A Multidimensional AI-powered Framework for Analyzing Tourist Perception in Historic Urban Quarters: A Case Study in Shanghai

Kaizhen Tan[1*], Yufan Wu[2]†, Yuxuan Liu[2]†, Haoran Zeng[2]†

[1] School of Economics and Management, Tongji University, Shanghai 200092, China
[2] College of Architecture and Urban Planning, Tongji University, Shanghai 200092, China

*Corresponding author. E-mail: wflps20140311@gmail.com; ORCID: 0009-0002-6459-578X
†These authors contributed equally to this work.

Abstract. Historic urban quarters play a vital role in preserving cultural heritage while serving as vibrant spaces for tourism and everyday life. Understanding how tourists perceive these environments is essential for sustainable, human-centered urban planning. This study proposes a multidimensional AI-powered framework for analyzing tourist perception in historic urban quarters using multimodal data from social media. Applied to twelve historic quarters in central Shanghai, the framework integrates focal point extraction, color theme analysis, and sentiment mining. Visual focus areas are identified from tourist-shared photos using a fine-tuned semantic segmentation model. To assess aesthetic preferences, dominant colors are extracted using a clustering method, and their spatial distribution across quarters is analyzed. Color themes are further compared between social media photos and real-world street views, revealing notable shifts. This divergence highlights potential gaps between visual expectations and the built environment, reflecting both stylistic preferences and perceptual bias. Tourist reviews are evaluated through a hybrid sentiment analysis approach combining a rule-based method and a multi-task BERT model. Satisfaction is assessed across four dimensions: tourist activities, built environment, service facilities, and business formats. The results reveal spatial variations in aesthetic appeal and emotional response. Rather than focusing on a single technical innovation, this framework offers an integrated, data-driven approach to decoding tourist perception and contributes to informed decision-making in tourism, heritage conservation, and the design of aesthetically engaging public spaces.

Keywords: Historic urban quarters; Tourist perception; Visual aesthetics; Sentiment analysis; Deep learning.

Statements and Declarations: The authors have no competing interests to declare that are relevant to this paper.

1 INTRODUCTION

Historic urban quarters represent concentrated embodiments of cultural heritage, bearing a city's collective memory and distinct local identity. They increasingly contribute to fostering social cohesion and a sense of place (Lynch, 1964; Assmann, 1995). In recent years, many of these quarters have undergone urban transformation and now function as vital spaces for daily activity, cultural consumption, and tourism. Therefore, understanding how tourists cognitively and aesthetically relate to historic quarters is not only essential for enhancing visitor experiences but also critical for human-centered urban planning and place-based design.

Tourists' perceptions and experiences directly determine the attractiveness, reputation, and sustainability of historic quarters as travel destinations. Their cognitive impressions and emotional responses collectively shape the perceived image of a place, which in turn influences the quarter's appeal, conservation priorities, and future planning strategies (Ginzarly et al., 2019). A robust evaluation framework that captures tourists' authentic preferences and concerns can help urban managers identify the most appealing physical elements, detect service deficiencies, and optimize spatial design. This facilitates both the improvement of visitor satisfaction and the continued transmission of cultural heritage.

Conventional studies on tourist experiences in historic quarters have largely relied on survey-based methods and in-situ observations to assess satisfaction and spatial imagery. While such approaches can offer rich insights, they are often limited by subjectivity, time delays, and scalability constraints. With the proliferation of user-generated content (UGC) on platforms such as Flickr and Instagram, researchers now have access to large volumes of geo-tagged photos, comments, and narratives that reflect how tourists interpret urban space (Zhang et al., 2019; Kang et al., 2021; Cho et al., 2022). Meanwhile, recent advances in artificial intelligence have demonstrated the considerable potential of deep learning to extract structured features from unstructured data and to reveal latent dimensions of urban experience. For example, the emergence of tools like ChatGPT illustrates how AI models can be applied to fields such as urbanism, history, and heritage studies, opening up new analytical possibilities (Batty, 2023).

However, most existing work in this space suffers from several key limitations. First, methodological fragmentation is widespread: few studies integrate deep-learning–based image analysis, color palette comparison, and sentiment modelling into a unified framework. Second, research on tourist perception rarely focuses explicitly on historic urban quarters as bounded cultural spaces. Among the literature reviewed, only Bai et al. (2023) provide a case study centered on a historic quarter, Testaccio in Rome, using multimodal social media data to visualize perceived urban heritage elements through pre-trained image recognition and natural language processing models. Nonetheless, their study does not address the role of subjective aesthetic preferences or emotional satisfaction.

While a small body of literature has begun to examine color characteristics in social media imagery, such studies are confined to user psychology. For instance, recent research has linked Instagram users' colorfulness, harmony, and diversity in their photos to their personality traits, loneliness, and attitudes toward online self-presentation (Kim & Kim, 2019). Although these findings support the idea that color carries latent affective and perceptual signals, they do not account for spatial context or how aesthetic expectations manifest in touristic representation.

When it comes to the urban imagery, existing studies have largely focused on the computational evaluation of color harmony in building façade. Recent work has proposed quantitative frameworks for assessing the harmoniousness of building colors in historic quarters, arguing that color harmony is a key factor in achieving coherent urban landscapes (Zhou et al., 2022; Yang et al., 2024). These models typically rely on formal aesthetic principles to evaluate whether façade colors appear balanced or coordinated. However, such approaches do not directly address whether higher levels of harmony translate into greater public satisfaction, nor do they capture how people subjectively judge urban beauty. In other words, they cannot explain why certain urban scenes are still perceived as visually unappealing, even when they conform to predefined aesthetic standards. In addition, recent research has shown that local color palettes evolve over time and accumulate complex meanings influenced by historicization, ethnicization, and commercialization (Xue et al., 2025). Their analysis, based on five dimensions including dominant colors, color complexity, color harmony, average saturation, and average value, provides a valuable retrospective perspective on chromatic transformation in heritage sites. However, this site-centred, retrospective approach largely overlooks how contemporary observers (e.g., tourists) perceive, interpret, or even anticipate these dynamic shifts in urban color.

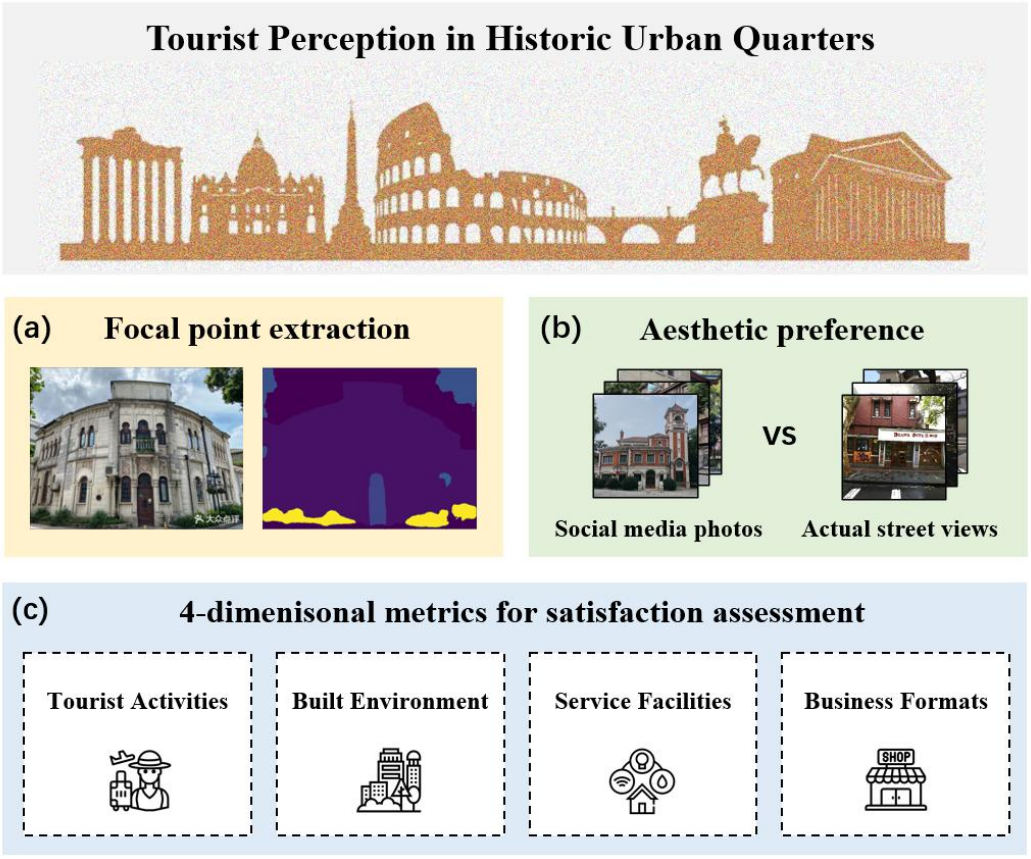


Fig. 1 Conceptual overview of the AI-powered framework **(a)** Semantic segmentation is used to identify the visual focus areas from tourist photos **(b)** A comparative analysis of dominant colors and composition between social media photos and actual street views reveals tourists' aesthetic expectations **(c)** Tourist satisfaction is evaluated across four dimensions: Tourist Activities, Built Environment, Service Facilities, and Business Formats

To move beyond this technical focus, it is necessary to explore the gap between the city's actual visual output and people's internal aesthetic expectations. Understanding this divergence requires not only measuring what is physically present, but also decoding what individuals desire or idealize in terms of urban visual experience. This study addresses this issue by comparing the color compositions of tourist-shared photographs with those of actual street views in historic urban quarters, thereby uncovering potential misalignments between perception and reality.

The study proposes a multidimensional AI-powered framework to decode tourist perceptions of historic urban quarters, with an exploratory case study conducted in Shanghai. As illustrated in Figure 1, the framework integrates three analytical modules: visual focus detection, aesthetic preference analysis, and satisfaction assessment. The main contributions are as follows:

- We propose the first framework for historic urban quarters that jointly analyzes visual content, color preferences, and emotional responses to reveal how tourists cognitively and emotionally engage with these environments.
- A semantic segmentation model is developed to detect focal attention areas in historic urban context, and a multi-task BERT model is developed to assess satisfaction across multiple perceptual dimensions based on textual reviews
- Color distribution is introduced as an abstract representation of aesthetic preference, and a novel comparative analysis is conducted between tourists photos and actual street views to uncover potential mismatches, offering new insight into unmet visual expectations.

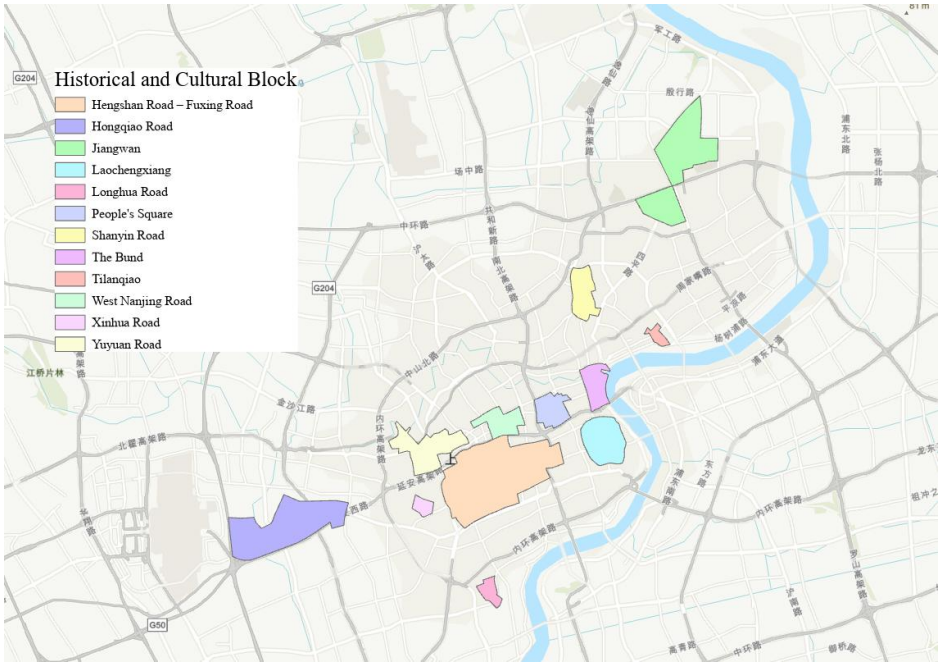


Fig. 2 Historical and cultural blocks in central Shanghai.

2 METHODOLOGY

2.1 Case Study

This study selects Shanghai as the study area. The historical and cultural blocks designated by the Shanghai municipal government refer to areas where historic buildings are densely distributed, and where architectural styles, spatial layouts, and streetscapes collectively reflect the cultural characteristics of a specific historical period in Shanghai. Figure 2 shows 12 historical and cultural blocks located in the central urban area of Shanghai, namely: People's Square, West Nanjing Road, The Bund, Shanyin Road, Yuyuan Road, Tilanqiao, Xinhua Road, Jiangwan, Laochengxiang, Hongqiao Road, Hengshan Road–Fuxing Road, and Longhua Road. These areas integrate distinctive styles from different stages of Shanghai's urban development, and reflect the city's modern achievements and evolution in terms of economy, culture, and everyday life.

2.2 Data Preparation

The overall workflow of the proposed framework is illustrated in Figure 3, which outlines the key steps of data collection, preprocessing, model training, and perceptual analysis. We collected tourist photos and review texts related to the selected historic urban quarters from the Dianping mobile app. Deep learning techniques were then applied to analyze tourists' visual focus areas, aesthetic preferences, and emotional responses. To support the color distribution comparison, street view images were also retrieved from Baidu Maps. Sampling was conducted at 20-meter intervals along the streets within each historic block, capturing views in four cardinal directions: 0°, 90°, 180°, and 270°.

To construct a semantic segmentation dataset, the pre-trained Segment Anything Model (SAM) (Kirillov et al., 2023) was used as an auxiliary tool to assist with manual annotation. A high-quality training set was created by manually labeling 1,000 tourist photos. A Yolov8-Seg model (Jocher et al., 2023) was then fine-tuned on this initial set to develop an automatic labeling model. This model was subsequently integrated into the annotation software to accelerate the labeling process, enabling the expansion of the training set and the construction of a complete dataset. Initially, 26 categories were defined based on preliminary discussions, aiming to cover the main visual elements present in social media imagery. However, during the annotation process, some categories proved difficult to distinguish or appeared too infrequently in the dataset. Based on descriptive statistics and practical applicability, the final taxonomy was refined to 22 categories. Definitions of these categories are provided in the appendix.

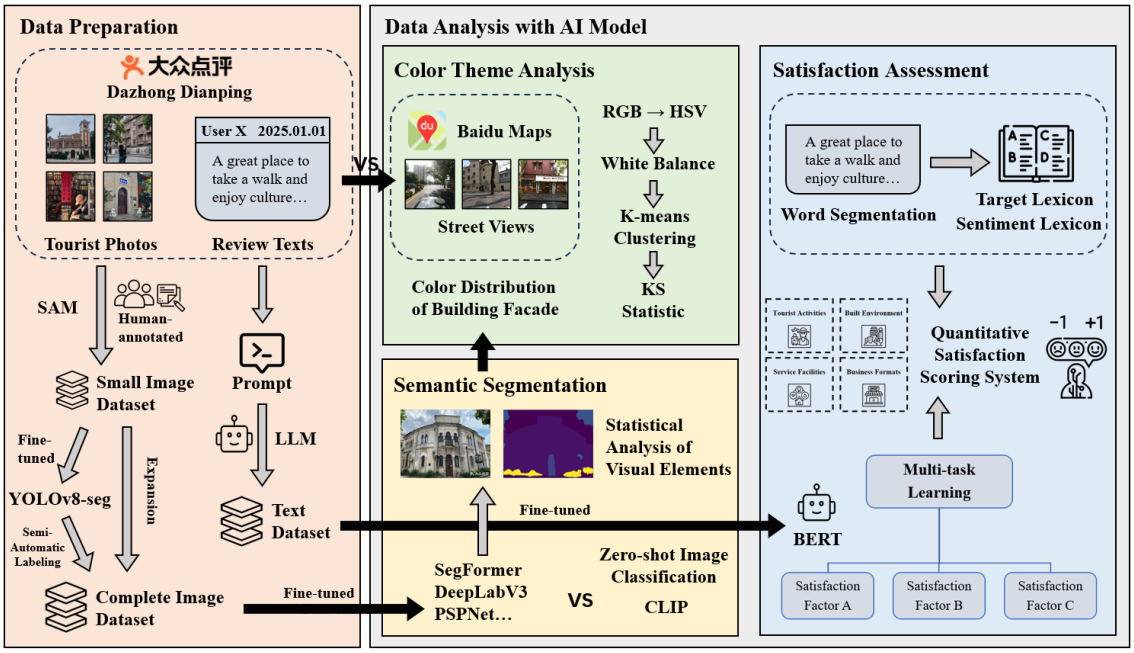


Fig. 3 Technical roadmap of the AI-powered framework

Before constructing the multi-dimensional sentiment scoring dataset for deep learning, a small dataset was built using rule-based named entity recognition and sentiment word matching methods to perform the same task on a limited sample. The four dimensions of satisfaction (Tourist Activities, Built Environment, Service Facilities, and Business Formats) were determined based on prior survey findings, reflecting commonly held visitor concerns. We used the DeepSeek API to generate sentiment labels automatically by leveraging large language model capabilities. For each dimension, sentiment was labeled as {-1, 0, 1}, where -1 indicates dissatisfaction, 1 indicates satisfaction, and 0 denotes either no mention or a neutral attitude. Prompt design details are included in the appendix.

2.3 Semantic Segmentation

No single model has been universally recognized as optimal for semantic segmentation of tourist photographs. Model performance largely depends on dataset characteristics and application contexts. Even minor differences between datasets can lead to noticeable variation in high-dimensional feature mapping. Therefore, several state-of-the-art pre-trained models were evaluated, including SegFormer (Xie et al., 2021), Mask2Former (Cheng et al., 2022), DeepLabV3 (Chen et al., 2018), and PSPNet (Zhao et al., 2017). Fine-tuning experiments showed that Deeplabv3+MobilenetV2 (Sandler et al., 2018) and Mask2Former+Swin-L (Liu et al., 2021) performed well during the first 100 epochs. Considering the trade-offs between parameter size, computational cost, and inference speed, we ultimately adopted Deeplabv3 as the segmentation framework and MobilenetV2 as the backbone network.

Descriptive statistics of the labeled data revealed a significant class imbalance, with certain categories (e.g., fountain, animal, vendor) having substantially fewer samples, which resulted in poor recognition performance for these classes. To address this, class weights were incorporated into the loss function, assigning higher weights to underrepresented categories to increase their training emphasis. The weighting strategy was informed by statistical analysis and inspired by the application of Focal Loss in long-tailed data distributions, aiming to improve small-sample class recognition.

To optimize memory efficiency and prevent overfitting due to excessive parameters, a two-stage training strategy was adopted. In the freezing stage, higher-level features were trained while freezing lower layers to reduce GPU memory usage. In the unfreezing stage, all layers were fine-tuned jointly to maximize model performance. Hyperparameter settings are detailed in the appendix. Finally, the trained segmentation model was applied to tourist photos collected from Dianping, generating statistics on the number and pixel proportions of each category. Results were compared with visual feature detection outputs from the CLIP model (Radford et al., 2021).

2.4 Color Theme Analysis

This study applies color space transformation and clustering algorithms to extract dominant color themes from images. These colors are classified based on the Chinese color system (Huang et al., 2007) to analyze spatial patterns and variations in color distribution across different quarters. Images were first preprocessed by converting from RGB to HSV color space, which better reflects human color perception. The Gray World algorithm was used for white balance correction to mitigate lighting effects and ensure consistency in color representation. K-means clustering was applied to the transformed color data to identify dominant colors in each image. The pixel count for each cluster center was computed, and the dominant colors were then categorized according to the Chinese color system. For each image, a top-5 color palette and corresponding textual data were generated, along with a top-20 color summary. Color distribution curves were fitted, and inter-quarter differences were evaluated using the Kolmogorov-Smirnov (K-S) divergence. Finally, semantic segmentation was applied to both tourist photos and actual street views to compare the distribution of façade colors. This allowed for a detailed analysis of aesthetic alignment or mismatch between perceived and real-world visual elements.

2.5 Satisfaction Assessment

To evaluate tourist satisfaction across four key dimensions, we adopt both traditional lexicon-based methods and a deep learning model. The traditional method involves customized word segmentation, a target lexicon, and a sentiment lexicon. We developed a quantitative satisfaction scoring system: adjectives and evaluation terms are assigned weighted scores, modified by negations and adverbs. Sentence segmentation is guided by fuzzy keyword matching to accurately extract sub-sentences containing sentiment and target terms, improving relevance over general-purpose tokenization.

In parallel, a multi-task BERT model is trained on labeled review texts to predict sentiment polarity in each dimension. We use BERT-base-Chinese as the backbone, with four independent classification heads to output three-level sentiment scores (positive, neutral, negative) per dimension.

3 RESULTS

3.1 Focal Point Extraction

After training the semantic segmentation model, we applied it to the tourist photos to analyze the dominant visual elements in social media imagery. The results demonstrate that the model performs well in identifying core object categories, showing generalization across scenes with different primary subjects such as people and buildings. Figure 4 presents a sample of segmentation results. To evaluate the overall distribution of visual elements, we selected photos from 12 typical historic urban quarters and calculated the total number of pixels and corresponding proportions for each semantic class, excluding the background category. The aggregated results are shown in Figure 5.

Fig. 4 A sample of segmentation results

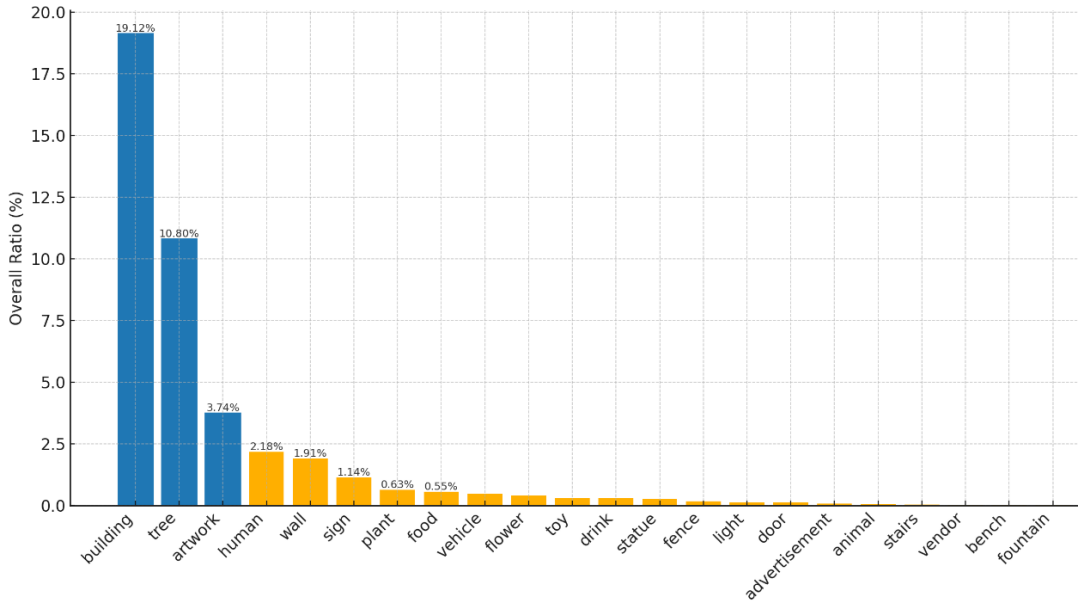


Fig. 5 Overall distribution of semantic classes

Overall, the distribution aligns with intuitive understanding and reflects the composition of real-world streetscapes in historic quarters. Buildings account for the largest share, comprising 19.12% of total pixels, followed by trees at 10.80% and artworks at 3.74%. Other relatively prominent categories include humans, walls, and signs, all exceeding 1%. Elements such as advertisements, street vendors, and fountains occupy minimal proportions, indicating their lower frequency in typical street scenes. Figure 6 presents a heatmap showing the semantic class ratios for each of the 12 quarters. For every quarter, we list the categories with pixel proportions exceeding 1%. Analysis of these distributions in relation to actual urban features reveals several patterns. In leisure-oriented commercial areas such as Hengshan Road, Huaihai Road, Tian'ai Road, and Xinhua Road, greenery and historic buildings dominate the visual composition. These quarters are typically characterized by a high concentration of trees and architectural heritage, which contribute to a comfortable urban atmosphere and highlight their cultural value.

In bustling commercial areas like Huaihai Road and Wukang Road, vehicles appear more frequently, reflecting the concentration of traffic in these zones. For example, Wukang Mansion, a popular photo spot, is often photographed from across the street, resulting in vehicles being captured within the frame and thus increasing the visual share of the vehicle category.

In high-traffic zones such as the Bund and Wujiang Road, humans rather than trees form the second most prominent visual element. This suggests distinct patterns of urban use. The Bund is dominated by waterfront cityscapes with limited greenery, making buildings and people the primary elements. Wujiang Road, known for its food culture, also shows a notably higher proportion of food-related imagery, reflecting its lively street life and commercial appeal.

Certain quarters also display strong artistic character. In particular, Tian'ai Road, Yuyuan Road, and Hengshan Road have significantly higher proportions of artworks compared to other areas. This reflects their role as cultural and creative hubs within the city. For instance, the graffiti wall on Tian'ai Road accounts for an artwork ratio as high as 12.81%. Similarly, Yuyuan Road and Hengshan Road are associated with cultural industries and public art installations. Yuyuan Garden and Shanyin Road also exhibit considerable artistic features, emphasizing their unique cultural atmosphere.

These findings suggest that the semantic segmentation model can adapt well to different street environments and effectively extract core visual elements. The statistical results align closely with the spatial characteristics of the quarters. For example, the heavy pedestrian flows at the Bund and Wujiang Road, the artistic prominence of Tian'ai Road, and the vehicular presence in commercial zones are all clearly reflected in the data. This method of analysis provides valuable insight for urban planning and commercial decision-making. Metrics such as greenery ratio can inform assessments of urban livability, while the proportion of food-related content may serve as a proxy for culinary attractiveness.

While the segmentation-based approach provides fine-grained quantification of visual elements, it has certain limitations. It does not account for overall scene semantics or the relative importance of different elements. To address this, we used the CLIP (Contrastive Language-Image Pre-Training) model, which identifies high-level semantic concepts by matching images with textual labels. Unlike pixel-based segmentation, CLIP offers a conceptual understanding of urban imagery. Although it may miss fine details, it complements segmentation by revealing the dominant meanings conveyed in tourist photos. The distribution of categories detected by CLIP is presented in the appendix.

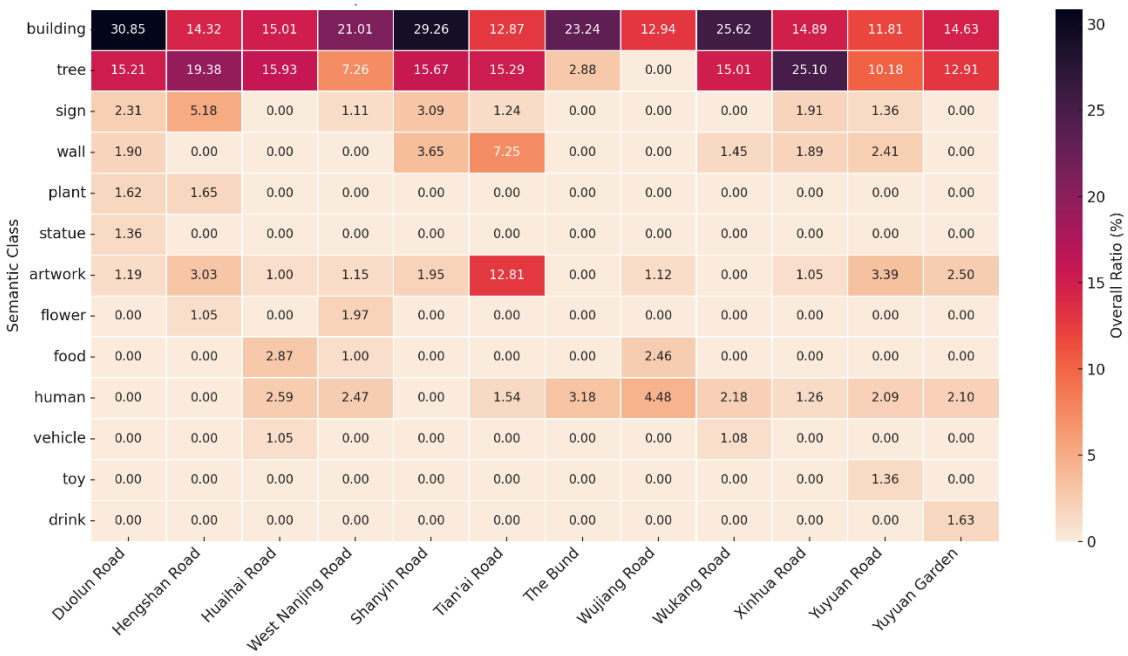
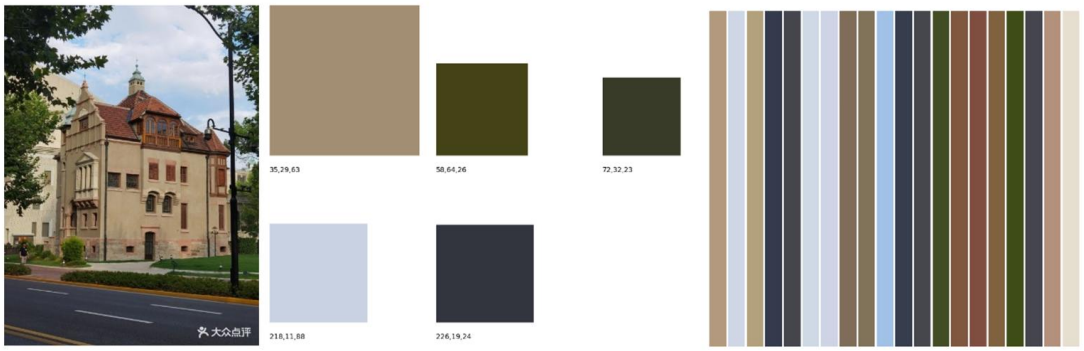


Fig. 6 Heatmap of semantic class ratios across 12 historic urban quarters

3.2 Aesthetic Preference: Color Composition and Expectations

For each tourist photo, we extracted five dominant colors using K-means clustering and visualized them as proportional color blocks. As this method is clustering-based, some color merging and representation errors may occur. In addition, we categorized all pixels into one of 1,000 predefined hues in the Chinese color system and selected the top 20 most frequent colors. While this classification method is less prone to error, it does not consider color grouping and focuses solely on pixel-level frequency. A sample result is shown in Figure 7, where the original image is compared with the K-means color palette and the top 20 classified colors.



(a) (b) (c)

Fig. 7 Example of color extraction results **(a)** Original photo **(b)** K-means clustered top-5 dominant colors **(c)** Top-20 dominant colors using Chinese color system classification

At the quarter level, we analyzed spatial patterns of color composition by examining the distribution of hue (H) values. Figure 8 displays the overall hue distribution across historic quarters along with fitted curves. Both the aggregate and quarter-level distributions show consistent bimodal patterns, with peaks around the orange-red (H ≈ 0–30) and blue-violet (H ≈ 100–120) ranges, indicating that these tones are particularly prevalent in the urban landscape.

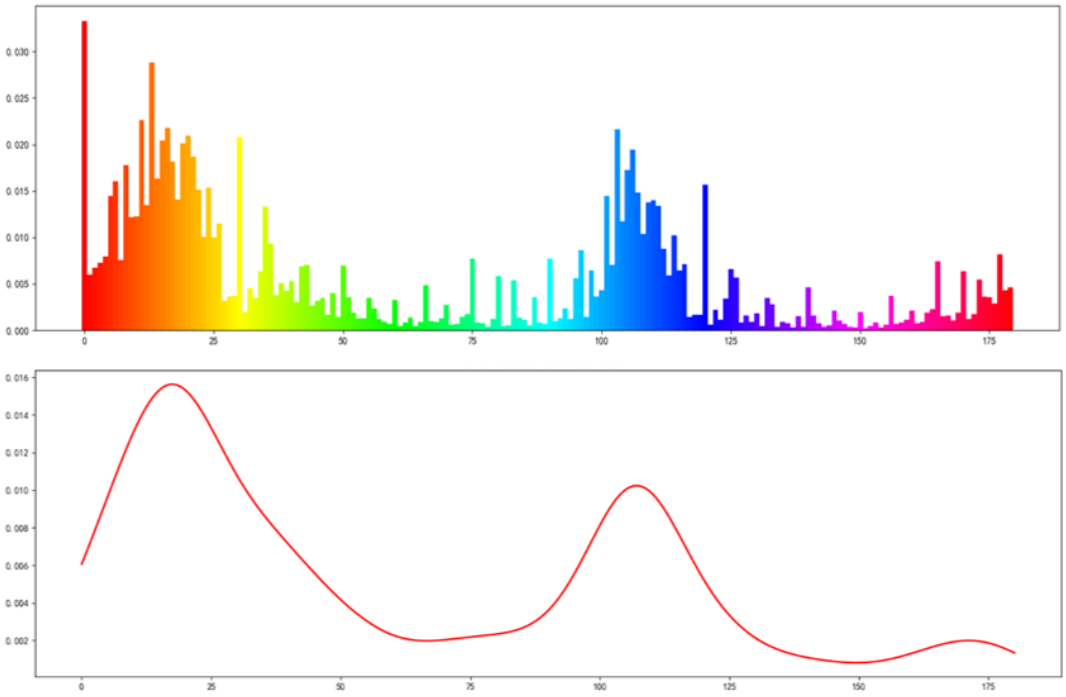


Fig. 8 Overall hue (H) distribution of tourist photos across historic urban quarters, with fitted curve

To quantify inter-quarter differences in color composition, we applied the Kolmogorov–Smirnov (KS) divergence metric. Figure 9 presents a KS divergence matrix of hue distributions across 12 historic urban quarters. A smaller KS value indicates greater similarity in color distribution. Figure 10 compares the hue distributions of Huaihai Road and the Bund. The results show that Fuxing Road and Huaihai Road, both located within the Hengfu historical block, exhibit significant color distribution differences when compared to the Bund, with KS divergence values reaching 0.244 and 0.247, respectively. This divergence may result from the functional characteristics. Quarters characterized by enclosed indoor spaces, such as cultural shops or food streets, tend to exhibit warmer tones (e.g., red, orange, yellow). In contrast, open outdoor areas such as waterfront promenades or historical plazas are dominated by cooler hues (e.g., blue, grey, green). The Bund, situated along the Huangpu River, shows a higher proportion of blue tones influenced by the sky and water bodies, which contributes to its distinctly cooler visual profile.

	Hengshan Road	Huaihai Road	Wukang Road	Yu Garden	Shanyin Road	Tian'ai Road	Duolun Road	The Bund	West Nanjing Road	Wujiang Road	Xinhua Road	Yuyuan Road
Hengshan Road		0.104	0.145	0.117	0.041	0.125	0.082	**0.244**	0.132	0.135	0.047	0.077
Huaihai Road	0.104		0.187	0.158	0.107	0.093	0.121	**0.247**	0.150	0.153	0.060	0.069
Wukang Road	0.145	0.187		0.055	0.110	0.142	0.069	0.126	0.073	0.093	0.137	0.135
Yu Garden	0.117	0.158	0.055		0.085	0.111	0.043	0.128	0.042	0.065	0.108	0.107
Shanyin Road	0.041	0.107	0.110	0.085		0.111	0.049	0.212	0.100	0.101	0.055	0.082
Tian'ai Road	0.125	0.093	0.142	0.111	0.111		0.101	0.171	0.085	0.077	0.081	0.060
Duolun Road	0.082	0.121	0.069	0.043	0.049	0.101		0.166	0.052	0.054	0.074	0.076
The Bund	**0.244**	**0.247**	0.126	0.128	0.212	0.171	0.166		0.118	0.123	0.234	0.205
West Nanjing Road	0.132	0.150	0.073	0.042	0.100	0.085	0.052	0.118		**0.023**	0.121	0.092
Wujiang Road	0.135	0.153	0.093	0.065	0.101	0.077	0.054	0.123	**0.023**		0.126	0.094
Xinhua Road	0.047	0.060	0.137	0.108	0.055	0.081	0.074	0.234	0.121	0.126		**0.034**
Yuyuan Road	0.077	0.069	0.135	0.107	0.082	0.060	0.076	0.205	0.092	0.094	**0.034**	

Fig. 9. KS divergence matrix of hue distributions across 12 historic urban quarters (quarters with similar color distributions are typically part of the same historical and cultural block)

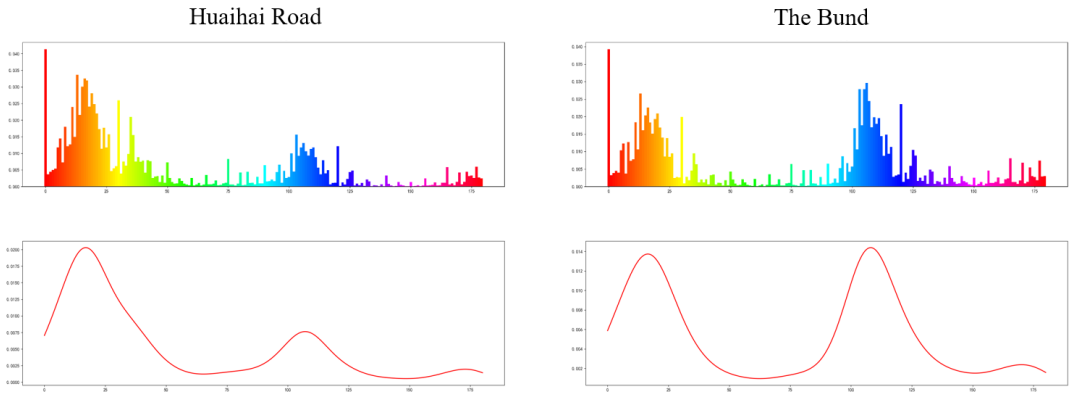


Fig. 10 Comparison of hue (H) distributions and fitted curves for Huaihai Road and the Bund

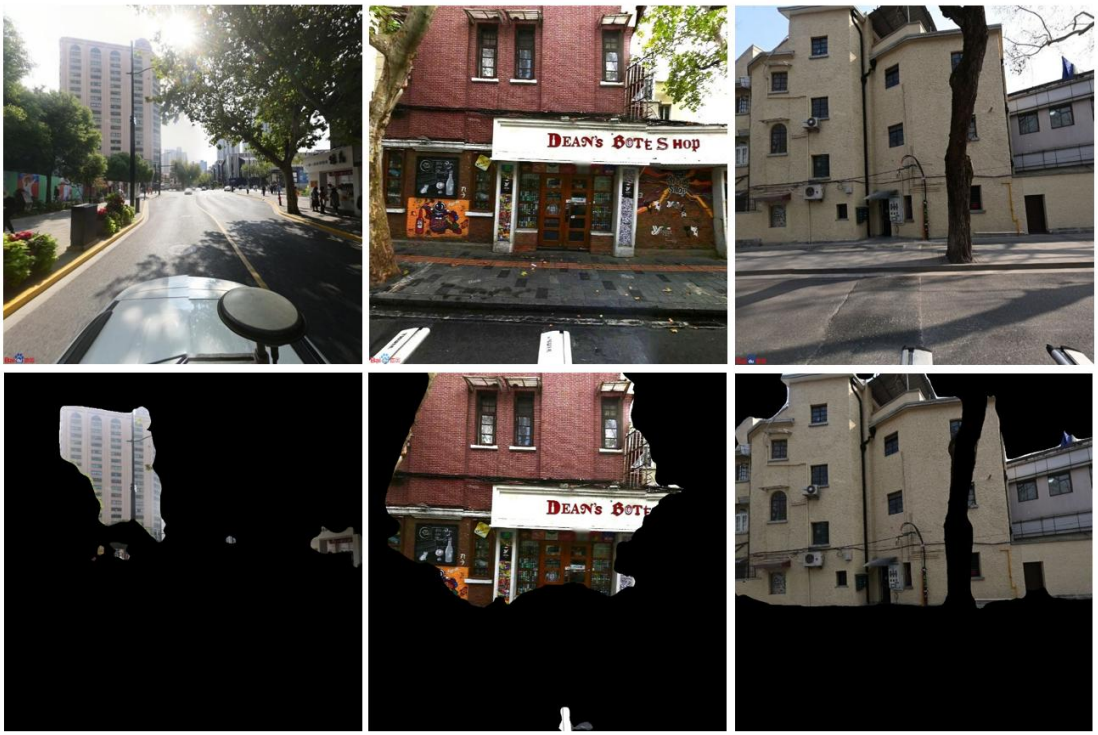


Fig. 11 Building façade segmentation results before and after processing

On the other hand, Wujiang Road and West Nanjing Road—two street quarters within the same historical and cultural block—have a KS divergence of only 0.023. These areas are located less than 50 meters apart, which likely contributes to their similar visual characteristics. A similar pattern is observed between Xinhua Road and Yuyuan Road, both of which are known for their artistic atmosphere and creative industries. Their similar functions may be reflected in their shared color distribution. Despite localized differences, the dominant tones of building façade across most quarters are variations of beige and ochre. This suggests that, at a citywide level, the architectural color palette retains a degree of visual consistency, contributing to a unified urban identity.

To further examine the differences in building color composition between social media photos and real-world street views, we applied semantic segmentation to isolate building façade. This step involved removing indoor scenes, road surfaces, vehicles, sky, and other irrelevant elements to ensure a direct comparison of architectural color. Examples of the segmentation process are shown in Figure 11, which illustrate the contrast before and after isolating building façade.

When comparing the hue distributions of façades extracted from social media photos and street view images, we observed noticeable deviations. As shown in Figure 12, cool hues—green (H = 60) and blue (H = 120)—are significantly more prominent in social media content. In contrast, real-world street views contain a higher proportion of warm tones, particularly within the orange-yellow range (H = 10–30). Several factors were considered to explain this divergence. One possible explanation lies in technical processing. During upload, social media platforms often apply JPEG compression or color space conversions, which may subtly alter color properties. However, such transformations typically affect all hues uniformly and are unlikely to cause systematic shifts in specific hue ranges. Another contributing factor may be environmental conditions at the time of image capture. Natural lighting, weather, and time of day can all influence perceived color. For instance, warm tones tend to dominate during sunrise or sunset, whereas cool tones are more prevalent under cloudy skies or at night. Reflective surfaces such as glass façades or water may also introduce blue and green hues into photos, especially if these elements are more commonly captured in social media imagery. Yet, these context-dependent effects do not fully account for the consistent bias observed across a large number of photos.

We therefore suggest that aesthetic preferences among photographers are the primary cause of the observed hue shift. Prior to uploading, users often enhance images with filters or manual adjustments to achieve a particular visual style. This may involve increasing the saturation of cool tones to create a

cleaner, more transparent look, or reducing warm tones to avoid a yellowish cast. Such stylistic interventions reflect subjective taste and may systematically influence the perceived color distribution in social media content.

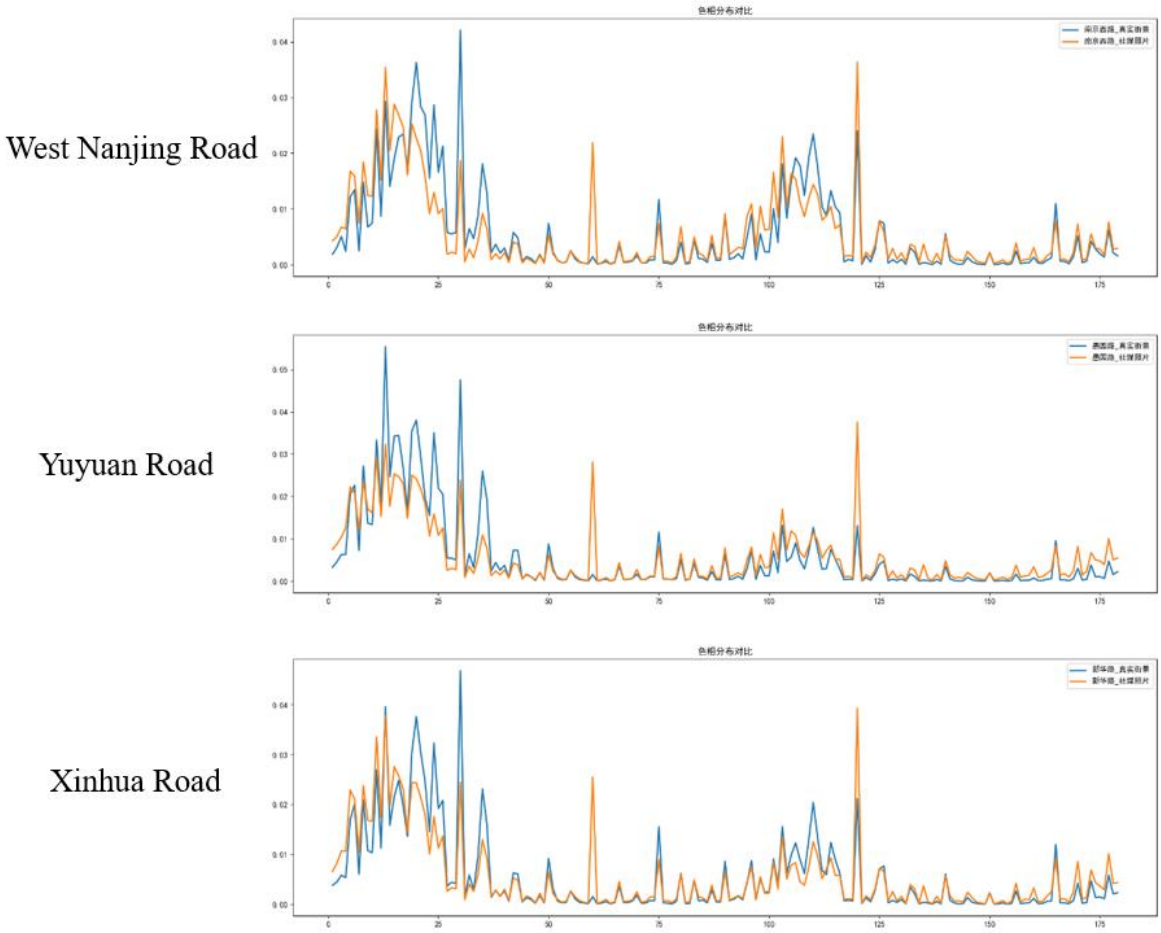


Fig. 12 Comparison of color distributions between social media tourist photos and real-world street views for the West Nanjing Road, Yuyuan Road, and Xinhua Road quarters

3.3 Multidimensional satisfaction assessment

To qualitatively evaluate the multi-task BERT model's performance, we selected several representative tourist reviews and compared the predicted sentiment scores across the four dimensions with manual interpretations. As shown in Table 1, the model provides sentiment predictions ranging from −1 (negative) to 1 (positive) across four dimensions: Tourist Activities, Built Environment, Service Facilities, and Business Formats. These predictions are then examined in Table 2, which presents rational explanations to assess their consistency with human judgment.

Overall, the multi-task BERT model demonstrates strong capability in capturing sentiment across four dimensions: Tourist Activities, Built Environment, Service Facilities, and Business Formats. The predictions are largely consistent with human interpretation, especially in identifying dissatisfaction with commercial composition and praise for specific event-based activities.

Table 1 Predicted sentiment scores for selected tourist reviews across four dimensions

No.	Review Text (Origin)	Review Text (Translated)	Tourist Activities	Built Environment	Service Facilities	Business Formats
1	"今天去了武康大楼，真的很无聊，就搞不懂一栋楼怎么那么多的人，就可能很无聊吧，毕竟在上海打卡地太多了，就旁边的风景还可以，其他的就自行体会，旁边有一个名人故居可以去看看，去散散心也可以，就反正也还行吧"	"I went to Wukang Mansion today. Honestly, it was quite boring. I really don't get why so many people are drawn to just a building. Maybe it's just not that interesting. After all, there are so many other popular spots in Shanghai. The scenery nearby was okay, I guess, but the rest is up to personal experience. There's also a celebrity residence nearby worth visiting if you want a peaceful stroll."	**1**	**-1**	**0**	**-1**
2	"一直没有时间来上海打卡只是刷视频刷到过，看着很不错的样子，过年放假跟男朋友来打卡啦。真的很热闹而且人特别多哦，各个地方人都是来来往往挤的不行，但是玩的地方真的很多，外滩人也特别多，上厕所都要排很长的队哦，下次继续打卡"	"I never had time to visit Shanghai before, only saw it in videos. It looked great, so I finally came here during the New Year holiday with my boyfriend. It was really lively, and there were so many people everywhere—totally packed. But there were also plenty of places to have fun. The Bund was especially crowded, and even the restroom lines were super long. I'll definitely come back again."	**1**	**1**	**-1**	**0**
3	"很商业化了，没什么意思里面很小，没什么可逛的挺没必要花门票进去的。"	"It's way too commercialized and not interesting at all. The place is very small, and there's nothing worth seeing. Honestly, it's not worth the ticket."	**-1**	**-1**	**0**	**-1**
4	"到处是人，今年的灯会还是不错的。小吃价格太贵，关键好不好吃，建议不要在豫园里消费。"	"People everywhere. But this year's lantern festival was actually quite nice. The snacks were way too expensive though—and to be honest, not even that tasty. I wouldn't recommend buying anything inside Yuyuan Garden."	**1**	**1**	**0**	**-1**

Table 2 Explanation of prediction rationality for each dimension

No.	Dimension	Score	Explanation of Prediction Rationality
1	Tourist Activities	**1**	The reviewer explicitly states it was "quite boring" and that there was nothing much to do, indicating low activity appeal. The positive score is thus unreasonable.
1	Built Environment	**-1**	Complaints about overcrowding ("so many people") and a dismissive comment about the building suggest a negative perception of the physical environment. The score is reasonable.
1	Service Facilities	**0**	No mention of specific facilities (e.g., restrooms, seating). The neutral score is reasonable due to insufficient information.
1	Business Formats	**-1**	The reference to "so many other popular spots in Shanghai" implies fatigue or criticism of homogeneity and commercialization. The score is reasonable.
2	Tourist Activities	**1**	The reviewer praises the abundance of fun activities ("plenty of places to have fun"), indicating high satisfaction. The score is reasonable.

2	Built Environment	**1**	Although the environment “looked great,” the reviewer also mentions severe crowding. A neutral rating may have been more appropriate, but the positive score is mostly reasonable.
2	Service Facilities	**-1**	“Long lines for the restroom” is a direct complaint about inadequate facilities. The score is reasonable.
2	Business Formats	**0**	No direct mention of business formats. The neutral score is reasonable.
3	Tourist Activities	**-1**	“Nothing worth seeing” directly reflects a lack of engaging activities. The score is reasonable.
3	Built Environment	**-1**	The phrase “very small” criticizes cramped space, implying poor environmental experience. The score is reasonable.
3	Service Facilities	**0**	The comment about the ticket price may imply dissatisfaction with value-for-money, but no explicit facility issue is raised. The neutral score is somewhat reasonable.
3	Business Formats	**-1**	“Too commercialized” suggests a negative view of homogeneous or overwhelming commercial presence. The score is reasonable.
4	Tourist Activities	**1**	The lantern festival is described positively (“quite nice”), indicating strong activity quality. The score is reasonable.
4	Built Environment	**1**	The review does not directly mention the built environment. The positive score is therefore unreasonable.
4	Service Facilities	**0**	No specific mention of facilities like restrooms or seating. The neutral score is reasonable due to lack of information.
4	Business Formats	**-1**	Complaints about overpriced and low-quality food, with a suggestion not to spend inside, reflect dissatisfaction with commercial offerings. The score is reasonable.

4 DISCUSSION

Although the current model has achieved satisfactory results in semantic segmentation, there is still room for improvement in overall mean Intersection over Union (mIoU). We adopted DeepLabV3+ with MobileNetV2 as the backbone, which offers advantages in speed and model size. However, its performance in capturing fine-grained features under complex scenes may be limited. Future work will explore recent state-of-the-art approaches such as MetaPrompt-SD and InternImage-H. We also plan to improve model robustness by expanding dataset diversity and applying more advanced data augmentation techniques.

In terms of satisfaction prediction, current analysis is based solely on textual reviews. However, user experiences are often closely tied to the physical urban environment. Future research could integrate visual features from street-level imagery—such as architectural style, cleanliness, and greenery—with sentiment scores derived from text. Combining multimodal inputs may improve the interpretability and contextual grounding of perception modeling. Additionally, features like crowd density or façade deterioration could be quantitatively extracted to support more comprehensive evaluations.

Our analysis of urban color schemes also offers several practical insights. First, the "filter effect" introduced by social media platforms should be carefully considered. Applying uncorrected social media images to landscape analysis risks misinterpreting color characteristics. Future studies could explore cross-cultural and cross-platform patterns of color shift, examining how different platforms’ visual styles influence perception. We also plan to develop a reinforcement learning–based method for calibrating social media color distributions against real-world street views, improving data accuracy and reliability.

From an applied perspective, the findings on urban color preferences can inform city planning, architectural design, and visual identity systems. Data-driven color analysis may support the development of locally adapted design guidelines, enhance streetscape coherence, and improve the visual comfort of public spaces. In the context of smart city development, understanding aesthetic preferences can also guide lighting design, public art deployment, and commercial atmosphere planning to create more engaging and human-centered urban environments.

5 CONCLUSION

This study presents an AI-powered framework for analyzing tourist perception in historic urban quarters, based on multimodal data collected from Dazhong Dianping platform. By combining deep learning techniques, the framework investigates tourists' visual attention, aesthetic preferences, and satisfaction across multiple perceptual dimensions.

To extract visual focus areas, we constructed a training dataset by annotating tourist photos with the help of the SAM model and YOLOv8-Seg. A series of state-of-the-art pre-trained models were fine-tuned for semantic segmentation and image classification, enabling the identification of salient elements within street-level imagery. Color distribution was introduced as an abstract representation of aesthetic preference, and K-means clustering was applied to both tourist photos and street view images to extract dominant hues. A comparative analysis of these distributions revealed perceptual mismatches, offering insight into unmet visual expectations in the built environment. For textual analysis, a rule-based approach combining named entity recognition and lexicons was used to calculate satisfaction scores across four dimensions. LLM was employed to generate labeled data, which was then used to train a multi-task BERT model capable of performing automated sentiment classification. Overall, the framework demonstrates strong potential for supporting data-driven urban analysis and offers practical implications for human-centered planning and urban design.

References

[1] Lynch, K. (1964). *The image of the city*. MIT Press.

[2] Assmann, J., & Czaplicka, J. (1995). Collective memory and cultural identity. *New German Critique*, *(65)*, 125–133. https://doi.org/10.2307/488538

[3] Ginzarly, M., Roders, A. P., & Teller, J. (2019). Mapping historic urban landscape values through social media. *Journal of Cultural Heritage*, *36*, 1–11. https://doi.org/10.1016/j.culher.2018.10.002

[4] Zhang, F., Zhou, B., Ratti, C., & Liu, Y. (2019). Discovering place-informative scenes and objects using social media photos. *Royal Society Open Science*, *6*(3), 181375. https://doi.org/10.1098/rsos.181375

[5] Kang, Y., Cho, N., Yoon, J., Park, S., & Kim, J. (2021). Transfer learning of a deep learning model for exploring tourists' urban image using geotagged photos. *ISPRS International Journal of Geo-Information*, *10*(3), 137. https://doi.org/10.3390/ijgi10030137

[6] Cho, N., Kang, Y., Yoon, J., Park, S., & Kim, J. (2022). Classifying tourists' photos and exploring tourism destination image using a deep learning model. *Journal of Quality Assurance in Hospitality & Tourism*, *24*(2), 1–29. https://doi.org/10.1080/1528008X.2021.1995567

[7] Batty, M. (2023). A new kind of search. *Environment and Planning B: Urban Analytics and City Science*, *50*(3), 575–578. https://doi.org/10.1177/23998083231165363

[8] Bai, N., Ducci, M., Mirzikashvili, R., Nourian, P., & Pereira Roders, A. (2023). Mapping urban heritage images with social media data and artificial intelligence: A case study in Testaccio, Rome. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLVIII-M-2-2023*, 139–146. https://doi.org/10.5194/isprs-archives-XLVIII-M-2-2023-139-2023

[9] Kim, J. H., & Kim, Y. (2019). Instagram user characteristics and the color of their photos: Colorfulness, color diversity, and color harmony. *Information Processing & Management*, *56*(4), 1494–1505. https://doi.org/10.1016/j.ipm.2018.10.018

[10] Zhou, Z., Teng, Z., Liu, M., & Ye, Y. (2022). Evaluating building color harmoniousness in a historic quarter intelligently: An algorithm-driven approach using street-view images. *Environment and Planning B: Urban Analytics and City Science*, *50*, Article 239980832211465. https://doi.org/10.1177/23998083221146539

[11] Yang, R., Deng, X., Shi, H., Wang, Z., He, H., Xu, J., & Xiao, Y. (2024). A novel approach for assessing color harmony of historical buildings via street view image. *Frontiers of Architectural Research*, *13*(4), 764–775. https://doi.org/10.1016/j.foar.2024.02.014

[12] Xue, X., Tian, Z., Yang, Y., Wang, J., & Cao, S.-J. (2025). Sustaining the local color of a global city. *Nature Cities, 2*, 400–412. https://doi.org/10.1038/s44284-025-00225-x

[13] Huang, Y., Li, A., Sun, X., He, W., & Zheng, Y. (2007). Introduction to the Chinese color system. *Print Quality & Standard*, *11*, 35–38.

[14] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). Segment anything. *arXiv*. https://arxiv.org/abs/2304.02643

[15] Jocher, G., Chaurasia, A., & Qiu, J. (2023). *YOLO by Ultralytics* (Version 8). Ultralytics. https://github.com/ultralytics/ultralytics

[16] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with Transformers. *arXiv*. https://arxiv.org/abs/2105.15203

[17] Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., & Girdhar, R. (2022). Masked-attention mask transformer for universal image segmentation. *arXiv*. https://arxiv.org/abs/2112.01527

[18] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(4), 834–848. https://doi.org/10.1109/TPAMI.2017.2699184

[19] Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6230–6239). IEEE. https://doi.org/10.1109/CVPR.2017.660

[20] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4510–4520). IEEE. https://doi.org/10.1109/CVPR.2018.00474

[21] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. *arXiv*. https://arxiv.org/abs/2103.14030

[22] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *arXiv*. https://arxiv.org/abs/2103.00020

[23] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv*. https://arxiv.org/abs/1810.04805

Appendix

Semantic Segmentation Category List

- Human
- Door
- Wall
- Building
- Stairs
- Sign
- Advertisement
- Light
- Tree
- Flower
- Plant
- Animal
- Toy
- Food
- Drink
- Vehicle
- Bench
- Fence
- Fountain
- Statue
- Vendor
- Artwork

Semantic Segmentation Hyperparameter Settings

- Initial Epoch: 0
- Frozen Training Epochs: 50
- Unfrozen Training Epochs: 400
- Batch Size (Frozen): 32
- Batch Size (Unfrozen): 16
- Optimizer: SGD
- Initial Learning Rate: 7e-3
- Minimum Learning Rate: 7e-5
- Momentum: 0.9
- Weight Decay: 1e-4
- Learning Rate Schedule: Cosine Annealing
- Mixed precision training (FP16) enabled

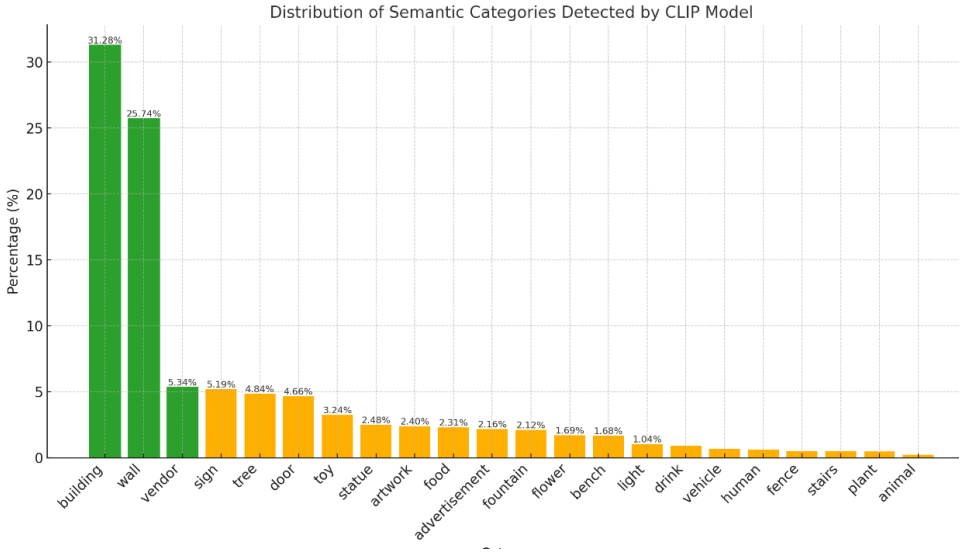


Fig. 13 Distribution of Semantic Categories Detected by CLlP Model

Chinese Color System

In order to standardize color classification, the Chinese color system is adopted with the following structure:

- Five primary hues: Red (R), Yellow (Y), Green (G), Blue (B), Purple (P)
- Five intermediate hues between adjacent primaries: Red-Yellow (YR), Yellow-Green (GY), Green-Blue (BG), Blue-Purple (PB), Purple-Red (RP)
- Ten basic hues are formed from the primaries and intermediates
- Each pair of basic hues is further divided into four equal intervals, producing 40 hue levels (H)
- Saturation (S) and Brightness (V) are each divided into 5 levels, resulting in a total of 40 × 5 × 5 = 1000 color categories

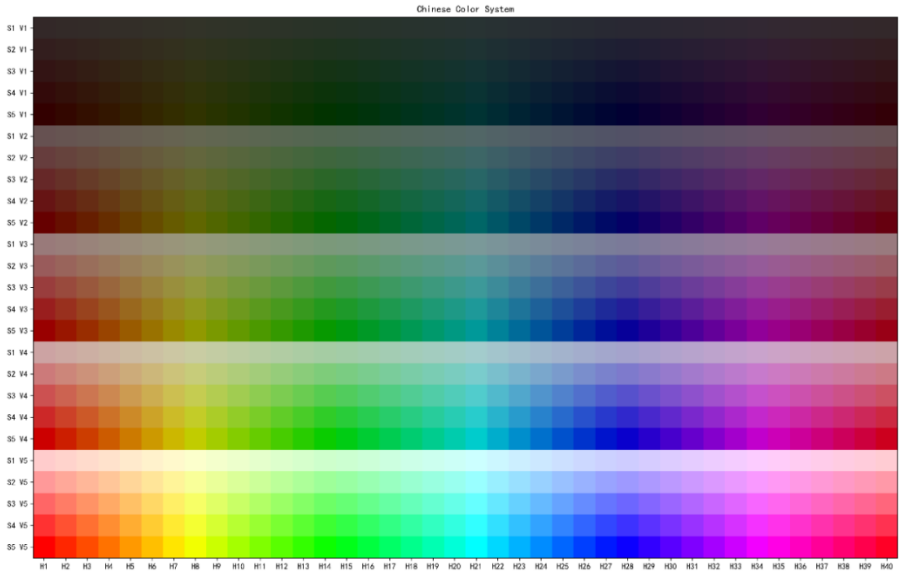


Fig. 14 Structure of Chinese color system

Prompt Design for Sentiment Annotation

请从[活动、建成环境、服务设施、业态]四个维度对评论内容进行满意度打分：

【维度定义】

1. 活动：涉及人群的自发性活动如（吃饭/拍照/游玩/购物等），或是有组织的特色活动（时装秀/展览/市集）

2. 建成环境：包含建筑景观、地标节点、道路环境、自然要素、小品等（如建筑美感/街道清洁/绿化程度）

3. 服务设施：指向公共服务配置、交通运输和游客服务（如公交车/厕所/休息区/指示牌）及服务态度

4. 业态：反映商业多样性及体验（如餐饮/珠宝/娱乐/服饰/住宿/文创零售种类、商品性价比、店铺特色）

【打分规则】

1. 满意=1，中立=0，不满意=-1

2. 直接返回4个数字，用空格分隔

3. 示例：

- "来步行街了，打卡成功！但街道有点脏乱差，而且说实话没啥太多的餐饮选择" → 1 -1 0 -1

- "人太多了 容易发生踩踏事件 不过景色真的很漂亮 建筑群超好看" → -1 1 0 0

请对以下评论打分：

"{text}"

Please rate the sentiment of the review content across the following four dimensions: [Activities, Built Environment, Service Facilities, Business Formats]

[Dimension Definitions]

1. **Activities**: Refers to spontaneous individual activities (e.g., dining, taking photos, sightseeing, shopping) or organized special events (e.g., fashion shows, exhibitions, markets).
2. **Built Environment**: Includes architectural landscapes, landmarks, street environment, natural elements, and small-scale urban features (e.g., aesthetic quality of buildings, street cleanliness, greenery).
3. **Service Facilities**: Refers to the provision of public services, transportation, and tourist assistance (e.g., buses, restrooms, seating areas, signage) as well as service attitude.
4. **Business Formats**: Reflects the diversity and quality of commercial offerings (e.g., food, jewelry, entertainment, fashion, accommodation, cultural/creative retail), perceived value, and shop uniqueness.

[Scoring Rules]

- 1 = Positive (satisfied), 0 = Neutral, -1 = Negative (dissatisfied)
- Return four numbers separated by spaces only
- **Examples**:

- "Visited the pedestrian street—check! But honestly, the street was a bit dirty and lacked dining options." → **1 -1 0 -1**
- "It was overcrowded and felt unsafe, but the view was stunning and the architecture was impressive." → **-1 1 0 0**

Please rate the following review:
"{text}"

BERT-Based Model Configuration

- Backbone: BERT-base-Chinese
- Tokenization: max_length=128, padding=True, truncation=True
- Label Mapping: {-1, 0, 1} → {0, 1, 2}
- Classifier: Four independent Linear(hidden_size, 3) layers with Dropout(0.3)
- Optimizer: AdamW, lr=1e-5, weight_decay=0.2, warmup_ratio=0.2
- Loss Function: CrossEntropyLoss (summed over four dimensions)
- Gradient Clipping: max_grad_norm=1.0
- Batch Size: 8 (train) / 32 (val)
- Epochs: 20
- Precision: Mixed precision (FP16)
- Evaluation Metric: macro_f1_score on each dimension

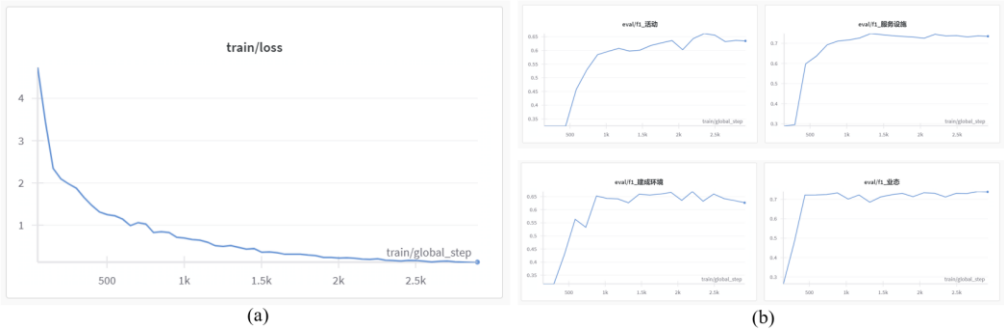


Fig. 15 Multi-task BERT training loss and evaluation F1 scores across 4 dimensions

Fig. 16 Chinese word clouds for the 12 historic urban quarters in Shanghai