MUSAJ
Minnesota Undergraduate Research & Academic Journal
MINNESOTA UNDERGRADUATE RESEARCH & ACADEMIC JOURNAL
VOLUME 8 ISSUE 1

Multi-View Learning with Multi-modal Fusion for Urban Walkability Prediction

Shilong Xiang[1], Janghyeon Lee[1], Yao-Yi Chiang[1]

Abstract

Urban walkability is central to public health, sustainability, and quality of life. Traditional assessments of walkability based on surveys and field audits are costly and difficult to scale. Recent computational approaches have used either satellite imagery, street view imagery, or population indicators, but capture only a partial view: satellites provide spatial layout, street views capture pedestrian conditions, and population data reflect activity patterns. We present WalkCLIP: a multimodal framework that fuses these complementary perspectives to predict urban walkability. Our approach adapts vision–language representations with walkability-focused captions, enhances them with a spatial aggregation module to capture neighborhood context, and integrates non-visual population dynamics embeddings. Evaluated on 4,660 locations in the Minneapolis–Saint Paul area, the model outperforms unimodal and multimodal baselines in both predictive accuracy and spatial consistency. These findings demonstrate that combining visual and behavioral signals provides a scalable and reliable method for assessing urban walkability.

Keywords

Urban Walkability, Multimodal Fusion, WalkCLIP

1 Introduction

Walkability refers to the ease with which people can reach daily destinations on foot while experiencing a safe and comfortable environment. Higher walkability is linked to improved public health, sustainability, and quality of life [5]. As a result, urban planners seek scalable tools to measure and improve walkable environments.

Measuring walkability, however, is challenging because it is multidimensional, involving quantitative (e.g., proximity to amenities), and qualitative factors (e.g., sidewalk conditions and perceived safety) [7, 2].

[1] Department of Computer Science, University of Minnesota

Walkability is also shaped by behavioral patterns and environmental factors, including pedestrian activity and air quality. Unfortunately, traditional assessments rely on field audits and surveys, which are costly and limited in coverage.

Recent advances in AI and geospatial data offer opportunities to address these limitations. Satellite imagery depicts urban structure from above [4], street-view imagery captures pedestrian conditions at eye level [3], and population dynamics data reflect neighborhood activity and environmental conditions [1]. Yet each source alone provides only a partial view.

This motivates multimodal approaches that integrate complementary data to provide a holistic view of the walking environment. Here, we present WalkCLIP: a framework that fuses satellite- and street-level visual features with population-dynamics embeddings to predict urban walkability.

2 Methods

WalkCLIP integrates satellite imagery, street view imagery, and population dynamics data to predict neighborhood walkability (Figure 1). Each modality is collected at the same geographic coordinates, ensuring that paired data share semantics across views. The framework has three stages:

Stage 1: Walkability-Aware Vision–Language Embeddings. CLIP [6] is a model that learns to associate images with text descriptions in a shared representation space. However, because CLIP is trained on general web data, it does not focus on walkability-related features. We adapt CLIP by fine-tuning it with paired satellite and street-view images alongside walkability-focused captions generated by GPT-4o. Each correct image-caption pair is treated as a positive example, while mismatched pairs serve as negatives, encouraging the model to learn features relevant to walkability. Satellite captions emphasize structural features such as block layout and street connectivity, while street-view captions describe pedestrian-level elements including sidewalks and building frontages. This process produces specialized embeddings for satellite (f_sat) and street view (f_street) imagery, aligned at their shared geographic coordinates.

Stage 2: Spatially-Aware Feature Enhancement (SAFE). Walkability depends not only on conditions at a specific location but also on the surrounding neighborhood. To capture this broader context, SAFE enriches each location's representation by incorporating information from nearby locations (Figure 2). It uses inverse-distance weighting so that closer neighbors have a stronger influence. When applied separately to satellite and street-view embeddings, SAFE produces spatialy enriched representations that reflect the character of the surroundings. street-view embeddings, SAFE produces spatially enriched representations that reflect the character of the surroundings.

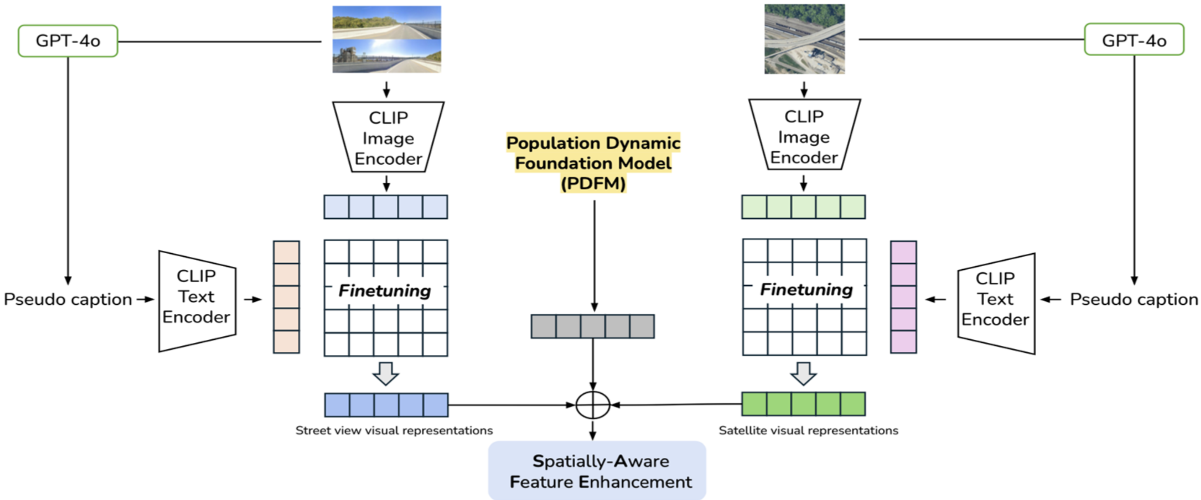

Figure 1. *The WalkCLIP framework*

Stage 3: <u>Multimodal Fusion and Prediction.</u> While visual features describe the built environment, walkability is also shaped by behavioral and environmental factors, such as how actively people use an area and local air quality. To incorporate these factors, we add embeddings from the Population Dynamics. Foundation Model (PDFM) [1], which encodes web search behavior and air quality in dicators. For each location, the spatially enriched visual embeddings are concatenated with the corresponding PDFM embedding to form the final representation. A multi-layer perceptron (MLP) regressor is then trained on this combined representation to predict walkability scores.

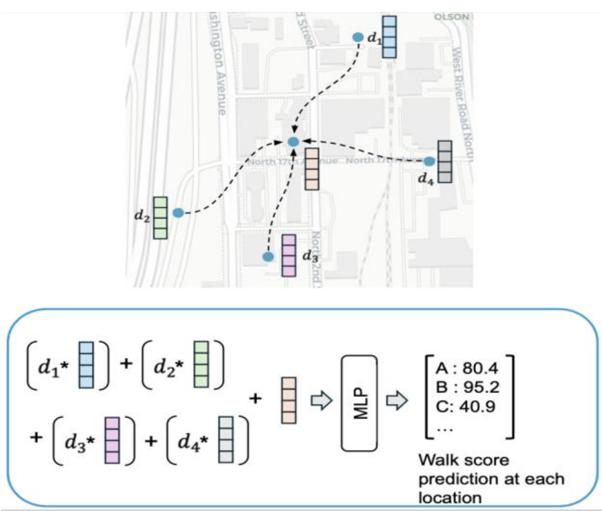

Figure 2 *Spatially Aware Feature Enhancement (SAFE)*

3 Results

3.1 Experiments

Setup. We evaluate on 4,660 geolocations in the Minneapolis–Saint Paul area, sampled from OpenStreetMap. Walk Score [2] is used as the downstream label. The dataset is split into 85% training and 15% test. We use 5-fold cross-validation, and performance is reported on the test set using R^2, RMSE, and sliced Wasserstein distance (SWD).

[2] https://www.walkscore.com

Table 1. Quantitative Results

	Methods	Test R^2 ↑	RMSE ↓	SWD ↓
(1)	CLIP (fine-tuned on streetview pairs)	0.350	11.620	4.860
(2)	CLIP (fine-tuned on satellite pairs)	0.538	9.795	3.442
(3)	PDFM	0.616	8.937	3.170
(4)	SS-CLIP (streetview + satellite)	0.597	9.154	2.936
(5)	SS-CLIP + PDFM	0.782	6.724	1.933
(6)	**WalkCLIP (SS-CLIP + PDFM + SAFE)**	**0.887**	**4.838**	**1.100**

Results. Table 1 shows that performance improves as the number of modalities and the spatial context increase. Among single modalities, PDFM performs best (R^2 = 0.616), indicating that behavioral context is important. Combining satellite and street views (SS-CLIP) improves accuracy (R^2 = 0.597), indicating that the two views are complementary. Adding PDFM embeddings (SS-CLIP + PDFM) yields a larger gain (R^2 = 0.782). WalkCLIP performs best overall (R^2 = 0.887).

3.2 Case Study

3.2.1 Modality Specific Patterns in Walkability Predictions

To illustrate how different modalities contribute to walkability prediction, we analyze three unseen locations (Figure 3). We compare a vision-only CLIP model that fuses satellite and street view representations, a context-only model based on population and environmental signals (i.e., PDFM), and the combined WalkCLIP model. These case studies highlight when each modality is informative, and how their combination provides balanced predictions.

Case 1 (44.95°N, 93.10°W): Redeveloped downtown core of St. Paul This location lies in the redeveloped downtown core of St. Paul, characterized by high-density housing, mixed land use, and upgraded pedestrian infrastructure. The satellite image (Figure 3a) shows compact urban blocks, while street view imagery highlights sidewalks, signalized intersections, and new building frontages. Vision-CLIP, which relies only on visual signals, assigns a very high score (92.4). strongly influenced by these visible cues of urban form. WalkCLIP predicts a slightly lower score (87.6; Walk Score: 90.0), as contextual information tempers the purely visual estimate and brings it closer to the proximity-based reference. By contrast, the context-only PDFM model assigns a much lower score (52.1), likely because population and activity signals lag recent redevelopment. This case illustrates how vision captures newly built pedestrian infrastructure, context-only inputs underrepresent rapid change, and WalkCLIP achieves balanced predictions.

Figure 3. Redeveloped Downtown Core of Minneapolis (paired satellite (left) and street view (right))

Case 2 (44.96°N, 93.19°W): Major commercial area in St. Paul This location lies along University Avenue, a major commercial area in St. Paul with light rail service and frequent public activity. The satellite image (Figure 4) highlights large blocks and surface parking, whereas street-view imagery shows wide roads, fencing, and sparse greenery, giving the area a car-oriented appearance. VisionCLIP, which relies solely on visual inputs, assigns a high score (85.7), emphasizing corridor density and development visible in imagery. WalkCLIP predicts a slightly lower score (80.1; Walk Score: 65.0), balancing visual impressions with contextual information. By contrast, the context-only PDFM model assigns a more moderate score (70.4), capturing transit accessibility and population activity that are not apparent in the imagery. This case illustrates how contextual inputs can temper visually optimistic predictions.

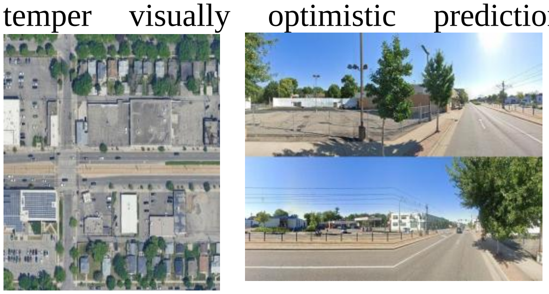

Figure 4. Major Commercial Area in St. Paul (paired satellite (left) and street view (right))

Case 3 (44.91°N, 93.29°W): Residential neighborhood in Southwest Minneapolis This location is a residential neighborhood in Southwest Minneapolis characterized by higher-density housing, tree-lined streets, and regular pedestrian activity. The satellite image (Figure 5) shows a connected street grid with greenery. Street-view imagery depicts continuous sidewalks, short block lengths, and closely spaced homes that facilitate everyday walking. VisionCLIP predicts a high score (83.6), highlighting the visibility of pedestrian infrastructure and the compact street network. WalkCLIP and the context-only PDFM model also predict high scores (WalkCLIP: 82.0; PDFM: 79.5; Walk Score: 80.0), resulting in strong alignment across modalities. This case illustrates that when both visual and contextual cues indicate pedestrian accessibility, all models converge on consistent high predictions.

Figure 5. Residential Neighborhood in Southwest Minneapolis (paired satellite (left) and street view (right))

3.2.2 Different Perspectives on Walkability

Walk Score primarily measures proximity to amenities, thereby overlooking the qualitative aspects of the pedestrian environment that visual and behavioral data can capture. For example, a high density of destinations does not guarantee the presence of safe, comfortable, or aesthetically pleasing infrastructure, such as sidewalks, crosswalks, and greenery. This section explores cases where WalkCLIP's predictions diverge from Walk Score. These disagreements should not be interpreted as model errors; rather, they highlight the different dimensions of walkability each method emphasizes. By analyzing these specific instances, we aim to illustrate how a multimodal approach provides a holistic assessment by capturing the qualitative, experiential aspects of walkability that proximity-based scores alone cannot. (Figure 5).

Case 1 (45.01°N, 93.26°W): **Industrial Zone** This location lies within an industrial zone and has a Walk Score of 41, indicating limited access to nearby amenities. The satellite view (Figure 6) shows large warehouses, expansive surface parking, and wide arterial roads. Street-level imagery highlights truck access points, construction activity, and the lack of sidewalks or crosswalks. WalkCLIP assigns a higher score of 73, likely because it interprets the concentration of large building footprints as a signal of urban density. In this case, Walk Score provides a more reliable assessment, since the dense industrial area does not translate into safe or functional pedestrian conditions.

Figure 6. *Industrial Zone*

Case 2 (44.96°N, 93.13°W): Cemetery This location lies within a cemetery but has a Walk Score of 82, reflecting its proximity to a nearby commercial district. The satellite image (Figure 7) shows a large, green parcel separated from the surrounding street grid. Street-view imagery reveals orderly, well-maintained grounds with narrow internal roads but no sidewalks or crosswalks for everyday connectivity. WalkCLIP predicts a lower score of 53. Although it identifies the park-like setting, the reduced score reflects the absence of functional pedestrian infrastructure. Both perspectives are reasonable: Walk Score emphasizes proximity to neighborhood destinations, while WalkCLIP reflects the local visual environment, underscoring the nuance of spaces that appear green and structured but are not functionally walkable.

Figure 7. *Cemetery*

Case 3 (44 .94°N, 93.05°W): Highway Bridge This location is on a highway bridge and has a Walk Score of 13, reflecting very limited proximity to everyday destinations. The satellite image (Figure 8) shows the bridge spanning rail lines and industrial land, with no nearby services. Street-view imagery highlights a multi-lane highway and reveals a fenced pedestrian path. WalkCLIP predicts a somewhat higher score of 41. Although it recognizes the presence of this dedicated path, the lower overall score reflects the surrounding car-oriented environment and lack of connected pedestrian infrastructure. Both perspectives are plausible: Walk Score emphasizes the absence of nearby amenities, whereas WalkCLIP highlights that some pedestrian facilities exist at the site.

Figure 8. *Highway Bridge*

4 Conclusion

We introduced WalkCLIP, a multimodal framework that integrates satellite imagery, street-view imagery, and population dynamics to predict urban walkability. By combining walkability-aware vision–language embeddings with spatial aggregation and contextual fusion, WalkCLIP outperforms both unimodal and multimodal baselines in predictive accuracy and spatial consistency. Beyond predictive gains, WalkCLIP demonstrates that visual and behavioral signals together provide a more comprehensive view of walkability than any single data source alone.

Several directions for future work emerge from this study. First, extending WalkCLIP to additional cities and regions would test its generalizability beyond the Minneapolis–Saint Paul area. Second, incorporating temporal data, such as seasonal imagery or time-varying population patterns, could capture how walkability changes over time. Third, integrating additional modalities such as LiDAR point clouds or noise level data may further enrich the representation of pedestrian environments. Finally, combining WalkCLIP predictions with urban planning tools could enable actionable recommendations for improving walkability in underserved neighborhoods.

References

[1] Mohit Agarwal, Mimi Sun, Chaitanya Kamath, Arbaaz Muslim, Prithul Sarker, Joydeep Paul, Hector Yee, Marcin Sieniek, Kim Jablonski, Yael Mayer, et al. General geospatial inference with a population dynamics foundation model. arXiv preprint arXiv:2411.07207, 2024.

[2] Ann Forsyth. What is a walkable place? the walkability debate in urban design. Urban design international, 20(4):274–292, 2015.

[3] Youngok Kang, Jiyeon Kim, Jiyoung Park, and Jiyoon Lee. Assessment of perceived and phys- ical walkability using street view images and deep learning technology. ISPRS International Journal of Geo-Information, 12(5):186, 2023.

[4] Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. Satclip: Global, general-purpose location embeddings with satellite imagery. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, pages 4347–4355, 2025.

[5] Kevin M Leyden, Michael J Hogan, Lorraine D'Arcy, Brendan Bunting, and Sebastiaan Bierema. Walkable neighborhoods: Linkages between place, health, and happiness in younger and older adults. Journal of the American Planning Association, 90(1):101–114, 2024.

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PMLR, 2021.

[7] Brian E Saelens and Susan L Handy. Built environment correlates of walking: a review. Medicine and science in sports and exercise, 40(7 Suppl):S550, 2008.