynamic state that can be updated with heterogeneous observations.

Multimodal artificial intelligence offers a route beyond image-only inference. Transformer architectures provide a general mechanism for modeling long-range dependencies across tokens and modalities [10], while vision-language pretraining has shown that jointly learned image-text representations can transfer across recognition and retrieval tasks [11]. Multimodal machine learning further distinguishes representation, alignment, fusion, and translation problems, each with different failure modes [12]. In parallel, digital twin engineering has developed state estimation, simulation, and lifecycle management concepts for physical systems [13], and urban digital twin research has begun to connect these ideas to city-scale monitoring and decision support [14].

The central gap is that urban digital twins generally lack a probabilistic perception state. Existing perception models are often offline, image-centric, and evaluated as static prediction tasks, whereas social sensing is dynamic, noisy, and shaped by platform and demographic selection. The manuscript therefore develops a digital twin-based framework for real-time urban perception modeling that treats perception as a latent spatiotemporal field, social posts as biased observations, multimodal encoders as partial evidence, and the twin as an uncertainty-aware assimilation system. The contribution is an integrated formalization of representation, fusion, regularization, calibration, and update logic, rather than a claim of empirical performance.

2. Problem Formulation and Research Context

2.1 Urban perception as a latent spatiotemporal state

Urban perception is treated here as a latent, place-specific evaluative state that summarizes how people experience a street segment or public place along dimensions such as safety, comfort, beauty, and liveliness. This state is not identical to objective built-environment quality. A well-maintained street may still be perceived as unsafe if lighting, visibility, or social cues signal risk, and a visually plain street may be perceived as comfortable because of quietness or familiarity. Prior perception studies demonstrate that visual form and contextual cues carry measurable information about such judgments [6], and large-scale learning models show that these relationships can be estimated from imagery [7], [8], [9]. The modeling problem is nevertheless broader than image classification because perception varies over time, across user groups, and with transient conditions such as weather, events, or construction.

Let v index a spatial unit such as a street segment, grid cell, or named place, and let t index a time window. The perception state p(v,t) may be scalar for a single dimension or vector-valued for multiple dimensions. The state is maintained by the digital twin together with a posterior variance, provenance metadata, and an observation history. The twin therefore represents not only a best estimate but also the evidential basis for that estimate. This distinction matters because real-time social sensing produces uneven coverage: some places receive frequent posts, many receive none, and the absence of observations is not equivalent to neutral perception.

2.2 Social sensing and multimodal evidence

Real-time social sensing consists of geotagged posts, captions, comments, images, reactions, and associated metadata. These signals are attractive because they are voluminous, temporally resolved, and often expressive of experience, but they are not designed as measurement instruments. Their spatial precision varies, their language and imagery reflect platform cultures, and their frequency depends on user behavior. Multimodal learning provides tools for extracting information from text and images jointly, including attention-based fusion [10], contrastive vision-language representations [11], and structured multimodal fusion strategies [12]. Applied to urban perception, such models can align a post caption such as a complaint about lighting with a nighttime street image and a geospatial context containing road type, land use, and points of interest. The resulting representation is richer than any single modality, but it remains partial evidence about a latent state.

2.3 Digital twin representation

A digital twin for urban perception should not be conflated with a high-fidelity three-dimensional scene model. The relevant twin is an updatable estimation layer that binds spatial units, multimodal observations, learned representations, uncertainty estimates, and decision rules. Digital twin engineering emphasizes synchronization between a physical system and its virtual counterpart, including state estimation, model updating, and lifecycle management [13]. Urban digital twin research extends these ideas to city-scale monitoring and planning [14]. For perception modeling, synchronization means that the twin maintains a current posterior over p(v,t) and revises it as social and visual streams arrive. The twin can then support queries such as which street segments have a high probability of low perceived safety, where evidence is too weak to support intervention, and how a proposed lighting change might alter perception under an explicit simulation model.

3. Methodological Framework

3.1 State, observation, and selection bias

The perception state is generated by a context-dependent function of multimodal evidence and unobserved influences. A general state equation is

$$p(v,t) = g_\theta\big(z(v,t), a(v,t), h(v,t)\big) + \varepsilon(v,t)$$

where p(v,t) is the latent perception state at location v and time t, z(v,t) is a visual representation derived from street-level or user-generated imagery, a(v,t) is a social sensing representation derived from text and interaction metadata, h(v,t) is geospatial and temporal context, g_θ is a parameterized mapping, and ε(v,t) is a stochastic residual representing unmodeled influences. The equation is a modeling assumption rather than an empirical result: it states that perception is conditionally predictable from multimodal context but retains irreducible variation across individuals and situations.

Social posts do not observe p(v,t) directly. They provide noisy, selected signals. A minimal observation model is

$$y_i(t) = m\big(p(v_i, t)\big) + b_i(t) + \eta_i(t)$$

where y_i(t) is the signal extracted from post i at time t, v_i is its associated location, m is a measurement link, b_i(t) is a selection and expression bias term, and η_i(t) is observation noise. The bias term is critical. Large-scale social media studies are vulnerable to representativeness problems and behavioral distortions [15], platform and demographic selection effects are difficult to ignore [16], and volunteered geographic information quality varies with contributor, device, and context [17]. If b_i(t) is correlated with p(v_i,t), then treating y_i(t) as a direct label produces a biased perception surface even when the predictive model fits the observed posts well. The framework therefore requires explicit bias modeling, external anchors, or sensitivity analysis rather than assuming that social signals are unbiased measurements.

3.2 Multimodal encoding and reliability-weighted fusion

The twin receives multiple modality-specific representations. Visual encoders process street-level imagery and user photographs; text encoders process captions and comments; geospatial encoders represent road hierarchy, land use, points of interest, and visibility; temporal encoders represent time of day, season, and event context. These representations are fused through reliability-weighted attention:

$$z(t) = \sum_\{k = 1\}\text{^}K\ \alpha_k(t)\ W_k\ x_k(t),\ \alpha_k(t) = softmax(r_k(t))$$

where K is the number of modalities, x_k(t) is the representation from modality k, W_k is a modality-specific projection, r_k(t) is a reliability score, and α_k(t) is the resulting fusion weight. The reliability score can depend on geolocation precision, image resolution, text length, language coverage, temporal freshness, and source history. This design follows the general logic of attention-based and multimodal fusion models [10], [12], while using vision-language pretraining where paired image-text evidence is available [11].

The reliability weighting is not merely a performance optimization. It encodes the fact that modalities fail differently. A nighttime image may have low visibility, a short caption may be sarcastic, and a geotag may be coarse. If fusion weights are learned only from predictive loss, the model can become overconfident on a modality that is systematically biased in a subregion. The twin should therefore store both the fused representation and the modality-level reliability trace, so that a perception estimate can be audited when new evidence contradicts it.

3.3 Uncertainty-aware inference and calibration

Uncertainty is a first-class component of the perception state. For a scalar perception dimension, the posterior variance over repeated stochastic samples can be summarized as

$$Var(p) = \ (1/T)\sum_\{t = 1\}\text{^}T\ p_t\text{^}2\ -\ ((1/T)\sum_\{t = 1\}\text{^}T\ p_t)\text{^}2$$

where p_t is the t-th sample from an approximate posterior and T is the number of samples. This expression is a diagnostic summary, not a claim that sampling alone captures all uncertainty. In practice, the twin should distinguish aleatoric uncertainty, arising from genuine variation in perceptions and noisy observations, from epistemic uncertainty, arising from limited data and model ambiguity. Modern neural networks can be miscalibrated even when accurate on average [18], and Bayesian approximations provide one route to

representing model uncertainty in deep learning [19]. Calibration is essential because the twin is intended to support decisions under uncertainty. A perception estimate with a narrow interval that is empirically wrong is more dangerous than a wider interval that honestly reflects weak evidence.

3.4 Spatiotemporal graph regularization

Urban perception is spatially dependent. Adjacent street segments often share visual character, land use, and pedestrian experience, and named places may be connected by movement flows rather than Euclidean distance. A graph regularizer encodes this dependence:

$$L_g = \sum_\{(u,v) \in E\} w_uv \, ||p(u,t) - p(v,t)||\text{^}2$$

where E is the set of spatial or network edges, w_uv is an edge weight representing adjacency, network connectivity, or learned similarity, and the term penalizes large perception differences between connected units. Graph neural networks provide a scalable way to propagate information across such structures [20]. The regularizer should not be interpreted as forcing spatial uniformity. Its role is to stabilize estimates where observations are sparse while allowing genuine discontinuities, such as a change from a well-lit commercial street to an unlit service area, to be preserved through learned edge weights and residual terms.

3.5 Learning objective and digital twin assimilation

The model is trained with a composite objective that balances supervised evidence, spatial regularization, multimodal consistency, and parameter regularization:

$$L = L_sup + \lambda_1 \, L_g + \lambda_2 \, L_con + \lambda_3 \, ||\theta||_2\text{^}2$$

where L_sup is the supervised loss on available perception labels or anchor surveys, L_con encourages agreement between modality-specific predictions and the fused prediction, λ_1, λ_2, and λ_3 control the relative contributions, and θ denotes model parameters. The consistency term is important because a multimodal model can achieve low average error by relying on a single dominant modality while ignoring evidence that would be decisive in rare conditions. The regularizer discourages this failure mode without requiring that all modalities agree at every location.

After training, the digital twin assimilates new observations sequentially. A generic update is

$$\hat{p}_\{t+1\}(v) = \hat{p}_t(v) + K_t(v)(y_t(v) - H\,\hat{p}_t(v))$$

where p̂_t(v) is the current posterior estimate, y_t(v) is a new observation or batch of observations, H is an observation operator that maps the state to the observation space, and K_t(v) is a gain that depends on posterior variance, observation noise, and reliability. The update follows the logic of state assimilation used in digital twin engineering [1], [13], [14]. When observations are absent or unreliable, K_t(v) should shrink toward zero and the twin should retain its prior with increased uncertainty rather than inventing a change. When observations are precise and consistent, the gain should allow a meaningful revision.

3.6 Decision rule

The twin becomes useful for planning when it converts probabilistic estimates into decisions with explicit error tolerance. A generic decision rule is

$$D(v,t) = 1 \; if \; P(p(v,t) \geq \tau \mid y_\{1:t\}) \geq \gamma \; else \; 0$$

where D(v,t) is a binary indicator for whether a location exceeds a perception threshold τ, $y_{1:t}$ is the observation history, and γ is a confidence requirement. The threshold τ may represent a minimum acceptable level of perceived safety or comfort, while γ controls the risk of acting on weak evidence. The rule can be generalized to multiple dimensions or cost-sensitive decisions, but the core principle remains: a perception deficit should be flagged only when the posterior probability is sufficiently high and the underlying evidence is adequately calibrated.

4. Analytical Assessment and Technical Discussion

4.1 Identification under nonrandom social sensing

The observation model makes clear that identification depends on the relationship between bias and perception. If $b_i(t)$ is independent of $p(v_i,t)$ conditional on observed covariates, then aggregated social signals can provide useful evidence after noise reduction. If bias is correlated with perception, however, no amount of model capacity can fully recover the latent state without external anchors or strong assumptions. For example, posts about nighttime safety may be more frequent in places where residents already feel unsafe, producing a selection pattern that amplifies the observed relationship. Conversely, satisfied residents may post less often, attenuating visible problems. The framework therefore treats anchor surveys, participatory audits, and targeted street-level observations as calibration instruments rather than optional additions. Sensitivity analysis should report how conclusions change under plausible bias assumptions, and the twin should mark regions where social evidence is too selective to support strong claims.

4.2 Temporal latency and state consistency

Real-time social sensing is not instantaneous. Posts may be uploaded after an event, geotagged at a different time from the experience, and processed with variable latency. The twin should maintain event-time and processing-time distinctions so that a delayed observation is assigned to the correct temporal window. The assimilation gain provides a natural mechanism for handling staleness: an old observation should receive lower weight than a fresh one unless the underlying perception state is known to be stable. Uncertainty should increase when a location has not been observed for an extended period, because the twin's estimate is then increasingly dependent on the prior and on spatial regularization. This behavior is preferable to presenting a static perception map as if it were continuously valid.

4.3 Multimodal failure modes

Each modality introduces characteristic errors. Visual models can be confounded by weather, occlusion, camera angle, and image quality. Text models can misread sarcasm, local slang, and code-switching. Geolocation can be imprecise or misleading when users post about a place without being there. Fusion can compound these errors if reliability weights are learned from biased training data. The reliability-weighted fusion design addresses this risk only partially. The twin should also include modality dropout tests, provenance flags, and contradiction detection. When text and image evidence disagree strongly, the appropriate response is not necessarily to average them but to lower confidence, request additional evidence, or defer the decision.

4.4 Calibration and decision thresholds

The decision rule depends on calibrated probabilities. A model that ranks places correctly but outputs overconfident probabilities can still produce poor decisions when γ is high.

Calibration should therefore be evaluated on held-out places and time periods, not only on random splits that share the same spatial and temporal dependencies as the training data. The choice of τ and γ is a policy decision as much as a statistical one. A low γ may flag many places for inspection but increases false alarms; a high γ reduces false alarms but may miss emerging problems. The digital twin should make this trade-off explicit rather than hiding it inside a single predictive score.

4.5 Implementation trade-offs and boundary conditions

A municipal digital twin must operate under computational, privacy, and governance constraints. Continuous multimodal inference can be expensive, so the architecture should support tiered processing: lightweight screening for all locations and deeper multimodal analysis for locations with high uncertainty or high decision stakes. Privacy protection is also relevant because geotagged social data can reveal sensitive information about individuals and communities. Aggregation, minimization, and access control are necessary conditions for responsible deployment. The framework is best suited to perception dimensions that can be defined consistently across data sources and validated with external anchors. It is less suited to claims about individual experience or to places where social data are systematically absent. These boundary conditions do not invalidate the approach, but they determine when its outputs should be treated as decision support rather than definitive measurement.

5. Implications for Urban Perception Modeling

The framework shifts urban perception modeling from static prediction to state estimation. Static models answer what a place is likely to look like in a training sample; a digital twin asks what is currently believed about a place, why that belief is held, and how it should change when new evidence arrives. This shift has scientific value because it makes bias, uncertainty, and temporal dependence explicit. It also has practical value because planners and community organizations often need to know where evidence is weak, not only where a model predicts a low score. A perception map that distinguishes strong evidence from sparse inference is more actionable than a single shaded surface.

The multimodal component is equally important. Urban perception is not reducible to imagery, and social text often captures dimensions such as belonging, harassment, or social atmosphere that visual models miss. At the same time, social text is shaped by platform dynamics and cannot be treated as a neutral sensor. The proposed fusion logic keeps both claims in view: multimodal evidence improves coverage, while reliability weighting and bias modeling prevent the system from mistaking visibility for truth. The graph regularizer and assimilation rule then connect local observations to the broader spatial and temporal structure of the city.

6. Conclusion

This manuscript developed a digital twin-based framework for urban perception modeling through real-time social sensing and multimodal artificial intelligence. The framework represents perception as a latent spatiotemporal state, models social posts as biased observations, fuses visual, textual, geospatial, and temporal evidence through reliability-weighted attention, regularizes estimates over spatial and network graphs, and updates the twin through uncertainty-aware assimilation. Its central contribution is the integration of these elements into a coherent inference logic that can distinguish a confident perception estimate from a weakly supported one. The analytical discussion shows that identification depends on selection bias, that temporal latency must be handled explicitly, that multimodal failure modes

require provenance and contradiction checks, and that decision thresholds must be calibrated to policy costs. These properties make the framework relevant to municipal digital twins, participatory sensing platforms, and human-centered urban analytics, while also defining the conditions under which its outputs should be treated as provisional evidence rather than settled fact.

References

1. Batty, M. (2018). Digital twins. Environment and Planning B: Urban Analytics and City Science, 45(5), 817–820.

2. Kitchin, R. (2014). The real-time city? Big data and smart urbanism. GeoJournal, 79(1), 1–14.

3. Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. GeoJournal, 69(4), 211–221.

4. Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., Chi, G., & Shi, L. (2015). Social sensing: A new approach to understanding our socioeconomic environments. Annals of the Association of American Geographers, 105(3), 512–530.

5. Aman, J., Matisziw, T. C., Kim, J. B., & Luo, D. (2022, April). Sensing the city: Leveraging geotagged social media posts and street view imagery to model urban streetscapes using deep neural networks. In Proceedings of the International Conference on Computer-Aided Architectural Design Research in Asia (pp. 595-604). Association for Computer-Aided Architectural Design Research in Asia.

6. Salesses, P., Schechtner, K., & Hidalgo, C. A. (2013). The collaborative image of the city: Mapping the inequality of urban perception. PLOS ONE, 8(7), e68400.

7. Dubey, A., Naik, N., Parikh, D., Raskar, R., & Hidalgo, C. A. (2016). Deep learning the city: Quantifying urban perception at a global scale. In Computer Vision – ECCV 2016 (pp. 196–212). Springer.

8. Naik, N., Philipoom, J., Raskar, R., & Hidalgo, C. (2014). Streetscore: Predicting the perceived safety of one million streetscapes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 779–785). IEEE.

9. Ordóñez, V., & Berg, T. L. (2014). Learning high-level judgments of urban perception. In Computer Vision – ECCV 2014 (pp. 494–510). Springer.

10. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (Vol. 30, pp. 5998–6008). Curran Associates.

11. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning (Vol. 139, pp. 8748–8763). PMLR.

12. Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443.

13. Tao, F., Zhang, H., Liu, A., & Nee, A. Y. C. (2019). Digital twin in industry: State-of-the-art. IEEE Transactions on Industrial Informatics, 15(4), 2405–2415.

14. Li, D., Yu, W., & Shao, Z. (2021). Smart city based on digital twins. Computational Urban Science, 1(1), Article 4.

15. Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. Science, 346(6213), 1063–1064.

16. Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. Frontiers in Big Data, 2, Article 13.

17. Senaratne, H., Mobasheri, A., Ali, A. L., Capineri, C., & Haklay, M. (2017). A review of volunteered geographic information quality assessment methods. International Journal of Geographical Information Science, 31(1), 139–167.

18. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (Vol. 70, pp. 1321–1330). PMLR.

19. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the 33rd International Conference on Machine Learning (Vol. 48, pp. 1050–1059). PMLR.

20. Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In International Conference on Learning Representations.