Assessing Street-Level Emotional Perception in Urban Regeneration Contexts Using Domain-Adapted CLIP

Liyang Chu and Keting Zhou *

Posted Date: 3 February 2026

doi: 10.20944/preprints202602.0146.v1

Keywords: urban regeneration; street-level emotional perception; domain-adapted CLIP; vision–language model; street view image (SVI)

 doi:10.20944/preprints202602.0146.v1

Article

Assessing Street-Level Emotional Perception in Urban Regeneration Contexts Using Domain-Adapted CLIP

Liyang Chu [1] and Keting Zhou [2,*]

[1] Manchester School of Architecture, Manchester M1 7ED, UK
[2] Harvard University, Cambridge 02138, USA
* Correspondence: ketingzhou@alumni.harvard.edu

Abstract

As urban regeneration goals shift from physical improvement to pedestrian-level experience and emotional perception, existing assessment methods struggle to describe the emotional responses associated with renewed street environments. This paper proposes a framework for street-level emotional perception and analysis within the context of urban regeneration, enabling the computational representation of emotional perception based on Street View Image (SVI) and a vision-language model (VLM). The study constructs a six-dimensional emotion perceptual framework encompassing Comfort, Vitality, Safety, Oppressiveness, Nostalgia, and Alienation, and uses a lightweight domain-adapted Contrastive Language-Image Pre-training (CLIP) model to infer emotional perceptions from SVI. Building upon this, a dual-axis evaluation framework is introduced to structure and interpret basic spatial experience and regeneration-related perception. Using the Yuyuan Road and Wuding Road areas in Shanghai as a case study, the study combines emotional perception results with street-level spatial analysis, proposing a scalable and interpretable analytical method for diagnosing urban regeneration outcomes and supporting emotion-informed spatial interventions.

Keywords: urban regeneration; street-level emotional perception; domain-adapted CLIP; vision–language model; street view image (SVI)

1. Introduction

Nowadays, urban development has gradually entered a stage dominated by stock renewal, and the effectiveness of renewal is no longer only focused on the improvement of the physical environment, but increasingly depends on whether the renewal space can stimulate positive emotional resonance and continuous emotional experience. Urban regeneration involves not only the optimization of spatial form and functional structure, but also the emotional reshaping process related to social memory, place attachment, and identity. Emotional perception accumulates through long-term interactions between people and space and, as an intrinsic mechanism, persistently influences the urban regeneration process [1–3]. Existing urban research and environmental psychology theories show that the evaluation of urban space not only depends on the physical form, but also on the emotional experience, cognitive schema and memory accumulation formed through the everyday use of urban space by individuals [1,4–7]. These studies collectively indicate that urban spatial imagery is constructed in continuous human-space interaction, rather than determined by static physical structures alone.

In recent years, Street View Images (SVIs) has become an important data source for studying the urban perception due to their continuous coverage, high spatial resolution, and the ability to

 doi:10.20944/preprints202602.0146.v1

represent the urban environment from a pedestrian perspective [8]. Studies based on SVIs and deep learning methods have shown that visual elements such as greenery and street-scale characteristics are stably correlated with spatial perception such as safety, comfort, and vitality [9,10]. However, these studies mostly focus on the quantitative description of visual elements or spatial physical properties, and rarely explain how these features are perceived and translated into specific emotional experiences at the pedestrian level. This limitation becomes even more prominent in the context of urban regeneration, because improving people's daily spatial experience and emotional attachment is an important dimension to measure the effectiveness of urban renewal.

In the quantitative study of emotional perception in urban space, traditional methods mainly rely on textual data such as social media texts and survey interviews to infer the distribution of urban emotions [11,12]. Although such methods can directly reflect subjective expressions, the lack of stable spatial anchors in textual information makes it difficult to map the results to specific spatial environments. With the development of multimodal artificial intelligence and vision-language model (VLM), emotional perception results are closer to human perception in real spatial environments, making large-scale image data processing and emotional perception quantification possible [10,13] .

However, existing research on emotional perception still has limitation in its application to actual urban regeneration projects, lacking support for refined spatial diagnosis. Current methods rely heavily on qualitative approaches such as interviews and case studies, making it difficult to produce continuous and comparable street-level spatial evaluation results [1–3]. Furthermore, some studies are constrained by specific regions or historical contexts, resulting in limited spatial generalizability of their conclusions [14,15].

In response, this paper proposes a street-level emotional perception analysis framework for the urban regeneration, focusing on the following three key questions:

1. How are visual elements of street space captured in SVIs perceived and translated into different emotional experiences?
2. How can a scalable and interpretable method for spatial emotional perception be constructed based on SVIs in the context of urban regeneration?
3. How can the results of spatial analysis based on emotional perception be applied to support the street-level diagnosis and intervention in urban renewal?

In response to the research questions, this paper constructs a multi-dimensional emotional perception analysis method based on lightweight domain adaptation of CLIP using Street View Images (SVIs). Through a six-dimensional emotional perception framework and a dual-axis evaluation analysis, the relative position of street-level basic spatial experience and regeneration-context perception is described. Meanwhile, the perception results are spatially mapped at the neighborhood level, providing an interpretable and scalable spatial diagnosis framework for urban regeneration.

2. Related Work

2.1. Contrastive Language-Image Pre-training (CLIP) model

The development of VLM provides a new methodological basis for the study of urban spatial perception beyond traditional supervised learning approaches designed for specific tasks. Unlike traditional visual models based on fixed categories and relying on manual annotation [16,17], the CLIP model establishes a joint representation between visual content and natural language semantics by pre-training on large-scale image-text pairs. Therefore, it can perform zero-shot analysis of images through language prompts without task-specific annotation and training [18].

In early urban research, the application of CLIP model mainly focused on the identification and analysis of the built environment and functional attributes. StreetCLIP [19] improved the cross-domain zero-shot generalization capability of SVI geolocation through synthetic text descriptions, and confirmed the feasibility of street-view analysis without annotating data. On this basis, UrbanCLIP [20] constructed a prompt engineering framework for urban function inference. Its zero-

 doi:10.20944/preprints202602.0146.v1

shot performance outperformed traditional supervision methods in cross-city migration scenarios, reflecting CLIP's generalization ability to urban spatial heterogeneity. However, the above research are still limited to urban functions and geographical attributes, without engaging with people's emotional perception of the urban environment.

In parallel, EmotionCLIP [21] enhanced general image emotion recognition. Its approach was to construct complex emotional states by manipulating the latent representations of basic emotions. However, its generalized architecture for general visual content was difficult to accurately capture the street-level emotional perception. In contrast, UP-CLIP further performed domain adaptation for urban perception semantics, which enabled the model to more accurately capture urban perceptual features in SVIs by constructing domain-specific perceptual semantic datasets and supervised fine-tuning of CLIP [22].

Overall, existing CLIP-based studies mainly concentrate on improving model performance in specific semantic tasks, while paying limited attention to integrating model outputs with spatial analysis for decision support in the context of urban regeneration. To address this gap, this study adapts CLIP to enhance its alignment with textual descriptions of regeneration-oriented emotional semantics. Based on the adapted model, emotional inference is conducted on SVI and subsequently combined with spatial analysis methods to support street-level diagnosis of urban regeneration.

2.2. Spatial Visual Elements Influencing Emotional Perception

Emotional responses of urban spaces are non-arbitrary but closely related to perceptible visual elements in the street environment. Early research in environmental psychology and urban design has shown that spatial characteristics such as street greenery, spatial enclosure, interface continuity, and transparency significantly influence subjective judgments of safety, comfort, and vitality [4,23]. With the development of SVIs and computer vision techniques, related studies have further verified the correspondence between these visual elements and emotional perception at the street level. Research based on SVIs showed that streets with high visibility of greenery and transparent interfaces were more likely to evoke positive perceptions of pedestrian, while streets with enclosed spaces and cluttered interfaces were often associated with negative emotions such as anxiety and oppression [24,25].

In recent years, the development of SVIs acquisition and machine learning technologies enabled researchers to explore a broader range of emotional dimensions. It has been found that a higher proportion of natural elements in the visual environment tends to correlate with more positive emotional responses. Conversely, excessive use of artificial elements such as buildings, walls, and fences often corresponds to lower emotional perception scores [26]. Further research has revealed that visual indicators such as the green view index, spatial openness, and enclosure index not only have significant predictive power for positive emotions but are also closely related to negative emotions such as anxiety and discomfort [27]. Furthermore, experiments incorporating physiological feedback have shown that visual elements such as green visibility, sky openness, sidewalk visibility, and building facade proportions can induce observable emotional arousal and changes in emotional tendency. This providing evidence for physiological correlation between the visual environment and emotional state [28].

In summary, there is a stable correlation between the street visual environment and emotional perception. In urban regeneration, the emotional perception of street spaces is typically influenced by a combination of multiple perceptible visual cues. Therefore, these repeatedly validated visual elements provide a theoretical reference for subsequent street-level emotional perception analysis and a basis for constructing an emotional perception framework.

2.3. Conceptualization of Emotional perception in the Context of Urban Regeneration

Early psychological research laid the foundation for the study of emotion models around the emotional response of individuals to stimuli. Among them, Russell's [29] Circumplex Model of Affect characterizes emotional states in a continuous space composed of valence and arousal, which

 doi:10.20944/preprints202602.0146.v1

provides a key theoretical framework for subsequent researchers to understand the structural dimension of emotions. However, such models primarily focus on individual psychological mechanisms and pay limited attention to the relationship between emotional experience, spatial forms, place meaning, and socio-cultural contexts. As a result, they have limitations in explaining emotional perception in complex urban environments.

As research perspectives expand into urban space and environmental perception, scholars increasingly adopt spatially oriented emotional responses to emphasize the emotional perception as a result of human-environment interaction [30]. Studies based on the MIT Place Pulse framework operationalize urban perception into dimensions such as safe, beautiful, lively, and depressing, and quantify public perception based on SVIs and machine learning methods [24,25,30]. Although some studies apply ranking or continuous ratings, urban perception is still often modeled along a single valence dimension, limiting the representation of coexistence and interaction of multiple emotions.

A growing body of research shows that spatial emotional experience is inherently multidimensional and continuous rather than a simple positive – negative dichotomy [31]. In urban environments, different spatial elements may simultaneously evoke multiple emotional responses, with context-dependent and nonlinear effects [28]. Huang et al. [27] further demonstrate that positive and negative emotions are not mutually exclusive, but may coexist and jointly shaping overall spatial perception.

In the context of urban regeneration, this multi-dimensional coexistence and context dependence are more prominent. The same spatial intervention may elicit complex emotional responses while improving environmental quality, and the results are influenced by factors such as spatial historical attributes, renewal methods, and functional transformation [15]. Especially in the process of historical street regeneration, nostalgia has pronounced dual nature. This dimension may reflect positive recognition of historical continuity and local memory, while also relating to the perception of spatial decay or abandonment [14,15]. In addition, urban regeneration, as a socio-cultural process of reconstructing collective memory, identity, and place meaning, exacerbates the complexity of emotional perception in renewal space [2,32].

2.4. Urban Spatial Evaluation in the Context of Urban Regeneration

In practice, evaluation tools for measuring the effectiveness of urban regeneration still mainly rely on physical and functional-oriented indicators, such as green space ratio and building density. These indicators have played a role in urban development evaluations [23,33], but they often struggle to reflect the real impact of renewal measures on residents' daily experiences. Existing research indicates that while some regeneration projects achieve improvements at the morphological or facility level. Nevertheless, they may not receive positive feedback in terms of emotional perception, the regenerated space may still be perceived as unattractive, or even cause alienation or discomfort [34,35]. This inconsistency reveals the limitations of traditional evaluation methods in explaining the effects of urban regeneration.

In the fields of urban design and environmental psychology, macro-structural indicators are no longer sufficient to fully explain the impact of spatial interventions on daily life. An increasing number of scholars are turning their attention to pedestrian-level perceptual experience, regarding it as a key criterion for measuring the effectiveness of spatial transformation [33,35]. As a primary spatial carrier of urban regeneration, street quality is usually identified and evaluated through immediate perception, such as ease of walking and lingering, support for daily activities and social interaction, and clear and safe order [23,35,36]. Therefore, incorporating perceptual and experiential dimensions into the urban regeneration evaluation system has become an important trend [10,33,37].

Following this perspective, this article defines "good urban regeneration space" as a spatial state centered on pedestrian-level experience. Its assessment is not based on whether a single physical indicator meets the standards. However, it relies on whether the regenerated street can be continuously perceived as comfortable, safe, and vibrant in daily use, and whether it maintains or reshapes place meaning and social connections during the renewal process. Therefore, this paper

 doi:10.20944/preprints202602.0146.v1

takes the perceptible emotional experience of pedestrians in street spaces as the basis for evaluation and constructs a street-level emotional perception framework in the context of urban regeneration.

3. Methodology

3.1. Research Framework

This study focuses on the emotional perception of street spaces and spatial diagnosis in the context of urban regeneration, constructing a research framework that integrates SVIs, multimodal artificial intelligence, and spatial analysis methods (Figure 1). The overall research process consists of four interconnected stages.

The first stage focuses on street view data construction and spatial unit definition. Street view sampling points are generated along the street network, and a street view database is established using street segments as the basic analysis units. The second stage constructs a multidimensional street-level emotional perception framework for urban regeneration by synthesizing and screening emotional concepts from the literature on urban design, environmental psychology, and urban regeneration. The third stage addresses the adaptation and application of the CLIP model in the urban regeneration context. Through lightweight domain adaptation and comparative strategy testing, the model's ability to capture emotional semantics in SVIs is enhanced, enabling emotional inference based on predefined emotional dimensions and text prompts. The fourth stage organizes the inferred emotional results at the street level and conducts spatial analysis using mapping and GIS methods, translating image-based emotional inference into street-level spatial diagnosis.

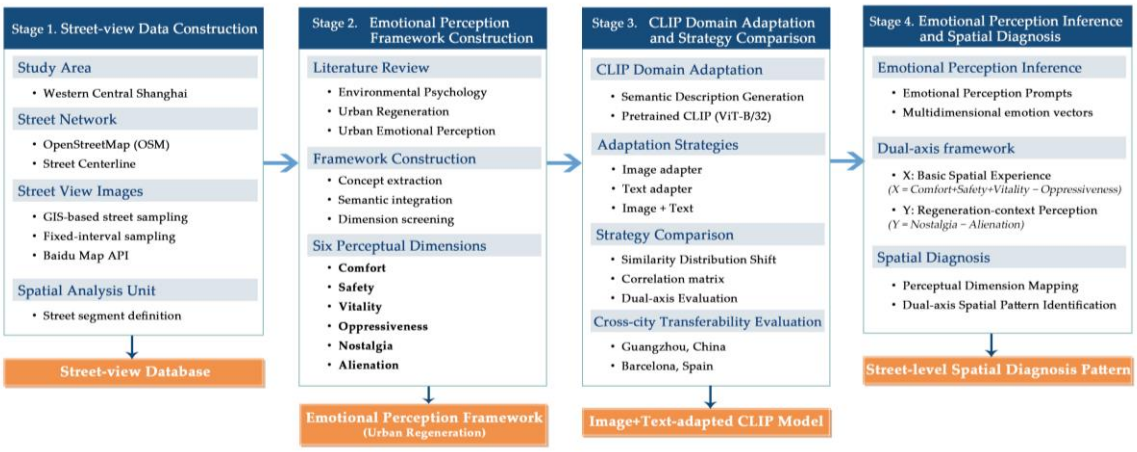


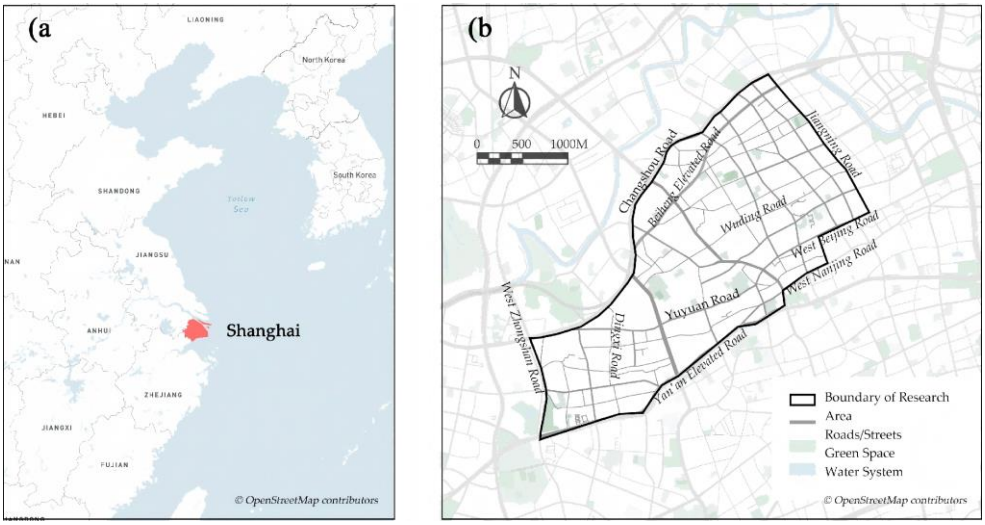
Figure 1. Four-stage research framework research framework.

3.2. Study Area and Data Collection

3.2.1. Study Area

The study area is located in the western part of Shanghai's central urban area, centered on Yuyuan Road and Wuding Road and extending to surrounding neighborhoods to form a continuous street network (Figure 2). The delineated boundary represents the scope of the study. To reduce the influence of building scale differences and functional heterogeneity, the eastern commercial core of the area was excluded from the study.

 doi:10.20944/preprints202602.0146.v1



(**a**) (**b**)

Figure 2. Location map of the study area: (**a**) Shanghai; (**b**) Research area. This map was created and edited based on data from OpenStreetMap (© OpenStreetMap contributors).

Unlike large-scale renovation projects, the regeneration of selected areas primarily involves small, gradual interventions such as facade renovations, adaptive reuse of historic buildings, introduction of street-level commercial uses, and improvements to public spaces [38,39]. The selected area features a dense road network with mixed functions and uneven renewal intensity of primary and secondary streets. This coexistence of older and renewed spaces provides a suitable sample for comparing emotional perception. Given that emotional experiences are often more pronounced in secondary streets with greater spatial heterogeneity [23,33,35], this study covered all public street network in the area in order to capture perceptual differences under different street conditions. This aligns with established methods in street-level perception research [24,30].

3.2.2. Data Collection

In this study, OpenStreetMap (OSM) is used as the basic data source of Shanghai's urban road network to construct a street view sampling framework. As an open geographic platform, OSM has been proven reliable in spatial analysis and street network research [40,41]. Using GIS, sampling points are generated every 20 meters along the centerline of the road. This distance is chosen to balance spatial detail with manageable data volumes. After cleaning the point data by removing duplicates and invalid entries, the total of 5959 valid sampling points were remained. All point data coordinate systems are uniformly converted according to the coordinate requirements of Baidu Maps API.

All street view images were retrieved at a resolution of 512×512 pixels with a quality parameter of 90 dpi. The field of view angle adopted a standard 90-degree angle, similar to the natural field of view of the human eye. The acquisition direction was configured with four basic directions: 0° (Road forward direction), 90° (Right side), 180° (Backward direction), and 270° (Left side). This design can completely record 360-degree environmental information around the sampling point. After removing coordinate sampling points without street view data, 4820 sets of valid street view data were obtained. Finally, the collected images from the four directions were combined to generate a panoramic image simulating the complete visual information available to a person at the sampling point location (Figure 3).

 doi:10.20944/preprints202602.0146.v1

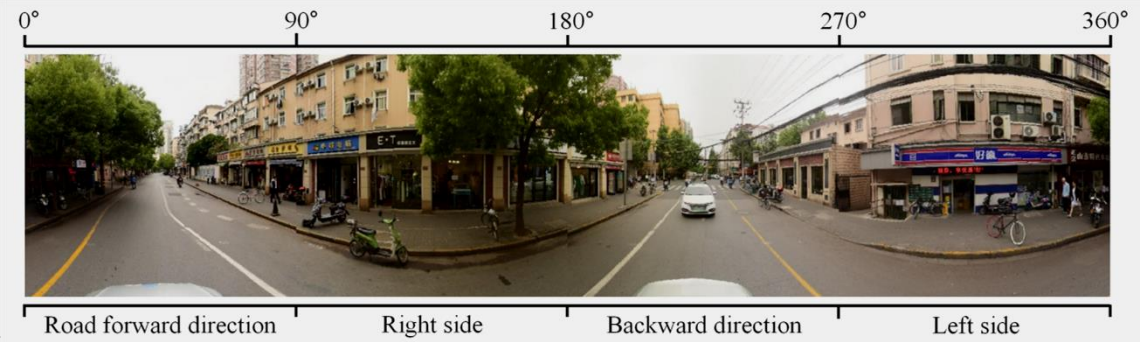


Figure 3. Example of a panoramic SVI from Baidu Maps. Image © Baidu Maps.

3.3. Construction of the Street Emotional Perception Framework

To construct a street emotional perception framework with literature consistency and computational operability in the context of urban regeneration, this study combines a systematic of literature review with semantic integration. Through systematic research on urban regeneration, environmental psychology, place experience and urban emotional perception, this study focuses on place-related affections, urban perception and street-level experience. Existing studies have generally pointed out that emotional perception dimension in urban space is closely related to factors such as place memory, identity, safety, vitality, and comfort, but the related emotional concepts are described in various ways in different studies [1,2,6,15].

Building upon this foundation, this paper extracts emotional and perceptual vocabulary directly related to street view user experience. An initial emotional vocabulary set is constructed. Similar concepts are merged semantically into several stable emotional clusters through semantic cleansing and integration to reduce conceptual redundancy and maintain clear semantic boundaries. To ensure this framework is applicable to SVI analysis, the selection of emotional dimensions follows the basic principles of existing street view perception research. Perceptions that can be stably inferred from visual features are selected while others that highly depend on individual experience, social relationships, or value judgments are weakened or excluded. [1024,25,30].

Unlike a simple dichotomy of positive and negative emotions, this paper retains emotional dimensions highly relevant to the urban regeneration context during its construction. For example, terms such as nostalgia, reminiscent, and historic-feeling are integrated into Nostalgia to represent emotional experiences triggered by historical remnants, traditional materials, and spatial memory cues. Existing research indicates that such emotions are closely related to the reconstruction of local memory and place identity during the regeneration process [6,14,15]. Meanwhile, emotional perceptions such as alienation, uneasy, and estrangement, which reflect spatial discontinuity or inadaptation to regeneration, are categorized into the Alienation dimension [2,32].

Table 1. Categories for Urban Regeneration-Oriented Street Perception Dimensions.

Perceptual Dimension	Perceptual Focus (Street-level)	Key Visual Elements	Final CLIP Text Prompts (example)
Comfort	Physical and psychological ease at street level	greenery, sidewalk, seating, human-scale buildings	a street scene with seating and resting spaces for pedestrians
Vitality	Level of public activity and social presence	pedestrian density, active storefronts, cafes, street activities	a street scene with cafes, shops, and outdoor street activity
Safety	Sense of order, visibility, and personal security	lighting, openness, visibility, orderly street edges	a street scene with open street edges and clear spatial structure

 doi:10.20944/preprints202602.0146.v1

Oppressiveness	Spatial pressure and environmental stress	narrow streets, high enclosure, visual clutter, dense buildings	a street scene with visual clutter and crowded surroundings
Nostalgia	Continuity with local history and place memory	traditional architecture, aged facades, historic materials	a street scene with old buildings and aged facades
Alienation	Lack of human scale and social connection	oversized spaces, empty plazas, abrupt stylistic contrasts	a street scene with blank facades and inactive street edges

Overall, this paper summarizes a street spatial perception framework applicable to the context of urban regeneration (Table 1). These dimensions are integrated from core concepts that have appeared repeatedly in existing research, have relatively clear semantic boundaries, and are street-level oriented.

The six dimensions can be divided into two categories. The first category consists of basic spatial experience dimensions at the pedestrian level, including Comfort, Safety, Vitality, and Oppressiveness. Safety and Comfort are commonly used dimensions in street perception research, reflecting key aspects of pedestrians' evaluation of street environments [24,25,33,35,42]. Vitality describes the level of activity and social interaction on the street, used to measure the quality of public life [23,36,42]. Oppressiveness summarizes the oppressive experience caused by visual forms such as spatial compression or visual crowding, serving as a negative perceptual dimension to supplement the spatial inhibition effect that is difficult to capture in the positive spatial experience dimension [4,23,25].

The second category is placed within the context of urban regeneration. Nostalgia, associated with historical continuity, local memory, and place identity, is often used to understand emotional responses in street regeneration [6,15]. Alienation, on the other hand, depicts a sense of exclusion or rejection brought about by regeneration, usually stemming from the imbalance of spatial scale, the fragmentation of street interfaces and activities, and the disappearance of local visual features [2,32].

The six dimensions, with their corresponding street visual elements, and representative text prompts are summarized in Table 1. Each dimension corresponds to six natural language descriptions, collectively forming a street-level emotional perception prompts library (Appendix Table A1 for the complete content). These prompts will be used to ensure that different street-view samples are measured for emotional perception at a uniform semantic scale, rather than for training the CLIP model. Overall, this framework achieves a balance between theoretical relevance and computational stability by controlling the number of perceptual dimensions, providing an operational basis for street-level emotional perception analysis in the context of urban regeneration.

3.4. Lightweight Domain Adaptation Strategy for CLIP Model

CLIP model learns a shared cross-modal embedding space through large-scale image–text contrastive learning and demonstrates strong generalization in open-domain visual understanding. However, its pretraining data are dominated by natural images and web-based text, leading to semantic representations that emphasize generic objects and scenes. These representations differ from urban street scenes, which are characterized by explicit spatial scale, functional constraints, and environmental context. When directly applied to street-level emotional perception, the original CLIP model tends to activate general visual semantics and may fail to stably capture perception-related street features.

To address this limitation, CLIP is regarded as an adaptable cross-modal backbone rather than a fixed inference model and introduces lightweight domain adaptation to adjust its embedding space for SVIs. The backbone parameters of CLIP are kept frozen, while small adapter modules are added to the image encoder, the text encoder, or both. To preserve CLIP's general visual-semantic capability

 doi:10.20944/preprints202602.0146.v1

and ensure efficient adaptation to urban regeneration contexts, updating only the task-specific adapter layers to mitigate overfitting.

This paper compares four adaptation paths: the original CLIP (Baseline), Image-adapted CLIP, Text-adapted CLIP, and Image+Text-adapted CLIP. All models are trained without any emotion supervision or perceptual dimensions, using only SVIs and their corresponding descriptive texts. Under consistent data scale, training epochs, and optimization settings, these configurations are compared in terms of representation stability, perceptual structure, and suitability for subsequent spatial analysis. The selected adaptation strategy provides a consistent and interpretable embedding space for multidimensional emotional perception inference and street-level spatial pattern analysis in the context of urban regeneration.

3.5. Dual-Axis Evaluation Framework for Emotional Perception

This study identifies the regeneration state of the street space through the dual-axis structure. Urban regeneration streets often evoke a variety of emotional responses, and different perceptual dimensions emphasize different aspects of street quality. Integrating these dimensions into a single comprehensive indicator may obscure their relationships and introduce additional weighting assumptions. Therefore, a dual-axis structure is adopted to describe basic spatial experience and regeneration-context perception separately (Figure 4).

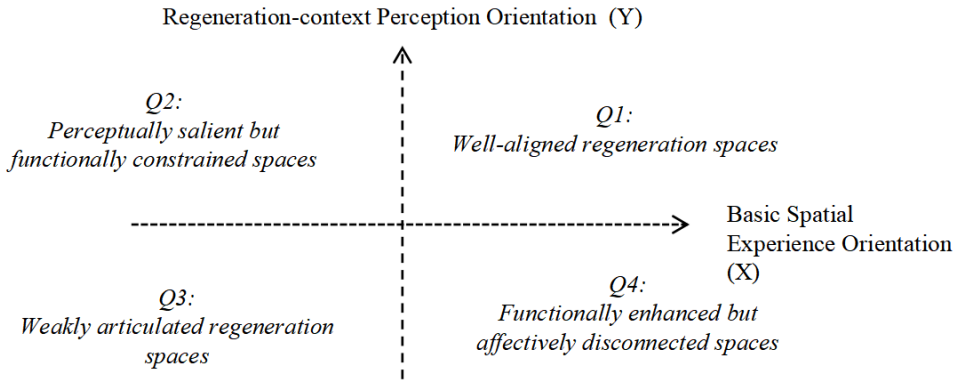


Figure 4. Dual-axis emotional evaluation framework for street space under urban regeneration.

3.5.1. Definition and Calculation of Two Axes

The dual-axis framework uses two independent axes to organize emotional perception at the street level. The basic spatial experience axis (X) represents the everyday user experience and integrates four dimensions: comfort, safety, vitality, and oppression. Comfort, safety, and vitality indicate the positive aspects of the street experience, while oppression captures the negative spatial pressure. For operational simplicity and interpretation, the X-axis is defined as the difference between the total score for comfort, safety, and vitality and the score for oppression (X= Comfort + Safety + Vitality − Oppressiveness). This equal-weighted formulation is an analytical choice and does not imply equal importance of each dimension in practice.

The regeneration-context perception axis (Y) reflects how streets maintain or reshape place meaning during urban regeneration. Nostalgia captures emotional responses related to local memory and historical continuity, while alienation indicates a sense of rupture or rejection triggered by renewal interventions. To represent the contrasting emotional orientations of these two dimensions, the Y-axis is defined as the difference between standardized nostalgia and alienation scores (Y = Nostalgia − Alienation), representing overall perceptual orientation in the regeneration context.

3.5.2. Explanation of Spatial Types in the Four Quadrants

By mapping street-level emotional perception results to the dual-axis coordinate space, four representative regeneration states are identified (Figure 4). This classification supports a structured

 doi:10.20944/preprints202602.0146.v1

explanation of how daily experiences and regeneration-context emotional perception interact, rather than ranking streets by a single performance score.

Quadrant 1 (Q1) represents well-aligned regeneration streets, characterized by strong basic spatial experience and positive regeneration-context perception, indicating consistency between daily perception and place meaning. Quadrant 2 (Q2) includes perceptually salient but functionally constrained streets, where regeneration-related meaning is evident, while daily spatial experience remains limited. Quadrant 3 (Q3) corresponds to weakly articulated regeneration streets, which perform poorly in both basic spatial experience and regeneration-context perception, reflecting limited effectiveness of renewal interventions. Quadrant 4 (Q4) represents functionally enhanced but affectively disconnected streets, where improvements in everyday spatial conditions are not accompanied by corresponding expression of local memory or emotional identification.

3.6. Street Emotional Perception Inference and Spatial Analysis Methods

After completing the domain adaptation of the CLIP model, a street-level emotional perception inference method is constructed based on cross-modal similarity. Focusing on the six perceptual dimensions of comfort, vitality, safety, oppressiveness, nostalgia, and alienation, this study constructs corresponding natural language text prompt collections to characterize the typical perceptual semantics that each dimension may present in street view scenes.

By calculating the similarity between image features and text prompt vectors, the relative scores of SVIs in each emotional perception dimension can be obtained. The score reflects the degree of matching of SVIs in a specific perceptual semantic direction, rather than a single numerical quantification of complex spatial experiences.

In order to introduce the image-level emotional perception results into the spatial analysis framework, further maps the inference results to geographic space. SVIs are sampled at a distance of about 20m along road network, and have clear spatial coordinate information. In spatial analysis phase, emotion scores of images are projected onto their corresponding spatial locations and aggregated using a regular grid. A 50m×50m regular grid is used as analysis unit to aggregate multiple adjacent images, thereby reducing viewpoint redundancy and local noise while preserving street-level spatial heterogeneity, resulting in a stable perceptual representation of street segments.

On this basis, this paper does not linearly integrate the multi-dimensional emotional perception results into a single comprehensive index, but explores the combination relationship between different perception dimensions through structured analysis. The emotional perception results are mapped to a dual-axis analysis framework to identify the relative position of street view scenes in the orientation of basic experience and regeneration-context perception, and to analyze its distribution characteristics at the spatial level. Through this process, a complete method path from SVIs perception inference to spatial pattern recognition is formed.

4. CLIP Adaptation and Diagnostic Evaluation of Emotional Perception

4.1. Semantic Description Generation of SVIs

In order to construct an image-text pairing dataset for CLIP lightweight domain adaptation, this paper uses a multimodal large language model to automatically generate semantic descriptions of SVIs in the study area.

Text generation is constrained based on a structured prompt framework. The GPT-4o multimodal API interface is called to set the model as an analyst with urban regeneration knowledge, and only describes the spatial visual elements that can be directly observed in the image. The length of the descriptive text is controlled within 80 English words. The content of the descriptive text revolves around the basic spatial experience at the pedestrian scale and the emotional perception of the urban regeneration context. Prompts explicitly prohibit speculative or evaluative suggestions to ensure a strict correspondence between the text and the visual content. Before batch generation, the prompt framework is iteratively optimized in combination with manual verification. About 50 SVIs

 doi:10.20944/preprints202602.0146.v1

were randomly selected during the iteration to check the consistency between the generated text and the image content. The prompt constraints were adjusted accordingly until the generated results reached a stable level in terms of accuracy and contextual relevance.

By integrating multimodal large language model and automated processing processes, 4,820 SVIs are semantically annotated. After deduplication of image content and multiple rounds of manual sampling review, a final standard dataset containing 4,428 high-quality image-text pairings was constructed.

4.2. Comparative Diagnostics of CLIP Adaptation Strategies

In order to systematically evaluate the impact of different lightweight domain adaptation strategies on the CLIP representation space, this paper constructs and compares four model configurations under unified training settings, including the original CLIP (Baseline) without adaptation, the Image Adapter with only the adaptation module on the image side, the Text Adapter with the adaptation module only on the text side, and the Image+Text Adapter (Both) with both sides of the image and text at the same time. All adaptation models employ the same CLIP pre-training weight (ViT-B/32), with the backbone parameter frozen. Only the adaptation module and the corresponding projection layer parameters are updated, in order to achieve efficient domain adaptation of parameters and avoid large-scale disturbances of the original semantic structure.

In terms of training mechanism, all models are optimized using CLIP's original image-text contrastive learning objective. The SVIs and its corresponding original descriptive text form a positive sample pair, without introducing any emotion labels, perceptual dimension prompts, or additional supervisory signals. All models are trained at the same data scale, training rounds (epoch = 6), and optimization settings to ensure that observed differences mainly stem from the adaptation strategy itself, rather than the difference in training conditions.

4.2.1. Similarity Distribution Shift across Adaptation Strategies

In order to investigate the overall impact of different CLIP adaptation strategies on the image–text similarity, the image-text cosine similarity distribution generated by each model under six perceptual dimensions prompts were counted on 500 randomly sampled SVIs, and their overall morphology was compared and analyzed by kernel density estimation (Figure 5). This analysis does not focus on the difference in a single dimension, but aims to test whether different adaptation strategies introduce systematic distribution shifts at the overall matching behavior level.

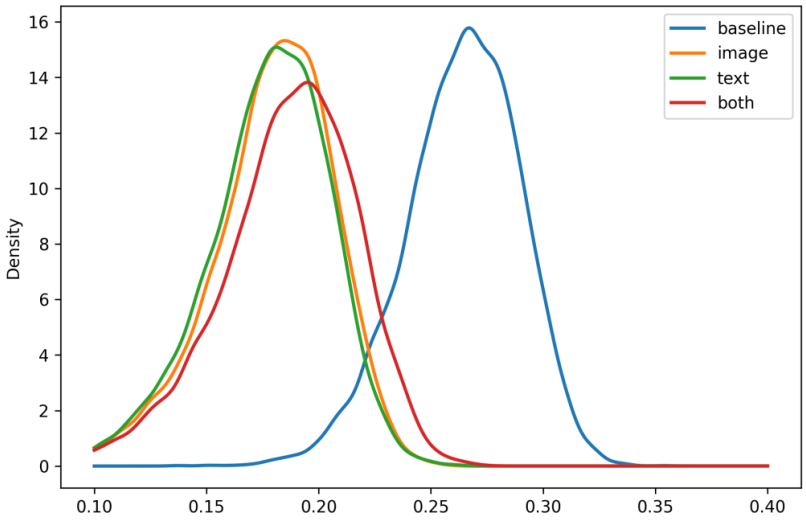


Figure 5. Similarity Distribution Shift (Image-Text Cosine Similarity).

From the perspective of distribution morphology, the similarity distribution of Baseline CLIP shifted to the right as a whole, and the peak was concentrated in the higher range, indicating that the model without domain adaptation generally gave a higher matching score between the SVI and the

 doi:10.20944/preprints202602.0146.v1

perceptual dimension prompt. This feature reflects the strong alignment tendency of the pre-trained CLIP model in the general semantic space, which tends to highly correlate multiple visual elements with abstract descriptions, thereby weakening the distinction between different perceptual dimensions. In contrast, after the introduction of lightweight adaptation, the similarity distribution of Image-adapted CLIP and Text-adapted CLIP showed a significant shift to the left. Simultaneously, the distribution presented a more concentrated form, indicating that the model became more conservative in the matching process.

In the joint adaptation model (Both), the similarity distribution was between the single-sided adaptation and the baseline. It indicated the model avoided the overall high general matching tendency in the baseline and did not over-compress the matching strength as the single-sided adaptation. This intermediate distribution showed that joint adaptation retained a certain degree of semantic elasticity while suppressing general semantic over-alignment.

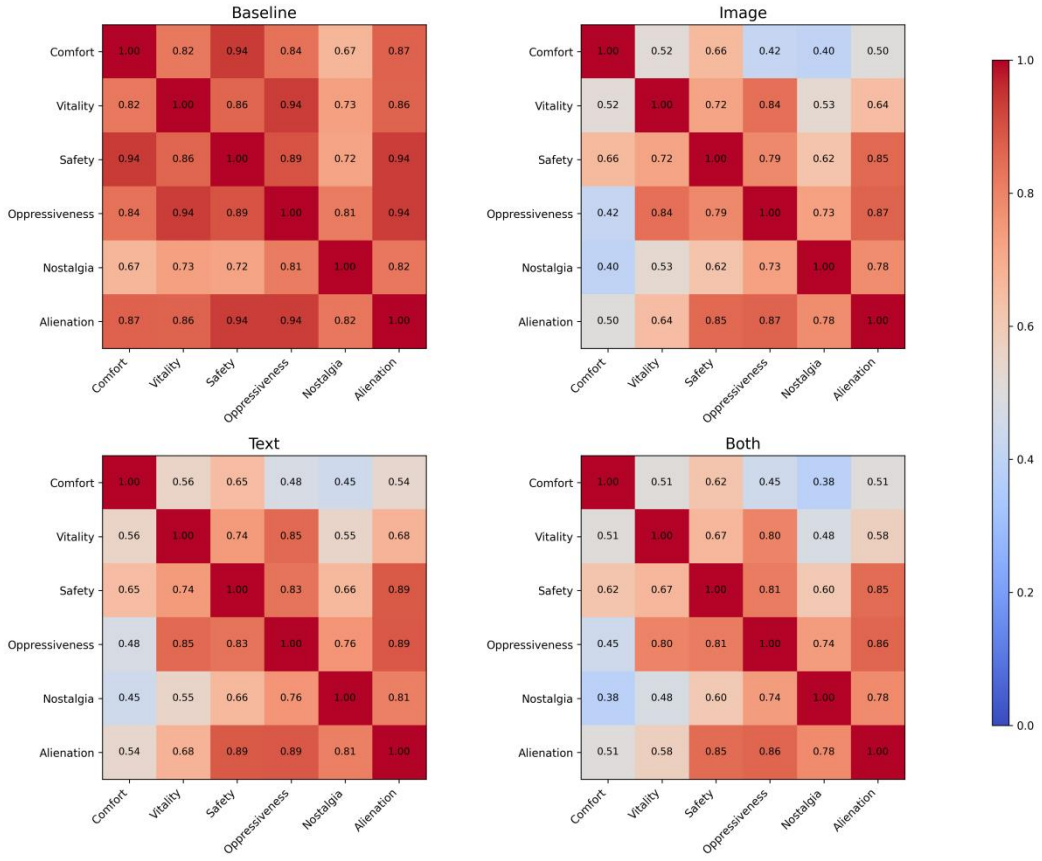


Figure 6. Correlation matrix of perceptual dimensions (n=500).

4.2.2. Dimension Correlation Analysis

To examine the discriminative power and structural stability of the six perceptual dimensions in real street view samples, this paper randomly selected 500 SVIs and performed Pearson correlation analysis on the inference scores of each dimension. The perceptual dimension correlation matrices were constructed based on Baseline CLIP model, Image-adapted CLIP model, Text-adapted CLIP model, and Image+Text-adapted CLIP model respectively (Figure 7). From an overall perspective, the correlation matrices under all four models exhibited a structure characterized by predominantly moderate to high correlations, but different models showed significant differences in the distribution and hierarchical distinction of correlation intensity. Specifically, the correlation of Comfort with other dimensions was generally low across the four models, illustrating the independence of comfort experience at the visual perception level. Meanwhile, vitality, safety, oppressiveness, nostalgia, and

 doi:10.20944/preprints202602.0146.v1

alienation formed relatively stable high-correlation regions across multiple models. This result aligns with the fact that real-world perceptions of street spaces are often composed of complex emotions.

Further comparison of the correlation matrices of different models revealed that Baseline CLIP exhibited a generally high and relatively uniform correlation distribution across the six dimensions, with most off-diagonal elements in the darker red region. This characteristic indicated that the original CLIP model tended to use the general visual-linguistic semantic structure formed during the pre-training phase to characterize street scene perceptual attributes, thus compressing the hierarchical differences between different perceptual dimensions at the structural level. However, this unified relevant structure limited the model's ability to distinguish subtle differences between perceptual dimensions.

In contrast, the introduction of lightweight adaptation resulted in a more differentiated relevance matrix in terms of color distribution and hierarchical structure. Both Image-adapted CLIP and Text-adapted CLIP introduced variations in relevance strength between different dimensional pairs, resulting in an alternating distribution of high and moderate relevance in the matrix. This indicated that one-sided adaptation began to reweight the perceptual structure. Further observation revealed that image-side adaptation primarily affected dimensional pairs related to spatial morphology and visual saliency, while text-side adaptation adjusted dimensional combinations with strong semantic connections in a more significant way. Building on this, Image+Text-adapted CLIP, while maintaining multidimensional perceptual co-occurrence relationships, presents a clearer hierarchical gradient in correlation strength. It avoided the overall correlation collapse occurred in the baseline, as well as excessive decoupling, thus providing a more robust representational foundation for subsequent multidimensional perceptions analysis.

4.2.3. Dual-axis Evaluation Framework

In this paper, the perception inference results of four CLIP adaptation models on randomly sampled 500 SVIs were mapped into the dual-axis evaluation framework (Figure 7). The purpose was to test whether different models could form a stable and interpretable perceptual structure under a unified coordinate system. The dual-axis framework was used to characterize the relative position of the street scene in the two directions of basic spatial experience (X) and regeneration-context orientation (Y). The value range of biaxial coordinates is naturally generated by linear combination and difference relationship after the multi-dimensional perception score is normalized by z-score. Since these two axes are composed of different numbers of perceptual dimensions, their numerical ranges are different and should be interpreted as relative perceptual positions rather than absolute emotional intensity. For each model, the degree of dispersion along the X-axis and Y-axis is quantified through Equation (1).

$$Var(X) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2, \qquad Var(Y) = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 \tag{1}$$

From the results of overall distribution patterns, the four models all showed continuous distribution characteristics in dual-axis quadrants. None of them formed clear discrete boundaries or quadrant segmentation, which indicated no simple binary opposition between basic user experience and regeneration-context perception in the analyzed SVIs. This phenomenon was consistent with the multi-dimensional perceptual co-occurrence characteristics revealed in the previous dimensional correlation analysis. It indicated the dual-axis framework's suitability in describing the combined state of street-level perceptions, rather than making threshold-based or category-based judgments.

 doi:10.20944/preprints202602.0146.v1

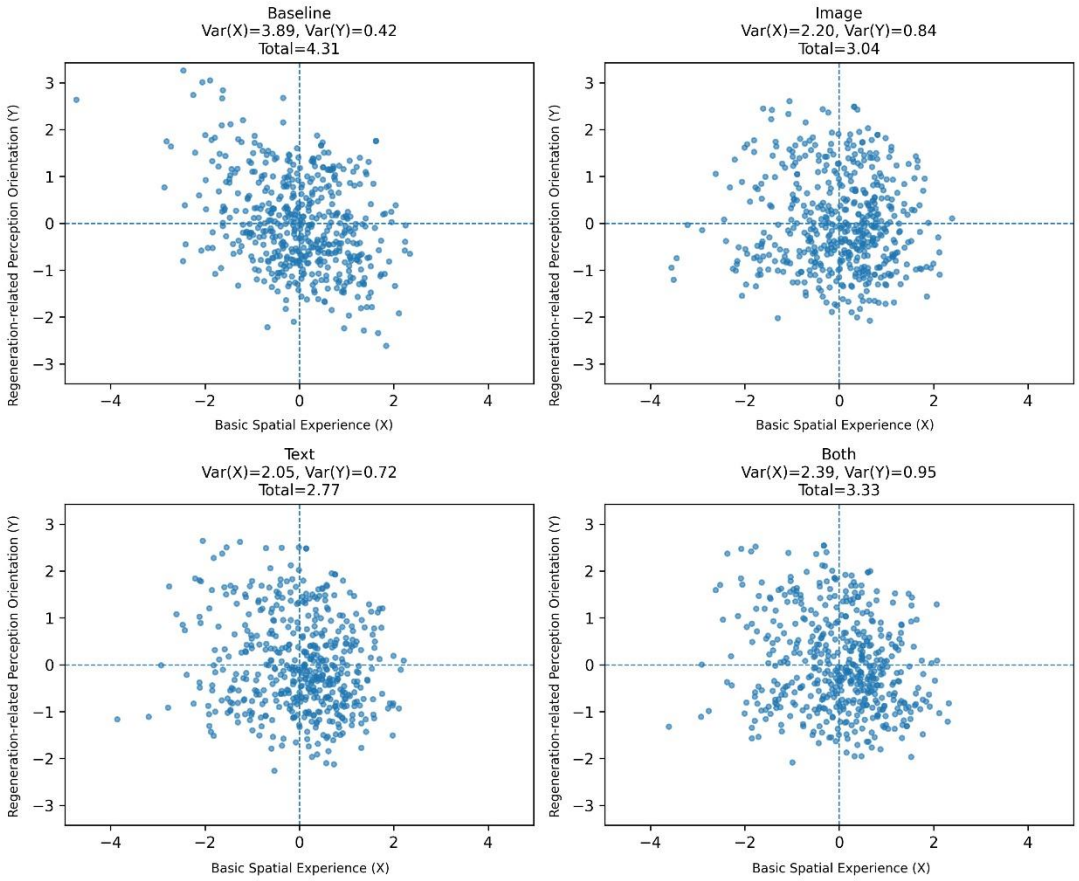


Figure 7. Dual-axis Evaluation Framework with four CLIP Adaptation Strategies.

The dual-axis evaluation framework was also used for comparing the four models. The study investigated whether different adaptation strategies could avoid the collapse or unilateral dominance of the perceptual structure. The sample distribution of Baseline CLIP was relatively concentrated, and a large number of SVIs were gathered near the origin of the coordinates, reflecting the model's limitation in discriminative ability. After the introduction of lightweight adaptation, Image-adapted CLIP and Text-adapted CLIP stretched the sample distribution across the Y axis with a higher degree of dispersion. However, both models presented a lower degree of dispersion along the X axis and shifted the distribution center of gravity in the same time. This indicated that unilateral adaptation may weaken the structural expression of the other axis while enhancing a certain type of perceptual cue. It also revealed a higher sensitivity of the Baseline model in differentiating basic spatial experience. In contrast, the two-sided adaptation model (Both) presented a even higher degree of dispersion along the Y axis while maintaining a relatively high degree of dispersion along the X axis. It also formed a more balanced distribution in the dual-axis quadrants.

4.3. Diagnostic Synthesis and Model Selection

Based on the comparative analysis results of similarity distribution shift, dimensional correlation structure and dual-axis evaluation framework, there were clear differences in the ability of different CLIP lightweight adaptation strategies in perceptual modeling. Image-adapted CLIP showed a more significant distribution adjustment effect in the overall image-text matching intensity control, reflecting its greater sensitivity to single-dimension perceptual variation and visual feature sensitivity. However, this enhancement was mainly reflected in the level of matching strength, and did not form the same degree of stability in the organization and structural expression of cross-dimensional relationships. Text-adapted CLIP reweighted the perceptual structure at the semantic mapping level, but its impact on the overall perceptual structure was relatively limited.

 doi:10.20944/preprints202602.0146.v1

It should be noted that the adaptation process in this paper aimed to improve CLIP's domain alignment ability in the context of urban regeneration, rather than optimizing the predictive performance for specific perception dimensions. Under this premise, a comprehensive consideration of the distribution stretching ability, multi-dimensional perceptual relationship, structural consistency is required.

Image+Text-adapted CLIP showed more balanced and stable structural characteristics in the three types of diagnosis above. On the one hand, the model suppresses the general matching tendency of Baseline CLIP through improving the discrimination of similarity distribution. On the other hand, it maintains the co-occurrence structure and axial discrimination ability of multi-dimensional perception.Based on the above comparison results, Image+Text-adapted CLIP was selected as the model for subsequent street scene emotional perception analysis.

4.4. Cross-City Transferability of the Image+Text-adapted CLIP Model

4.4.1 Area selection and sampling instructions

To test the cross-city transferability of Image+Text-adapted CLIP in different urban contexts, Guangzhou and Barcelona were selected as comparative cities covering different regeneration patterns and street spatial characteristics.

The study area in Guangzhou was the historical and cultural block of Yongqingfang-Enning Road, which included typical arcade streets and urban villages. Compared with Shanghai, the spatial interfaces were more heterogeneous, and the regeneration approaches were mainly gradual and hybrid intervention. The study areas in Barcelona included El Raval, a historical regenerated district and 22@ Barcelona Innovation District. In both cities, an 800m × 800m zone was defined to sample points at 50m spacing. A total number of 50 (Guangzhou) and 100 (Barcelona) SVIs from these points were randomly selected for verification.

4.4.1 Analysis of cross-city output consistency and difference

Image+Text-adapted CLIP was applied to perform six-dimensional emotional perception inference on the SVIs of the sampled points, and the output was verified against the corresponding images and text descriptions. At the comparison level, the perception results were compared between Guangzhou and Shanghai, and between two areas in Barcelona, to examine whether the model could generate interpretable differential expressions in different cities and different regenerated areas.

The results comparing Guangzhou and Shanghai showed that the overall scores across the six emotional dimensions in Guangzhou were generally lower than those in Shanghai, with a relatively higher proportion of negative perceptions in Oppressiveness and Alienation. Some street sections with arcade-style facades or urban village spatial characteristics did not show a significant advantage in scores on Comfort, Safety, and Vitality dimensions which remained consistent with the model output and the corresponding SVIs.

In comparative analysis of Barcelona, the model established a distinguishable emotional perception structure between the historical regenerated area El Raval and the transformational regenerated area 22@ Barcelona Innovation District. Two areas showed consistent differences in output across basic spatial experience and regeneration-context perceptual dimensions, without exhibiting significant dimensional collapse or random fluctuations.

While the overall score distribution differed between cities, the distribution of outcome remained structurally stable within the city. The results above demonstrated the Image+Text-adapted CLIP's transferability across cities, with an outcome of consistent and interpretable distribution of perception scores. Nevertheless, the model still exhibited systematic biases when dealing with strongly localized spatial elements, providing a clear direction for future improvements.

5. CLIP-Based Spatial Perception Recognition and Analysis

5.1. Perceptual Dimension Mapping at Street Level

 doi:10.20944/preprints202602.0146.v1

Based on the street emotional perception framework, this paper analyzed the perception results of 4,428 SVIs in the study area in six emotional dimensions. The SVIs were sampled along the road network at a spacing of about 20 m. Spatial aggregation was carried out through regular grids of 50 m × 50 m, and the perceptual score for each grid cell was generated by the median statistic to reduce the influence of adjacent viewing angle redundancy and local occlusion. The score of each emotion dimension reflects the relative match between the SVI and the corresponding emotion prompt text in the CLIP semantic embedding space. Figure 8 shows the spatial distribution of the six emotional perception dimensions across the study area. In order to avoid the interference of extreme values on color mapping and enhance the comparability of different dimensions in spatial structure, the scores of each dimension are truncated in the 5%–95% quantile range for presentation. The nuances in perception intensity are represented by continuous color bands, so that the spatial gradient and local differences can be presented at the same time.

From the perspective of overall distribution characteristics, all the six emotional dimensions formed spatial patterns with clear continuity along the street network. This phenomenon suggests that the emotional perception results based on visual-linguistic semantic similarity can show a stable spatial structure at the neighborhood level, rather than reflecting only the local visual differences of a single image. Different dimensions show significant differences in spatial continuity, concentration and local fluctuation, reflecting the diversity of spatial expression of emotional perception.

In terms of spatial distribution, Alienation and Oppressiveness show a similar pattern. High values show a linear distribution along major streets, with clusters in the study area. Comfort and Safety demonstrate a more homogeneous and stable pattern throughout the region. Although there are gradient changes in local street sections, high values usually form a primarily continuous positive perception section along the road network. Unlike the other dimensions, Nostalgia displays a more discrete pattern of high and low values, confined to localized street segments without forming broad clusters. Vitality shows continuity in some main streets or key nodes, and fluctuating more pronounced in secondary roads.

 doi:10.20944/preprints202602.0146.v1

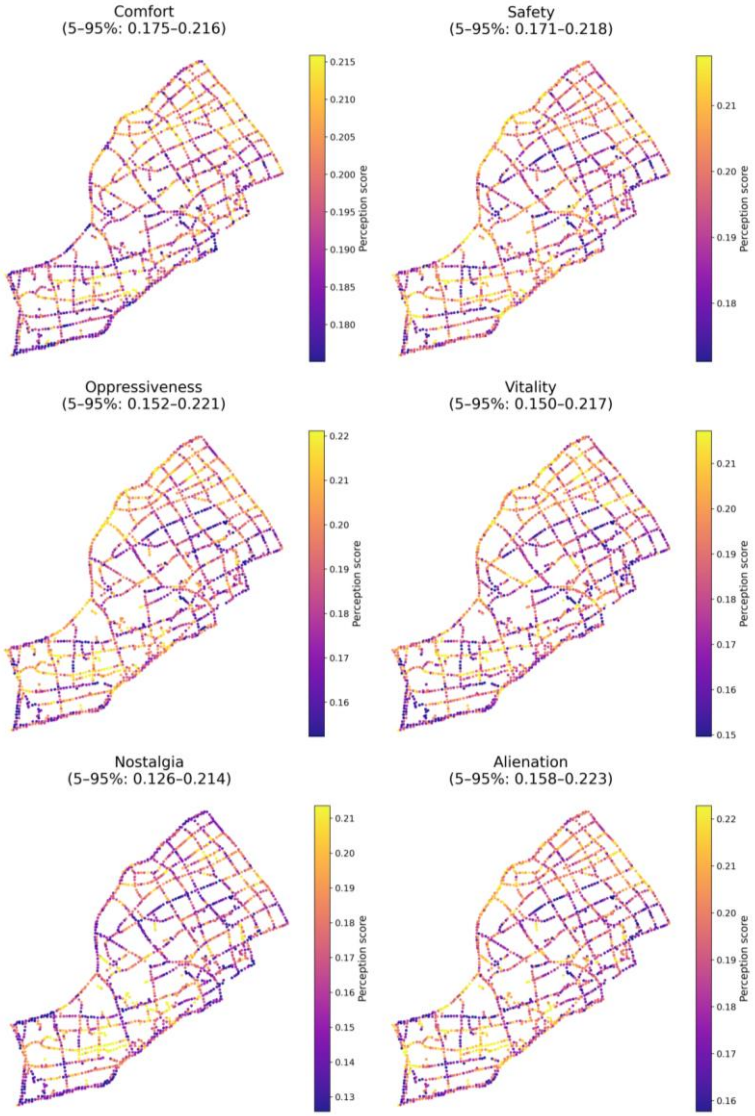


Figure 8. Perceptual Dimensions Mapping at Street level.

Overall, the six emotional dimensions exhibit a differentiated spatial distribution structure along the street network, with some dimensions showing strong continuity while others are characterized by local heterogeneity.

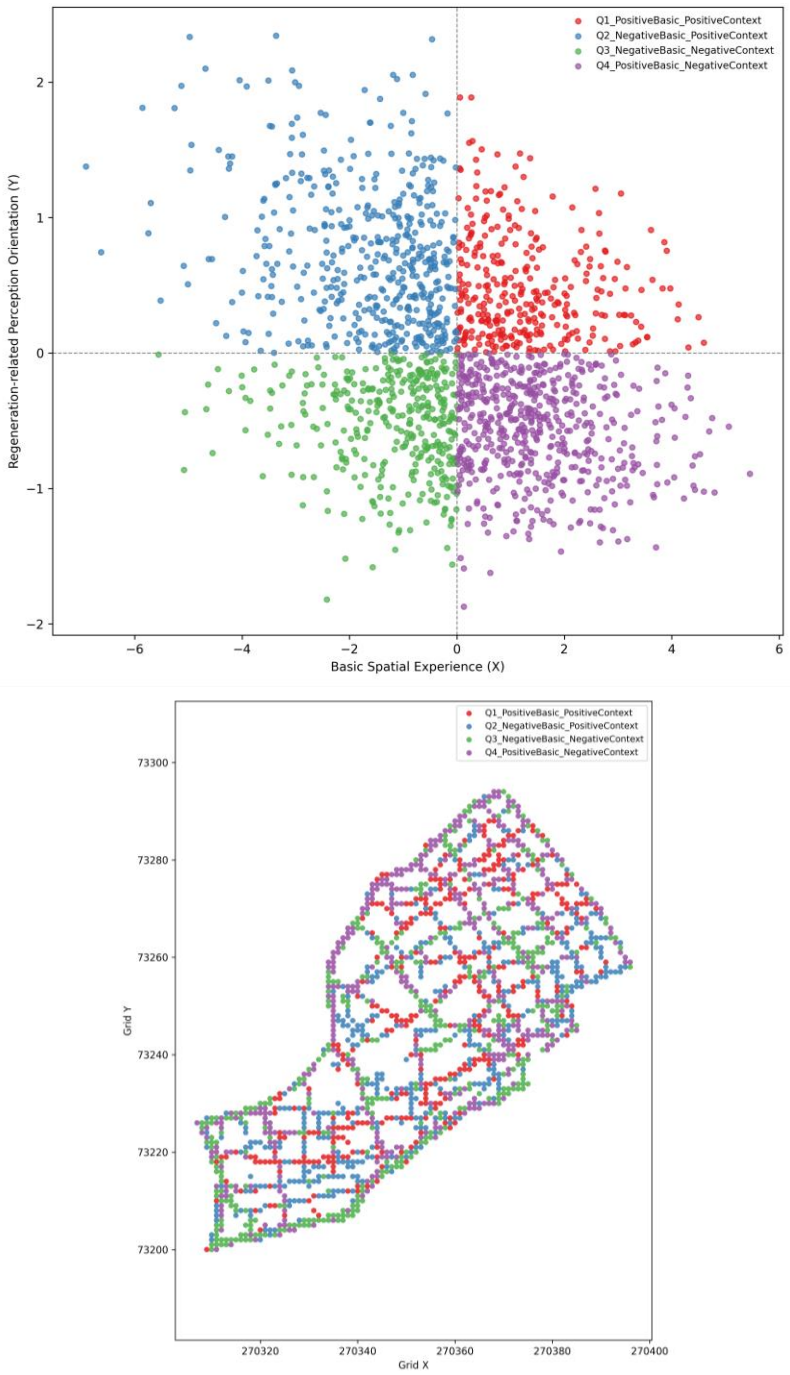
5.2. Dual-Axis Spatial Pattern Identification

Based on the dual-axis perception framework, the perceptual scores at the grid scale of 50 m × 50 m were projected into the dual-axis coordinate system of basic urban spatial experience (X)-regeneration-context perception (Y) (Figure 9a).

The different value ranges of numerical dispersion between the two axes is related to the different number of dimensions they contain. This pattern reflects the distinction between basic user experience and regeneration-context perception in terms of structural organization and fluctuation amplitude, rather than representing a direct comparison of perceptual strength. Further, differences in street space are more visible and distinguishable at the level of basic experience. In contrast, the regeneration-context perception is manifested through changes around the existing space rather than the independent perceptual dimension. Samples in different quadrants are interconnected by a large number of points close to the coordinate axes, presenting a smooth transition in distribution. This

 doi:10.20944/preprints202602.0146.v1

result indicates that emotional responses at the neighborhood level is closer to a continuous spectrum.



(**a**) (**b**)

Figure 9. Dual-axis spatial pattern: (**a**) Dual-axis scatter plot; (**b**) Spatial mapping.

When the dual-axis mapping results are projected to the geographic space (Figure 9b), the above continuity characteristics are intuitively reflected at the spatial level. Different quadrant types show a highly staggered and mosaic distribution pattern along the street network, rather than forming a large-scale and homogeneous continuous area. In some street sections, points from all quadrant types can be observed to be continuously distributed along the road, while in places where street turns, function mixes, or spatial interface changes, the perception types are more likely to switch quickly.

 doi:10.20944/preprints202602.0146.v1

This phenomenon indicates that the capacity of dual-axis perception framework to detect micro-scale perceptual fluctuations by responding to variations in street layout and renewal efforts.

Consequently, this framework is not used to classify space discretely, but provides a structured perspective for understanding the distribution and change of emotional perception in street space by revealing the relationship between basic spatial experience and regeneration-context perception.

6. Discussion

6.1. Interpreting Spatial Perception Patterns in Urban Regeneration Context

The results of dual-axis analysis and spatial mapping revealed the patterns of multi-dimensional perceptions based on SVIs at the street level. Through combining the emotional perception results with the visual features that can be directly observed in SVIs, the spatial characteristics of different emotional dimensions are diagnosed, so as to examine their correspondence with the specific street environment.

From the perspective of overall distribution, multiple emotional dimensions show clear spatial continuity at the street level. This feature may be related to the pedestrian's experience of the street environment, which is usually formed gradually through the continuous walking process. When similar spatial features appear repeatedly in a continuous street segment, the emotional perception results appear as smooth changes spatially rather than random fluctuations.

For specific emotional dimensions, both Comfort and Safety show high spatial stability, but their corresponding visual cues are different. Higher Comfort scores is mostly found in streets with more pedestrian-friendly conditions, which usually include continuous and relatively open sidewalk space, good paving quality, clear walking paths, and a certain proportion of green elements. In contrast, the distribution of safety is more related to the readability and sense of order in the street. Its corresponding streetscape characteristics also plays an important role, including clear spatial boundaries, a clear distinction between pedestrian and vehicular spaces, continuous building façade definition, and less visual occlusion. Vitality exhibits a localized and directional distribution pattern in space. Streets with high perceived vitality are typically accompanied by visible pedestrian activity, high openness of streetfront facades, and diversity of street functions. This indicates that perceived vitality is closely related to the intensity of street use and visibility of social interaction.

Oppressiveness and Alienation are both negative emotional dimensions, but their spatial meanings are different. Oppressiveness is typically found in street environments where the proportion of street space and interfacial conditions impose a strong sense of oppression on pedestrians, such as narrow pedestrian space, high building height, low sky openness, or chaotic street interface elements. In contrast, Alienation emphasizes the lack of social interaction and emotional connection. Street sections with high alienation scores often lack clear pedestrian activities. For example, the interface along the street is closed or remains mono-functional, and despite the large spatial scale, there is a lack of facilities that support staying or social interaction. Nostalgia perception is often concentrated in specific street segments or nodes, and it is difficult to form a continuous structure along the street network. The corresponding SVIs mostly contain building facades with age characteristics, traditional street and alley scales, historical materials or detailed elements related to local memory.

Under the framework of dual-axis analysis, the above six emotional dimensions are further organized into the relationship structure between basic urban spatial experience and regeneration-context perception orientation. This structure helps to identify the combined state of streets in different perceptual directions, revealing that the improvement of the underlying user experience is not necessarily accompanied by the simultaneous change of regeneration-context perception. This inconsistency reflects the real process of asynchronous adjustment of different spatial elements in gradual urban regeneration.

6.2. Implications for Urban Regeneration Processes

 doi:10.20944/preprints202602.0146.v1

The emotional perception analysis framework proposed in this paper is applicable to all stages of urban regeneration practice. Its core significance is not to draw specific design conclusions or policy suggestions, but to provide a scalable and explanatory analysis method to complement conventional planning evaluation indicators.

In the early stage of urban regeneration, the multi-dimensional emotional perception mapping inferred from SVIs could complement the existing planning method relying on land use indicators, traffic conditions or architectural scale analysis. By presenting the continuous distribution characteristics of emotional perception at the street scale, this method helps to identify street segments or areas with clear uneven perception performance, and provides clear analysis clues for further verification and in-depth investigation.

The framework can be used to support the design phase of urban renewal. When the street interface, spatial scale or street activity state are visually changed in different design schemes, the framework can present the potential emotional perception of schemes under the same analysis system. This could provide a comparative basis from the user's perspective for scheme discussion. The comparative analysis also helps to identify the emotional reaction of different districts in the city, thereby informing the degree of regeneration intervention required. Meanwhile, the multi-dimensional emotional perception and dual-axis structure provide a reference for the potential orientation of street renewal, such as focusing more on enhancing the vitality of commercial and public activities, or emphasizing the continuation of historical spatial characteristics and local significance.

In the later stage of design evaluation, the framework has the potential to carry out temporal comparison. SVIs obtained before and after the regeneration could be statistically examined with emotional perception scores. The nuances in each perceptual dimension and the distribution characteristics of these changes in street space can be systematically observed. This method emphasizes whether regeneration bring perceptible changes in experiences at the street level, rather than relying solely on metrics such as greening rate, building density, or spatial heat maps.

6.3. Methodological Reflections and Limitations

The multi-dimensional emotional perception analysis framework based on SVI and CLIP domain adaptation model proposed in this paper aims to provide a scalable and interpretive analysis path for street space diagnosis in the context of urban regeneration. The emotional perception score obtained belongs to the relative representation based on the semantic matching of vision and text, and its value reflects the relative difference of SVIs at a unified semantic scale, rather than a direct measurement of individual emotional perception. Therefore, the approach is more suitable for comparative analysis between street level or spatial segments, rather than for accurate identification of emotional states at the individual level.

While lightweight domain adaptation enhances the CLIP model's semantic alignment capabilities within the context of urban regeneration and demonstrates some transferability in cross-city validation, this transfer primarily manifests in the emotional perception structure and relative distribution. The local meanings attached to the urban landscapes in different cities are not fully captured. For emotional perceptual dimensions that heavily rely on specific historical contexts or cultural symbols, the model output may suffer from semantic compression. Therefore, the model cannot replace the in-depth interpretation guided by local perception.

The analysis results depend on the sampling strategy and image acquisition conditions of SVIs to a certain extent. Although this paper reduces the effects of single image noise and angle of view redundancy by sampling at 20m spacing along the road network and 50 m × 50 m grid aggregation, SVIs can still be affected by factors such as shooting time, lighting conditions, or temporary occlusion. Therefore, this paper describes the visible spatial characteristics of street space under specific sampling conditions and its corresponding emotional perception. It has limited coverage of the perceptual changes of the same space at different times or usage states.

 doi:10.20944/preprints202602.0146.v1

In terms of results interpretation, this research mainly analyzes the emotional perception pattern with the visual information observable in SVIs to enhance the comprehensibility of spatial diagnosis. Relevant conclusions should be regarded as explanatory associations rather than causal inferences. Future studies can be conducted through combining objective spatial indicators or behavioral data.

7. Conclusions

This paper proposes and validates a street-level emotional perception analysis method by performing lightweight domain adaptation on the CLIP model for urban regeneration contexts, enabling computational analysis of six-dimensional emotional perception without emotion annotations. On this basis, the emotional perception results are systematically introduced into the spatial analysis of the city, providing a quantitative path for analyzing the emotional perception evoked by street environment in the context of urban regeneration.

The result shows that the emotional perception based on SVI is not randomly distributed, but presents a continuous and organized spatial pattern in the street network, demonstrating the feasibility and stability of the proposed method. By organizing the multidimensional perceptual results to the dual-axis framework, this paper avoids the simplified processing of compressing complex spatial experiences into single indicators or discrete types, making it possible to diagnose the current perceptual condition of the streets.

The main methodological contribution of this paper is to construct a interpretable and scalable street-level emotional perception analysis framework for urban regeneration contexts. By correlating the emotion assessment results with recognizable visual elements in SVIs, this method enhances the interpretive potential of the results while maintaining computational operability. In practical applications, the framework supports large-scale screening of street networks in the absence of survey data. It helps identify street segments with imbalanced emotional perception and provides information support for subsequent field investigations or regeneration strategies.

Overall, this paper presents a research approach that integrates SVIs, multidimensional emotional perception and structured spatial analysis, providing a methodological reference for emotion-oriented street diagnosis in the context of urban regeneration. Future study could research with social surveys, behavioral data or SVIs of different years to cross-validate the model inference results, thus extending the applicability of the framework across different urban contexts and regeneration stages.

Author Contributions: Conceptualization, L.C.; methodology, L.C. and K.Z.; software, L.C.; validation, L.C. and K.Z.; formal analysis, L.C.; investigation, L.C.; resources, L.C. and K.Z.; data curation, L.C. and K.Z.; writing—original draft preparation, L.C.; writing—review and editing, L.C. and K.Z.; visualization, L.C.; supervision, K.Z.; project administration, L.C. and K.Z.; funding acquisition, L.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The road network data used in this study are publicly available from OpenStreetMap via Geofabrik (https://download.geofabrik.de/, accessed on 25 January 2026). SVIs were obtained through the Baidu Maps Street View API and Google Maps API under each platform's terms of use. A limited number of representative SVIs are displayed in this paper for illustrative and analytical purposes only. These images are not redistributed as part of the dataset. Street sampling point coordinates, grid aggregation units, image-text semantic descriptions generated for CLIP domain adaptation, and emotional perception inference results were produced by the authors following the procedures described in the Methods section. Derived datasets and analysis results are available from the corresponding author upon reasonable request, subject to licensing restrictions of the original data sources.

Acknowledgments: I would like to express my deepest gratitude to Professor Wei Wang for her invaluable guidance and continuous encouragement throughout this research. Her mentorship and advice were essential to the successful completion of this study.

 doi:10.20944/preprints202602.0146.v1

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CLIP	Contrastive Language-Image Pre-training
VLM	Vision-language Model
SVI	Street View Image
OSM	OpenStreetMap
GIS	Geographic Information System
API	Application Programming Interface

Appendix A

This table provides the complete set of text prompts used for CLIP-based street-level perceptual analysis. For each perceptual dimension, multiple descriptive prompts were designed to capture different visual elements observable in street-view imagery.

Table A1. Text prompts used for emotion-oriented perceptual dimensions.

Perceptual Dimension	Prompt	ID
Comfort	a street scene with greenery providing shade along pedestrian sidewalks	C1
	a street scene with clean, even pavement suitable for comfortable walking	C2
	a street scene with seating and resting spaces for pedestrians	C3
	a street scene with human-scale buildings and generous walking space	C4
	a street scene with trees and a calm pedestrian-oriented environment	C5
	a street scene with smooth sidewalks and low physical walking effort	C6
Vitality	a street scene with many pedestrians and visible street activity	V1
	a street scene with active storefronts and people walking	V2
	a street scene with cafes, shops, and outdoor street activity	V3
	a street scene with frequent social interaction and pedestrian movement	V4
	a street scene with lively commercial activity at street level	V5
	a street scene with people gathering and moving along the street	V6
Safety	a street scene with good lighting and clear visibility of surroundings	S1
	a street scene with clear spatial boundaries and predictable pedestrian paths	S2
	a street scene with visible pedestrian routes and unobstructed sidewalks	S3
	a street scene with open street edges and clear spatial structure	S4
	a street scene with well-lit sidewalks and unobstructed sightlines	S5
	a street scene with an easily readable and navigable street layout	S6
Oppressiveness	a street scene with narrow sidewalks and strong spatial enclosure	O1
	a street scene with dense buildings and limited open space	O2
	a street scene with visual clutter and crowded surroundings	O3
	a street scene with tall buildings creating a closed street environment	O4
	a street scene with little sky visibility and strong enclosure	O5
	a street scene with crowded street space and visual congestion	O6
Nostalgia	a street scene with traditional architecture and visible historic character	N1

 doi:10.20944/preprints202602.0146.v1

	a street scene with old buildings and aged facades	N2
	a street scene with historic materials and architectural details	N3
	a street scene with visible traces of past urban development in buildings	N4
	a street scene with traditional storefronts and historic street elements	N5
	a street scene with an old urban character reflected in architecture	N6
Alienation	a street scene with oversized spaces and few pedestrians	A1
	a street scene with blank facades and inactive street edges	A2
	a street scene with large-scale buildings and limited human presence	A3
	a street scene with isolated functions and little street activity	A4
	a street scene with wide empty spaces and lack of social interaction	A5
	a street scene with inactive surroundings and minimal pedestrian presence	A6

References

1. Lynch, K. *The Image of the City*; MIT Press: Cambridge, MA, USA, 1960.
2. Manzo, L.C. Beyond house and haven: Toward a revisioning of emotional relationships with places. *Journal of Environmental Psychology* **2003**, *23*, 47–61. https://doi.org/10.1016/S0272-4944(02)00074-9
3. Seamon, D. Place Attachment and Phenomenology: The Dynamic Complexity of Place. In *Place Attachment: Advances in Theory, Methods and Research*, 2nd ed.; Manzo, L. C., Devine-Wright, P., Eds.; Routledge: New York, NY, USA, 2021; pp. 29–44. https://doi.org/10.4324/9780429274442-2.
4. Cullen, G. *The Concise Townscape*; Architectural Press: London, UK, 1961.
5. Hillier, B.; Hanson, J. *The Social Logic of Space*; Cambridge University Press: Cambridge, UK, 1984.
6. Relph, E. *Place and Placelessness*; SAGE Publications Ltd: London, UK, 2008.
7. Tuan, Y.-F. *Space and Place: The Perspective of Experience*; University of Minnesota Press: Minneapolis, MN, USA, 2001.
8. Biljecki, F.; Ito, K. Street view imagery in urban analytics and GIS: A review. *Landscape and Urban Planning* **2021**, *215*, 104217. https://doi.org/10.1016/j.landurbplan.2021.104217
9. Tang, F.; Zeng, P.; Wang, L.; Zhang, L.; Xu, W. Urban perception evaluation and street refinement governance supported by street view visual elements analysis. *Remote Sensing* **2024**, *16*, 3661. https://doi.org/10.3390/rs16193661
10. Zhang, J.; Li, Y.; Fukuda, T.; Wang, B. Urban safety perception assessments via integrating multimodal large language models with street view images. *Cities* **2025**, *165*, 106122. https://doi.org/10.1016/j.cities.2025.106122
11. Mitchell, L.; Frank, M.R.; Harris, K.D.; Dodds, P.S.; Danforth, C.M. The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PLOS ONE* **2013**, *8*, e64417. https://doi.org/10.1371/journal.pone.0064417
12. Zhou, J.; Kuang, Y.; Wu, S.; Zhang, H.; Wei, J. Research on emotional space identification and regeneration strategies for urban neighborhoods based on large language models: A case study of the North Sichuan Road neighborhood in Shanghai. *Shanghai Urban Planning Review* **2025**, *2*, 32–39. https://doi.org/10.11982/j.supr.20250205
13. Ma, H.; Wu, D. A natural language processing-based approach: Mapping human perception by understanding deep semantic features in street view images. *arXiv* **2023**, arXiv:2311.17354. https://arxiv.org/abs/2311.17354
14. Pendlebury, J.; Short, M.; While, A. Urban World Heritage Sites and the problem of authenticity. *Cities* **2009**, *26*, 349–358. https://doi.org/10.1016/j.cities.2009.09.003
15. Lewicka, M. Place attachment, place identity, and place memory: Restorations of attachment. *Journal of Environmental Psychology* **2008**, *28*, 209–231. https://doi.org/10.1016/j.jenvp.2008.02.001

 doi:10.20944/preprints202602.0146.v1

16. Kang, J.; Körner, M.; Wang, Y.; Taubenböck, H.; Zhu, X.X. Building instance classification using street view images. *ISPRS Journal of Photogrammetry and Remote Sensing* **2018**, *145*, 44–59. https://doi.org/10.1016/j.isprsjprs.2018.02.006
17. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
18. Radford, A.; Kim, J.W.; Hallacy, C.; et al. Learning transferable visual models from natural language supervision. *arXiv* **2021**, arXiv:2103.00020. https://doi.org/10.48550/arXiv.2103.00020
19. Haas, L.; Alberti, S.; Skreta, M. Learning generalized zero-shot learners for open-domain image geolocalization. *arXiv* **2023**, arXiv:2302.00275. https://arxiv.org/abs/2302.00275
20. Huang, W.; Wang, J.; Cong, G. Zero-shot urban function inference with street view images through prompting a pretrained vision-language model. *International Journal of Geographical Information Science* **2024**, *38*, 1414–1442. https://doi.org/10.1080/13658816.2024.2347322
21. Zhang, S.; et al. Learning emotion representations from verbal and nonverbal communication. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 19267–19278. 10.1109/CVPR52729.2023.01821
22. Huang, G.; Jiao, H. UP-CLIP: A domain-specialized multimodal framework for fine-grained urban perception extraction from street-view imagery. *SSRN* **2025**. https://doi.org/10.2139/ssrn.5895454
23. Ewing, R.; Handy, S. Measuring the unmeasurable: Urban design qualities related to walkability. *Journal of Urban Design* **2009**, *14*, 65–84. https://doi.org/10.1080/13574800802451155
24. Naik, N.; Philipoom, J.; Raskar, R.; Hidalgo, C.A. Streetscore—Predicting the perceived safety of one million streetscapes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Columbus, OH, USA, 23–28 June 2014; pp. 779–785. https://doi.org/10.1109/CVPRW.2014.121
25. Dubey, A.; Naik, N.; Parikh, D.; Raskar, R.; Hidalgo, C.A. Deep learning the city: Quantifying urban perception at a global scale. In Computer Vision—ECCV 2016; Springer: Cham, Switzerland, 2016; pp. 196–212. https://doi.org/10.1007/978-3-319-46448-0_12
26. Zhong, W.; Wang, L.; Han, X.; Gao, Z. Spatiotemporal analysis of urban perception using multi-year street view images and deep learning. *ISPRS International Journal of Geo-Information* **2025**, *14*, 390. https://doi.org/10.3390/ijgi14100390
27. Huang, Y.; et al. Dynamics of street environmental features and emotional responses in urban areas: Implications for public health and sustainable development. *Frontiers in Public Health* **2025**, *13*, 1589183. https://doi.org/10.3389/fpubh.2025.1589183
28. Zhao, W.; Tan, L.; Niu, S.; Qing, L. Assessing the impact of street visual environment on the emotional well-being of young adults through physiological feedback and deep learning technologies. *Buildings* **2024**, *14*, 1730. https://doi.org/10.3390/buildings14061730
29. Russell, J.A. A circumplex model of affect. *Journal of Personality and Social Psychology* **1980**, *39(6)*, 1161–1178. https://doi.org/10.1037/h0077714
30. Zhang, F.; Zhou, B.; Liu, L.; Wang, Y.; Ratti, C. Measuring human perceptions of a large-scale urban region using machine learning. *Landscape and Urban Planning* **2018**, *180*, 15–27. https://doi.org/10.1016/j.landurbplan.2018.08.020
31. Mehrabian, A.; Russell, J.A. *An Approach to Environmental Psychology*; MIT Press: Cambridge, MA, USA, 1974.
32. Smith, L. *Uses of Heritage*; Routledge: London, UK, 2006.
33. Mehta, V. Evaluating public space. *Journal of Urban Design* **2014**, *19*, 53–88. https://doi.org/10.1080/13574809.2013.854698
34. Carmona, M. The place-shaping continuum: A theory of urban design process. *Journal of Urban Design* **2014**, *19*, 2–36. https://doi.org/10.1080/13574809.2013.854695
35. Gehl, J. *Life Between Buildings: Using Public Space*, 6th ed.; Island Press: Washington, DC, USA, 2011.
36. Montgomery, J. Making a city: Urbanity, vitality and urban design. *Journal of Urban Design* **2007**, *3*, 93–116. https://doi.org/10.1080/13574809808724418

 doi:10.20944/preprints202602.0146.v1

37. Chan, S.H.G.; Lee, W.H.H.; Tang, B.M. Legacy of culture heritage building revitalization: Place attachment and cultural identity. *Frontiers in Psychology* **2023**, *14*, 1314223. https://doi.org/10.3389/fpsyg.2023.1314223
38. Roberts, P.; Sykes, H. *Urban Regeneration: A Handbook*; SAGE Publications Ltd, 2008. https://doi.org/10.4135/9781446219980
39. Wu, F. *Planning for Growth: Urban and Regional Planning in China*; Routledge: New York, USA, 2015. https://doi.org/10.4324/9780203067345
40. Haklay, M. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Urban Analytics and City Science* **2010**, *37*, 682–703. https://doi.org/10.1068/b35097
41. Barrington-Leigh, C.; Millard-Ball, A. The world's user-generated road map is more than 80% complete. *PLOS ONE* **2019**, *14(10): e0224742*. https://doi.org/10.1371/journal.pone.0180698
42. Jacobs, J. The Death and Life of Great American Cities; Knopf Doubleday Publishing Group, 1992.
43. Zhang, Y.; Li, X.; Wang, A.; Bao, T.; Tian, S. Density and diversity of OpenStreetMap road networks in China. *Journal of Urban Management* **2015**, *4*, 135–146. https://doi.org/10.1016/j.jum.2015.10.001