Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

How do agricultural heritage landscape elements influence tourist perception? Evidence from online text and images of the Longji Rice Terraces

Maojun Xiong
Beijing Forestry University
Xiyue Wang
wangxy402@bjfu.edu.cn

Beijing Forestry University

Article

Keywords: Longji Rice Terraces, agricultural cultural heritage, visitor perception, online text-and-image reviews, aspect-level sentiment analysis, interpretable machine learning

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11202815/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

How do agricultural heritage landscape elements influence tourist perception? Evidence from online text and images of the Longji Rice Terraces

Abstract: Although tourism development and utilisation help raise public awareness of agricultural cultural heritage, they also run the risk of simplifying its multidimensional value into easily consumable scenic symbols in tourism communication. To analyse this contradiction, this study aims to explore how specific landscape elements of agricultural cultural heritage shape visitors' perceptions and, in turn, influence their overall emotional responses. Taking the Longji Rice Terraces in Guangxi as a case study, this research analyses user-generated online reviews and their accompanying images. By integrating thematic clustering, aspect-level sentiment analysis, and visual language models, the study compares the distribution of attention toward different landscape elements in text and images, as well as their cross-modal co-occurrence relationships, and further examines the association between these features and visitors' overall evaluations. The results indicate significant differences in text-image representation: images primarily capture the seasonal colours and spatial textures of the terraces, while text tends to describe the overall scenery. Visitors' attention is primarily focused on visual features, while agricultural production practices, local knowledge, and sociocultural values are relatively marginalised in text-and-image communication. The study reveals the selective expression characteristics of agricultural cultural heritage in tourism communication, which helps identify the heritage values that are highlighted and those that are downplayed, and provides empirical evidence for optimising the interpretation of agricultural cultural heritage and its sustainable management.

Keywords: Longji Rice Terraces; agricultural cultural heritage; visitor perception; online text-and-image reviews; aspect-level sentiment analysis; interpretable machine learning

1 Introduction

Globally Important Agricultural Heritage Systems (GIAHS) are defined as agricultural socio-ecological systems formed and continuously evolving through long-term human-land interactions, embodying multiple values such as biodiversity, traditional agricultural knowledge, local culture, community organisation, and distinctive landscapes (Fuller et al. 2015; Tian et al. 2016; FAO 2023). Unlike static heritage centred on tangible remains, agricultural cultural heritage relies on ongoing agricultural production, community life, and ecological management practices, and thus exhibits distinct characteristics of vitality, complexity, and adaptability (Rössler and Lin 2018; Min 2021). As tourism at heritage sites continues to develop, the tourism market, visitor flows, and digital media have gradually become significant forces influencing the conservation and dissemination of agricultural cultural heritage. On the one hand, tourism can provide resources for heritage conservation and community development by creating jobs, generating revenue, and raising public awareness; on the other hand, tourism presentations may reorganise heritage content according to market demands and visitor expectations, simplifying complex agricultural production systems and sociocultural relationships into landscape symbols that are easy to view, photograph, and share (Cheng et al. 2019; Sun et al. 2019; Lin et al. 2025). Therefore, determining whether the cultural connotation of heritage is accurately perceived and identifying the mechanisms of selective presentation of living heritage values across multidimensional media have become key issues in balancing heritage conservation and sustainable tourism.

Theories of the tourist gaze and authenticity offer complementary perspectives for understanding this selective presentation. The theory of the tourist gaze argues that tourists do not passively receive all information about a destination but rather

selectively view, document, and disseminate specific objects based on preexisting cultural imaginations, media representations, and tourism expectations (Urry and Larsen 2011). At agricultural cultural heritage sites, landscape elements with strong visual distinctiveness—such as the contours of terraced fields, seasonal colours, and village architecture—are typically more likely to capture tourists' attention and be captured in photographs; in contrast, heritage values that are difficult to convey directly through a single image—such as agricultural production practices, local knowledge, community collaboration, and ecological management—may be downplayed in tourism communication. Research on authenticity further indicates that the presentation of heritage sites to tourists is not a complete replication of local reality, but rather a negotiated expression shaped by the combined influence of local actors, tourism operators, visitor expectations, and communication media (MacCannell 1973; Chhabra et al. 2003; Kim and Jamal 2007; Moore et al. 2021; Gardiner et al. 2022). Visitors' perceptions thus reflect not only the objective landscape characteristics of the destination but also the visibility of heritage values within the tourism context and the ways in which they are communicated.

Existing research has primarily examined tourists' landscape perceptions, place experiences, and behavioural attitudes through the analysis of questionnaires, interviews, and travelogues (Santoro et al. 2020; Yang et al. 2022; Zhang and Xu 2024), and has gradually begun to apply topic modelling and sentiment analysis to process large-scale text data such as online reviews (Tieskens et al. 2018; Chen et al. 2024). However, questionnaires and interviews are limited by sample size and retrospective bias, while traditional text analysis tends to compress evaluation dimensions, making it difficult to achieve precise matching at the element level (Zube et al. 1982; Wu and Li 2014; Jia et al. 2024; Zhang et al. 2024). In recent years, image recognition, vision-language models, and multimodal analysis have begun to be applied to research on destination imagery, visitor preferences, and cultural landscapes (Zhang et al. 2019; Gosal and Ziv 2020; Egarter Vigl et al. 2021; Ma and Furuya 2024; Ren et al. 2024; Wang et al. 2024; Zheng et al. 2025; Chen et al. 2025; Lin et al. 2026). However, most studies have focused on content recognition or multimodal fusion predictions, with few

comparing “what visitors photographed” and “what they discussed” within a unified system of heritage elements. In other words, existing research has not quantitatively compared, within a unified framework of heritage values, the heterogeneity between visitors’ visual captures and textual expressions of perception, nor assessed its impact on overall perception.

The Longji Rice Terraces in Guangxi (third batch of GIAHS) provide a representative case for examining the aforementioned issues. The Longji Terraced Fields are a composite cultural landscape formed through long-term interactions between adaptation to the mountainous environment, rice cultivation, and multi-ethnic settlement. The terraced fields, mountainous water systems, village architecture, and seasonal changes constitute its visually apparent natural-productive landscape, while agricultural knowledge, festive rituals, community collaboration, and livelihood practices embody the sociocultural values underlying the landscape (Shen and Li 2012; Fu 2017; Lu et al. 2017; Bao et al. 2021). Tourism development has provided new resources and opportunities for the maintenance of the terraces, community participation, and local management (Zhu et al. 2021), while simultaneously driving the transformation of local production and living spaces into spaces for tourist viewing and consumption, resulting in uneven visibility of different heritage elements as they are viewed by tourists and disseminated through online platforms (Hua et al. 2018; Cao et al. 2024). Based on this, this study utilises user-generated online reviews and their accompanying images to construct a unified classification system for agricultural cultural heritage landscape elements. It innovatively integrates aspect-level sentiment analysis with visual language models to quantitatively analyse the co-occurrence patterns and degrees of deviation between text and images in heritage representation, and to examine the relationship between this cross-modal interaction and visitors’ overall perceptions.

This study focuses on three core questions: First, to identify differences in the perceived distribution of agricultural cultural heritage elements between images and text; second, to reveal which heritage values are emphasized and which are downplayed in tourism communication; third, to construct a model linking the cross-modal

presentation of landscape elements to overall perception. By addressing these questions, this study expands research on visitor perception from the analysis of text or images alone to cross-modal comparisons at the element level, thereby providing a basis for the interpretation of value, tourism communication, and conservation management of agricultural cultural heritage sites.

2 Methods

This study uses online reviews paired with images as the unit of analysis to examine differences in the representation of agricultural cultural heritage landscape elements in text and images, their cross-modal correspondences, and their association with visitors' overall emotional evaluations. The analysis proceeds in five steps: using BERTopic to identify perception themes and align them with the GIAHS value framework to construct a perception evaluation framework for the Longji Rice Terraces landscape; within the same framework, measuring the attention given to each aspect in both text and images, and validating the measurements through expert blind evaluation; comparing the attention allocation structures of the two modalities; using conditional probability and aspect-level regression to examine cross-modal co-occurrence and its emotional associations; and finally, assessing the incremental value of image features in predicting overall text sentiment (Figure 1).

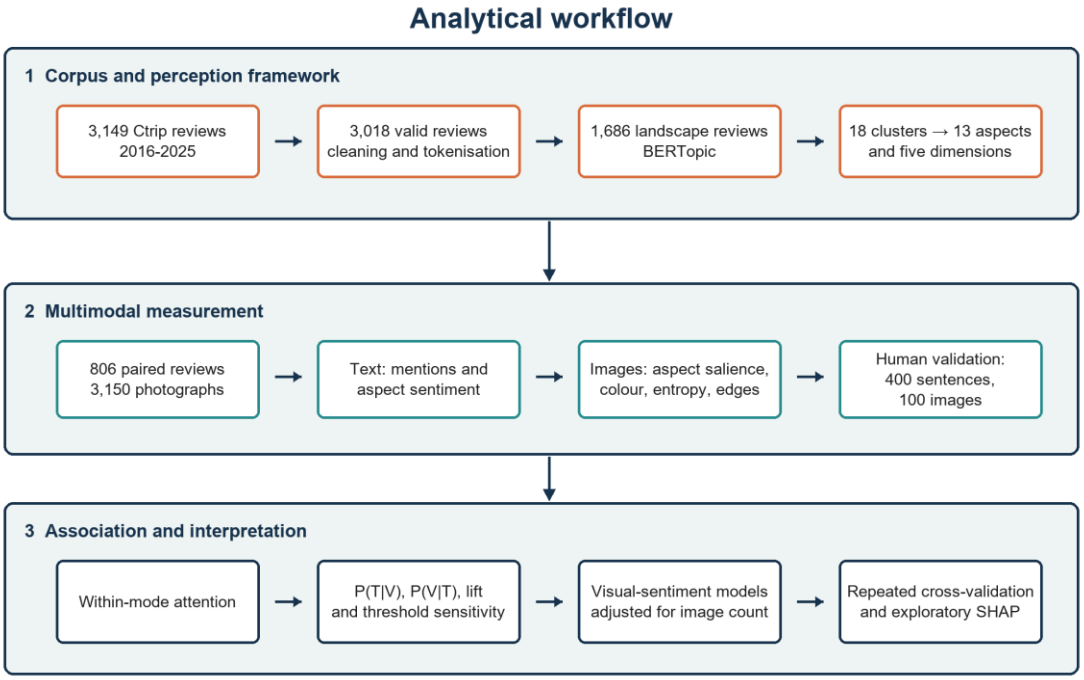


Figure 1. Analytical workflow from review collection and heritage-framework construction to cross-modal comparison and exploratory modelling.

2.1 Case study and data

The case study focuses on the Longji Rice Terraces in Guangxi (the third batch of GIAHS). The heritage value of this site encompasses highly visual landscape features—such as the morphology of the terraces, mountainous water systems, village architecture, and seasonal phenology—as well as living elements—such as agricultural knowledge, festive rituals, and community collaboration—that are difficult to capture in a single image. This makes it suitable for comparing the selective representation of heritage elements in text and images.

The data consisted of user-generated reviews and accompanying images from the Ctrip platform. This study used a distributed web crawler to collect 3,149 raw reviews from 2016 to 2025; duplicate reviews were removed and reviews containing no Chinese text were discarded, leaving 3,018 reviews for the Chinese-language pipeline. To ensure that topic identification focused on landscape perception, a custom stopword list tailored to landscape perception was constructed. Irrelevant terms such as “services,” “accommodations,” and “tickets” were iteratively removed, yielding 1,686

reviews for constructing the perception framework; a random sample of 100 excluded reviews was checked to confirm they contained no landscape perception content. Based on this, 806 comments with images and 3,150 images were selected as the sample for subsequent comprehensive analysis.

Since the decision to upload images may be related to the content of the reviews, this study used Fisher's exact test to compare differences in mention rates across various aspects between the 806 image-containing reviews and the 880 image-free reviews, and applied the Benjamini–Hochberg correction (Benjamini and Hochberg 1995) to define the scope of applicability of the conclusions (see Results for details).

2.2 Construction of the perception evaluation framework

To enable a comparison of text and images across the same set of heritage elements, this study first used BERTopic (Grootendorst 2022) to identify data-driven perception themes, which were then aligned with heritage value theory to avoid omitting elements that might result from directly applying existing scales. Semantic embeddings for Chinese comments were generated using paraphrase-multilingual-MiniLM-L12-v2 (Reimers and Gurevych 2020), dimension-reduced via UMAP (McInnes et al. 2018), and then clustered using HDBSCAN (Campello et al. 2013); Given the limited corpus size and the need for fine-grained thematic analysis, the minimum theme size was set to 10; other parameters are provided in the Supplementary Information.

Topic identification was conducted iteratively. Each round, the top 15 keywords for each topic were examined; non-landscape terms were added to the stopword list, and clustering was performed again until no obvious non-landscape terms appeared. In the final round, 1,686 comments were input, yielding 51 topics. Among these, outlier topics (topic −1) accounted for 18.6%; since texts in this category often contained mixed, multidimensional evaluations, they were retained but not treated as independent perception dimensions. The remaining 50 topics were merged into 18 topic clusters based on semantic similarity of keywords (Supplementary Figure S2).

Subsequently, the topic clusters were theoretically aligned with the value dimensions of GIAHS and the classification of elements in agricultural cultural heritage landscape research (FAO 2023; Fuller et al. 2015; Tian et al. 2016; Santoro et al. 2020). Through manual induction, a perception evaluation framework for the Longji Rice Terraces landscape was developed, comprising 5 dimensions and 13 aspects (Figure 2; Table 1). This framework serves as a common yardstick for both textual and visual measurements, enabling comparability of the attention levels across the two modalities.

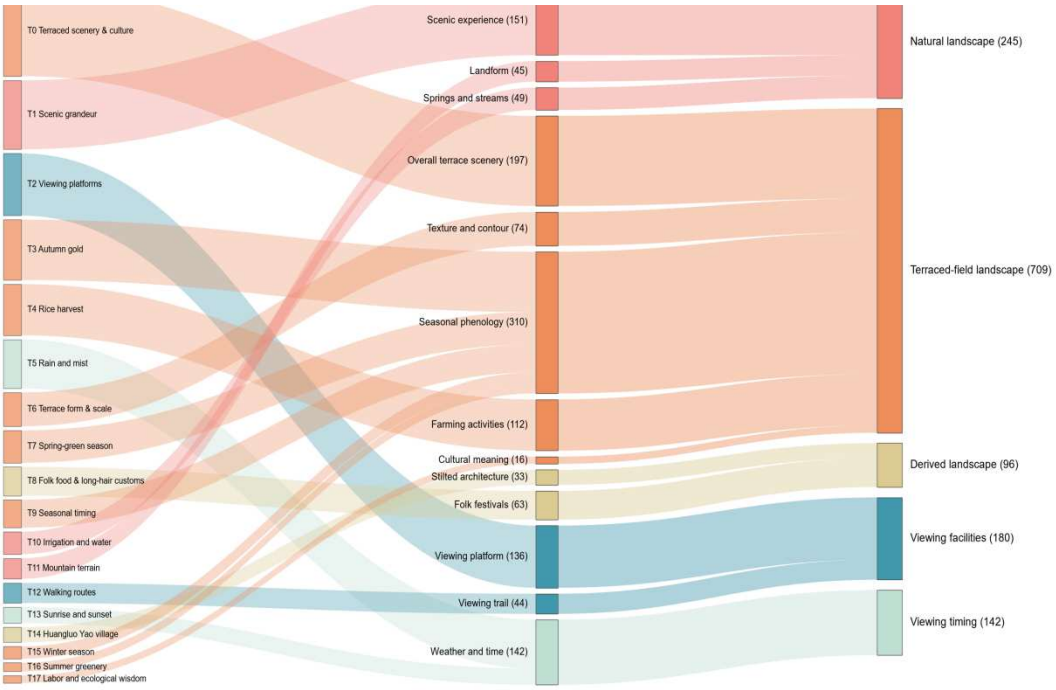


Figure 2. Theory-guided mapping from 18 BERTopic clusters to 13 landscape-perception aspects and five heritage dimensions. Numbers in parentheses are review counts.

Table 1. Perception Evaluation Framework for the Longji Rice Terraces.

Dimensions	Aspect	Topic	Description
Natural landscape	Scenic experience	1	Scenic Beauty
	Topography and landforms	11	Rolling Mountains and Steep Terrain
	Springs and streams	10	Gurgling springs, trickling streams, and stream valleys
Terraced landscape	Overall scenery	0	Spectacular terraced fields and pastoral scenery
	Texture and contour	6	Layered, distinctly tiered rice paddies; naturally curved, winding, and gracefully meandering
	Seasonal phenology	3, 7, 9, 15, 16	Spring: mirror-like rice paddies; Summer: lush green rice fields; Fall: golden rice fields; Winter: snow-covered landscapes
	Farming activities	4	A farming system featuring “two plowings and three harrowings,” a farming method comprising “twelve agricultural tasks,” farming tools, land reclamation and conservation techniques, and farmland irrigation systems
	Cultural connotation	17	The Ecological Wisdom of the Working People
Derived landscape	Stilted architecture	14	Deeply Rooted in Agricultural Production, Cultural Symbols Shared by Multiple Ethnic Groups, Such as Stilt Houses, etc.
	Folk festivals	8	Festive customs closely tied to agricultural life, such as the Plowing Festival, Torch

Dimensions	Aspect	Topic	Description
			Festival, Clothes-Drying Festival, and mountain songs (bend songs, drinking songs), etc.
Viewing facilities	Viewing platform	2	Nine Dragons and Five Tigers, Seven Stars Accompanying the Moon, Golden Buddha's Summit, Thousand-Tiered Heavenly Staircase, and Western Mountain Music
	Viewing trail	12	Summit Trail, Descending Trail
Viewing timing	Weather and time of day	5, 13	Sunrise and sunset, misty conditions, continuous rain

2.3 Measurement and validity testing of text perception

To characterise visitors' linguistic expressions regarding landscape aspects, this study constructed a semantic dictionary of landscape aspects to perform sentence-level matching on reviews to identify the mentioned landscape aspects, and used the Erlangshen-Roberta (Zhang et al. 2022) pre-trained model to estimate the sentiment polarity of the mentioned aspects.

To avoid conflating "not mentioned" with negative evaluations, both descriptive and predictive analyses separately encode whether an aspect is mentioned (D_{ia}) and the conditional sentiment when mentioned (S_{ia}), and use these to construct the normalised aspect sentiment S_{ia}^{norm}.

$$S_{ia}^{\text{norm}} = D_{ia}\frac{S_{ia}+1}{2} \tag{1}$$

Here, i denotes a review and a denotes a landscape aspect; $D_{ia} \in \{0,1\}$ indicates whether review i mentions aspect a; $S_{ia} = P_{ia}(\text{positive}) - P_{ia}(\text{negative}) \in [-1,1]$

is the conditional sentiment; $S_{ia}^{\mathrm{norm}} \in [0,1]$ is the composite variable. Descriptive sentiment analysis considers S_{ia} only when $D_{ia} = 1$.

Text attention measures the distribution of tourists' attention across different aspects in their written narratives. To account for the fact that longer comments may mention more aspects, we employ intra-comment standardisation: the limited linguistic attention in each comment is divided equally among the aspects it mentions, and then averaged across 806 image-text pairs.

$$K_i = \sum_{b=1}^{13} D_{ib}\,, \qquad w_{ia}^{T} = \frac{D_{ia}}{K_i}\ (K_i > 0;\ \text{otherwise } 0), \qquad A_a^{T} = \frac{1}{N}\sum_{i=1}^{N} w_{ia}^{T} \quad (2)$$

Here, b is the summation index over the 13 aspects; K_i is the number of aspects mentioned in comment i; w_{ia}^{T} is the share of textual attention for aspect a within comment i; A_a^{T} is its overall share of textual attention; $N = 806$. This metric reflects the relative distribution of attention within a modality, rather than the raw mention rate.

The validity of the text measurements was verified through a blind expert evaluation: 400 candidate aspect–sentence records produced by the model were drawn at random, and the experts independently judged, for each record, whether the assigned aspect was actually expressed in the sentence and, where it was, the sentiment polarity of that expression, without referring to the model's output; these annotations were then compared record by record with the model's predictions. Aspect identification is therefore evaluated at the candidate level, as the proportion of flagged records that the experts confirmed, and sentiment classification by accuracy, macro-average F1 and per-class precision and recall on the confirmed records; detailed results are provided in the Supplementary Information. Because the sample was drawn from records the model had already flagged, the aspect figures estimate precision but not recall or a full-detection F1, which would require a sample drawn independently of the model's output. This check reduces the risk of semantic misinterpretation in automated analysis.

2.4 Measurement and validity testing of image perception

To ensure that images and text are comparable within the same framework, this study used the open-source vision-language model Qwen2.5-VL-7B-Instruct (Bai et al. 2025) to score the visual salience of 13 landscape aspects in each image (greedy decoding, maximum 512 new tokens, run locally through a single-image API). Here, visual salience refers to the observable evidence supporting the identification of a particular aspect within the image and its prominence, rather than the aspect's importance in terms of heritage value; the scoring is based solely on the image itself, without using descriptive text, image metadata, or external background knowledge. The conceptual definitions of each aspect, observable cues, 0–1 scoring criteria, and complete prompt templates are provided in Supplementary Note 1.

In addition to aspect salience, this study calculates three low-level visual descriptors—colour richness C, Shannon visual entropy H, and Canny edge density E—to characterise the image's colour, information complexity, and structural density. These are included in the predictive model as control variables alongside the semantic features.

$$rg = R - G, \qquad yb = \frac{R+G}{2} - B, \qquad C = \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} + 0.3\sqrt{\mu_{rg}^2 + \mu_{yb}^2} \tag{3}$$

$$H = -\sum_{l} p_l \log_2 p_l, \qquad E = \frac{1}{P}\sum_{p} 1(\text{pixel } p \text{ is an edge}) \tag{4}$$

Here, R, G and B represent the colour channels; rg and yb are the opponent colour channels, and μ and σ denote their means and standard deviations, respectively; p_l is the empirical probability of grayscale level l; and P is the total number of pixels in the image.

Visual attention A_a^V measures the distribution of tourists' attention across different aspects in the image records. When a comment contains multiple images, the aspect salience is aggregated into both the comment-level mean Smean and the maximum value Smax; the former is used for continuous visual attention and regression analysis,

while the latter is used to determine whether a particular aspect is prominently recorded.

$$\mathrm{Smean}_{ia} = \frac{1}{n_i}\sum_{j=1}^{n_i} s_{ija}\,, \qquad w_{ia}^{V} = \frac{\mathrm{Smean}_{ia}}{\sum_{b=1}^{13}\mathrm{Smean}_{ib}}, \qquad A_a^{V} = \frac{1}{N}\sum_{i=1}^{N} w_{ia}^{V} \tag{5}$$

Here, j denotes an image in comment i, n_i is the number of images, s_{ija} is the salience of aspect a in image j, b is the summation index over the 13 aspects, w_{ia}^{V} is the share of visual attention within the comment, and A_a^{V} is the overall share of visual attention; for each valid comment, $\sum_a w_{ia}^{V} = 1$ holds.

The validity of the visual measurements was also verified through expert blind evaluation: 100 images were randomly selected and independently blind-evaluated by two landscape experts on 13 aspects, with scores ranging from 0 to 1 (in 0.1 increments); Inter-rater agreement was reported using ICC(A,1) (McGraw and Wong 1996) and linearly weighted Kappa. Model validity was assessed by calculating MAE, RMSE, and Spearman's ρ for each aspect relative to the mean of the two experts; 95% confidence intervals were obtained by performing 3,000 bootstrap repetitions on a per-image basis. Results for each aspect are presented in the Supplementary Information.

2.5 Cross-modal co-occurrence and its relationship to aspect-specific sentiment

To examine whether visual records and linguistic descriptions mutually manifest each other, this study defines a cross-modal co-occurrence metric based on individual comments. Let $T_{ia} \equiv D_{ia}$ denote text-mentioned events, and let $V_{ia} = 1[\mathrm{Smax}_{ia} > 0.7]$ indicate whether at least one image in comment i prominently depicts aspect a. Based on these definitions, we calculate the linguistic manifestation rate, the visual grounding rate, and Lift.

$$P(T_a|V_a) = \frac{\sum_i T_{ia}V_{ia}}{\sum_i V_{ia}}, \quad P(V_a|T_a) = \frac{\sum_i T_{ia}V_{ia}}{\sum_i T_{ia}}, \quad \mathrm{Lift}_a = \frac{P(T_a, V_a)}{P(T_a)P(V_a)} \tag{6}$$

$P(T_a|V_a)$ represents the linguistic manifestation rate, that is, the proportion of comments with a visually prominent aspect whose text also mentions that aspect;

$P(V_a|T_a)$ represents the visual grounding rate, that is, the proportion of comments mentioning an aspect whose images also depict it prominently. $\text{Lift}_a > 1$ indicates that co-occurrence exceeds the expected value based on independent occurrence at marginal rates. The 95% confidence intervals for conditional probabilities were constructed using 3,000 nonparametric bootstrap samples at the comment level.

The threshold of 0.7 is derived from the 0–1 rating scale in Supplementary Note 1 rather than from post-hoc fitting: on this scale 0.7 denotes an aspect that is clearly prominent but still shares visual attention with others, whereas an aspect close to the subject of the frame scores 0.8 or higher (Supplementary Table S3). A visual-grounding event is therefore defined as a salience score strictly above 0.7, which counts the 0.8–1.0 levels. Binary visual-grounding differences and the aspect-level sample support behind this definition are shown in Supplementary Figure S1, and sensitivity analyses using thresholds of 0.6 and 0.8 are reported in Supplementary Note 3.

To test whether visual salience is associated with specific aspect evaluations, this study constructed aspect-level regression models only for comments in which the text explicitly mentioned the aspect:

$$S_{ia} = \alpha_a + \beta_a Z(\text{Smean}_{ia}) + \gamma_a \log(1 + n_i) + \varepsilon_{ia}, \qquad \text{for } T_{ia} = 1 \tag{7}$$

where S_{ia} is the text sentiment score for aspect a in comment i; Smean_{ia} is the average visual salience for aspect a across all images in that comment; $Z(\cdot)$ denotes standardisation within the valid sample of comments for aspect a; n_i is the number of images; α_a is the intercept; γ_a is the coefficient controlling for the number of images; and ε_{ia} is the error term. β_a and its 95% confidence interval are estimated using 3,000 nonparametric bootstrap repetitions at the comment level; an interval spanning 0 is interpreted as indicating no stable association. Additionally, the significance of β_a is tested using conventional OLS, and the false discovery rate for the 13 aspects is controlled according to the Benjamini–Hochberg procedure. This method transforms qualitative visual information into continuous variables suitable for regression modelling, thereby achieving semantic alignment of multimodal data.

2.6 Overall sentiment and relationship modelling

To obtain an overall metric at the comment level that can be cross-validated with aspect-level perceptions, this study fine-tuned a Chinese three-class sentiment model based on BERT-Base-Chinese to estimate overall text sentiment, using a dataset of 500 manually annotated three-class comments (stratified by label into 400 training and 100 validation comments). The complete training setup is given in Supplementary Note 2.3, and per-class metrics are reported in Section 3.4. The deployed checkpoint, taken from the third epoch, achieved an accuracy of 69.0%, a macro-F1 score of 0.621, and a quadratic-weighted Kappa of 0.593 on the validation set. Misclassifications were concentrated in adjacent categories, while the model demonstrated reliable discrimination between positive and non-positive sentiments. Subsequently, this model was used to perform batch inference on 806 image-text paired comments, outputting probabilities for the three categories and constructing a continuous outcome variable.

$$Y_i = P_i(\text{positive}) - P_i(\text{negative}), \qquad -1 \le Y_i \le 1 \tag{8}$$

Here, Y_i represents the overall text sentiment of comment i, and P_i(positive) and P_i(negative) are the probabilities output by the model; a higher value of Y_i indicates a more positive text sentiment.

On the validation set, the Spearman correlation between Y_i and human-labelled sentiment was 0.718, and the convergence test using platform star ratings as an independent criterion showed consistent results (see Results 3.4 for details); therefore, Y_i is used to characterise the relative emotional position of comments, and its absolute value is not equivalent to a satisfaction rating.

To determine whether image information provides incremental value beyond the text, this study compared six feature combinations with Y_i as the prediction target: Text-only, Image-Smax, Image-Smean, Text + Image-Smax, Text + Image-Smean, and Text + Smax + Smean. Text features consist of normalised sentiment S_{ia}^{norm} across the 13 aspects (with D_{ia} and S_{ia} split into separate encodings for sensitivity analysis); image features include visual salience, as well as comment-level colour richness, visual entropy, and edge density.

Model evaluation consists of two stages: for algorithm selection, under a fixed "Text + Image-Smax" combination, a 70%/30% holdout split was used to evaluate and compare the mean baseline, Random Forest (Breiman 2001), and XGBoost (Chen and Guestrin 2016), based on R^2, RMSE, and MAE on the test set; Feature combination comparisons were conducted using the selected algorithm with 10 repetitions of 5-fold cross-validation under an identical data split.

No hyperparameter tuning was performed for any of the three algorithm classes; instead, a single set of fixed parameters with relatively strong regularisation was used across all feature combinations and cross-validation folds. This ensured that differences between combinations stemmed solely from features rather than parameters, while mitigating the risks of overfitting and selective reporting in small-sample scenarios (806 reviews, 13 to 42 features). The complete set of parameters is provided in Supplementary Table S6.

This study distinguishes between predictive and interpretable feature combinations: the combination with the best cross-validation performance was used to identify the most robust predictive scheme and the incremental value of image features; additionally, an exploratory model was fitted using "Text + Image-Smax" as a predefined interpretable feature set, and SHAP (Lundberg and Lee 2017) was employed to examine the relative contributions of text-based evaluations, image-based saliency, and low-level visual descriptors. This combination was selected not because of its superior predictive performance, but because it addresses the question of cross-modal feature contributions and is consistent with the previous analysis defining visual saliency based on Smax.

3 Results

3.1 Distribution characteristics of perceived attention to landscape elements

Supplementary analysis indicates that comments with images and those without images exhibit significant differences in the structure of landscape element mentions:

comments with images mention topography less frequently (7.4% vs. 11.4%, $q < 0.05$), stilted architecture (4.8% vs. 8.3%, $q < 0.05$), and folk festivals (7.4% vs. 12.6%, $q < 0.01$), and mentioned weather and time of day somewhat more frequently (21.6% vs. 16.9%, uncorrected $p = 0.016$), a difference that did not survive Benjamini–Hochberg correction ($q = 0.051$). These results indicate a certain degree of selectivity in the image-text paired comments; therefore, all conclusions in this paper based on the 806 image-text paired comments primarily apply to the group of tourists who actively uploaded images. After Benjamini–Hochberg FDR correction, the differences in mentions across the remaining aspects were not statistically significant.

Within the perceptual evaluation framework of the Longji Rice Terraces, the distribution of cross-modal attention showed significant differences: tourists' textual expressions primarily focused on the overall scenery, scenic experience, and seasonal phenology, while image records exhibited a more decentralised pattern, covering visible cues such as topography, texture, and weather; the values of living heritage—such as production, architecture, and festivals—occupied a marginal position in both modalities; text and images served complementary rather than redundant perceptual functions (Figure 3).

Figure 3 shows that text and visual attention exhibit different distribution patterns. Textual attention is highly concentrated, with overall scenery (27.9%), scenic experiences (19.0%), seasonal phenomena (13.0%), and weather and time of day (9.8%) collectively accounting for 69.7%, indicating that visitors' verbal expressions primarily revolve around comprehensive scenic experiences, seasonal changes, and viewing conditions—suggesting that these expressions rely heavily on holistic experiences and seasonal characteristics. In contrast, the distribution of visual attention is more balanced, with smaller differences in proportions:with overall scenery (12.3%), topography and landforms (11.8%), texture and contour (11.6%), scenic experiences (10.6%), seasonal phenology (10.1%), and weather and time of day (9.9%) all at relatively high levels. This suggests that, in addition to capturing the overall scenery, visitors' photos precisely preserve concrete spatial elements such as topographic structures and the texture of terraced fields.

The aspect with the greatest disparity in attention between text and images is the overall scenery, where the text share is 15.6 percentage points higher than the visual share (27.9% vs. 12.3%). This reveals that tourists tend to use text for highly generalised evaluations, while the focus of images is diluted by specific landscape elements. Scenic experiences also leaned toward verbal expression, with the text share exceeding the visual share by 8.4 percentage points (19.0% vs. 10.6%). Conversely, topography and texture contours showed a clear preference for visual documentation, with visual attention exceeding text mentions by 9.1 and 8.5 percentage points, respectively; the visual share of cultural connotations was also 6.7 percentage points higher than the textual share (7.9% vs. 1.2%), and that of stilted architecture was 4.4 percentage points higher (6.0% vs. 1.6%). Although these aspects possess high visual prominence and can enter the frame through topographical, textural, architectural, and cultural cues, they are rarely transformed by tourists into active narrative expressions.

The closest discrepancies in text-to-image attention were observed for weather and time of day (9.8% vs. 9.9%) and folk festivals (2.7% vs. 2.5%), with differences of less than 0.3 percentage points in both cases. The difference in text-to-image ratios for viewing trails was 0.8 percentage points, and for farming activities, 1.8 percentage points, indicating that the relative prominence of these aspects is largely consistent across the two modalities. However, although the text-to-image ratios for folk festivals are close, both are relatively low, suggesting that they occupy a generally marginal position in tourists' perceptual structure. The differences for seasonal phenology and viewing platforms are relatively limited, but they show a slight bias toward textual expression and verbal evaluation, respectively.

Overall, textual attention is highly general and emotionally evaluative, focusing primarily on the overall scenery, scenic experiences, and seasonal changes; visual attention, on the other hand, is more dispersed and emphasises the documentation of topography, the texture of terraced fields, architecture, and weather conditions. The overall scenery falls under the core category of "high attention in both text and images, but with stronger linguistic generalisation"; topography, texture and contour, cultural connotation, and stilted architecture exhibit "visual visibility but insufficient linguistic

articulation”; weather and time of day demonstrate high and largely symmetrical attention in both text and images; while folk customs and festivals show a low level of consistency between text and images. Visual visibility does not equate to linguistic articulation, suggesting a modal disconnect in the perception of heritage value.

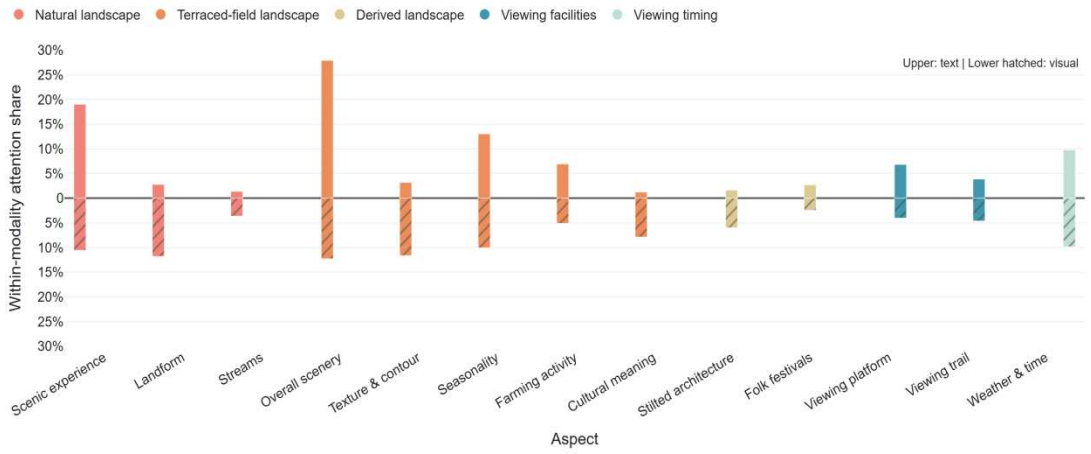


Figure 3. Within-modality allocation of textual and visual attention. Positive bars show text shares and hatched negative bars show visual shares; colours indicate heritage dimensions.

3.2 Consistency between textual and visual perception

Figure 4 shows significant differences in the linguistic manifestation rates and visual grounding rates across various aspects. Overall, the overall scenery, scenic experience, and seasonal phenology exhibit relatively high levels of bidirectional manifestation; topography, texture and contour, and weather and time of day primarily exhibit high visual grounding rates; while folk festivals exhibit high linguistic manifestation rates but low visual grounding rates.

Overall scenery, scenic experiences, and seasonal phenology exhibit high levels of bidirectional manifestation, though the strength of their associations varies. The linguistic manifestation rate for overall scenery is 65.9%, and the visual grounding rate is 98.7%, indicating that when text mentions overall scenery, a prominent depiction can almost always be found in the images, and when images highlight overall scenery, approximately two-thirds of them are also described in the text. However, its Lift is

only 1.00, indicating that this high co-occurrence stems primarily from the widespread presence of overall scenery in both modalities, rather than from a specific text-image association exceeding baseline prevalence. The linguistic manifestation rate for scenic experiences is 47.2%, and the visual grounding rate is 82.9%; however, with a Lift of 0.99, this similarly suggests that the high co-occurrence is primarily explained by its prevalence in both modalities. The linguistic manifestation rate for seasonal phenology was 34.5%, the visual grounding rate was 59.9%, and the Lift was 1.08. Although the bidirectional manifestation levels were lower than those for overall scenery and scenic experience, its co-occurrence was slightly higher than expected by random independence, indicating a weak but relatively balanced correspondence between visual records and linguistic evaluations of seasonal phenology.

Topography, texture and contour, and weather and time of day exhibit an asymmetric structure characterised by “high visual grounding rates and low linguistic manifestation rates.” The visual grounding rates for topography and texture and contours were 100.0% and 97.4%, respectively, but their linguistic manifestation rates were only 7.5% and 9.5%.The Lift for both is close to 1, indicating that whenever the text mentions these aspects, there is almost always a corresponding visual record in the image; however, since the relevant visual aspects are already very common in the vast majority of images, this high visual grounding rate cannot be interpreted as a strong text-image association. On the other hand, although a large number of topographical features and terraced field textures appear in the photographs, they are not actively verbalised by tourists, representing a classic case of “seen but rarely spoken.”The visual grounding rate for weather and time of day was 65.5%, while the linguistic manifestation rate was only 18.3%, resulting in a Lift of 0.85. This indicates that although weather and time of day frequently appear in images, their co-occurrence in text and images is even lower than what would be expected by chance under marginal prevalence, suggesting a significant disconnect between visual documentation and verbal expression.

Although the absolute conditional probabilities for stilted architecture, viewing platforms, and farming activities were not the highest, their Lift was relatively high,

indicating selective over-co-occurrence. Specific elements have a stronger triggering effect on verbal expression, reflecting the high efficiency with which visual representations of specific scenes are transformed into cultural perceptions of the heritage site. The linguistic manifestation rate for stilted architecture is only 8.2%, and the visual grounding rate is 38.5%, but the Lift reaches 1.70. This indicates that although visual cues of the architecture are not frequently included in comments, once stilted architecture are mentioned in text, the probability of a corresponding image being recorded is significantly higher than the random expectation. For viewing platforms, the linguistic manifestation rate and visual grounding rate were 26.3% and 23.0%, respectively, with a Lift of 1.40, indicating a relatively balanced excess of text-image correspondence. For farming activities, the linguistic manifestation rate was 21.9%, the visual grounding rate was only 9.9%, and the Lift was 1.24. This suggests that while farming activities are mentioned in text, they are not necessarily prominently depicted in images; however, the limited co-occurrence of text and images still exceeds the random expectation based on their low marginal frequency.

Folk festivals exhibit a unidirectional manifestation pattern: “visual appearances are easily verbalised, but textual mentions are rarely supported by images.” Their linguistic manifestation rate reached 63.6%, second only to overall scenery, but their visual grounding rate was only 11.7%. Folk festivals also showed a descriptively elevated Lift (8.55), but this estimate is based on only 11 visually highlighted cases and 7 co-occurrences. The 95% bootstrap confidence interval was wide, so this pattern should be treated as a hypothesis-generating signal rather than stable evidence of a general text–image correspondence, and it cannot be directly compared with high-sample categories such as overall scenery. Cultural connotation and viewing trails occupy an intermediate position. The linguistic manifestation rate for cultural connotation was only 3.7%, with a visual grounding rate of 32.0% and a Lift of 1.21; for viewing trails, these figures were 11.1%, 29.9% and 1.16, respectively. This indicates a certain degree of excess co-occurrence for both aspects, but their linguistic manifestation is generally weak and insufficient to form a stable bidirectional correspondence.

Overall, the results confirm that visual presentation is not a sufficient condition for the recognition of heritage significance; the transformation process is subject to multiple constraints, including the recognizability of elements, their event-driven nature, and existing discursive frameworks. The cross-modal perception of visitors to the Longji Rice Terraces exhibits significant “visual-to-linguistic” conversion differences, reflecting variations in the ability of different landscape elements to transform from visual presence into cognitive and meaningful expression. The overall scenery falls within a basic perceptual framework jointly dominated by both modalities; seasonal phenology exhibits a strong capacity for visual-linguistic connection; topography, texture, and cultural cues, despite their high visual visibility, tend to remain at the level of an unarticulated background; whereas festivals, architecture, and viewing facilities primarily trigger the co-occurrence of images and text through specific events or scenes. This suggests that visual presentation is not a sufficient condition for linguistic articulation and the cognitive understanding of heritage meaning; whether visitors can transform the landscapes they see into explicit expressions also depends on the recognizability and event-driven nature of the elements, as well as whether existing heritage interpretations provide a meaningful framework.

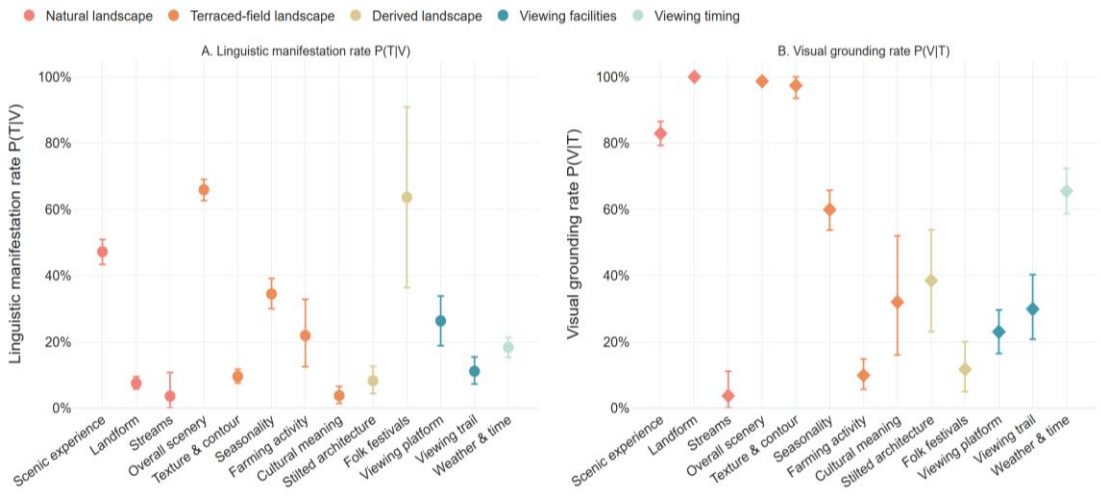


Figure 4. Conditional cross-modal co-occurrence at Smax > 0.7 with review-level bootstrap 95% confidence intervals. Panel A shows the linguistic manifestation rate

P(T|V); Panel B shows the visual grounding rate P(V|T). Absolute probabilities should be read together with Lift and sample support.

3.3 Emotional characteristics in terms of perceptual attention

The overall conclusion is that emotions across aspects are highly concentrated in the positive range, but there is no universal positive correlation between visual salience and aspect evaluations; after conventional OLS testing and Benjamini–Hochberg correction for 13 factors, only the “Overall scenery” aspect retained a weak positive association; the weak negative result for “Seasonal phenology” is not robust to inferential methods (Figure 5).

Figure 5 Panel A shows that the median aspect sentiment scores across the 13 aspects range from 0.982 to 0.994, with the upper quartile approaching 1 for most aspects, while a small number of neutral and negative records still exist. The raincloud plots show that the scores are strongly left-skewed and pile up near the upper bound of 1, that is, they exhibit a ceiling effect. Manual verification further indicates that this concentration of high scores is partly influenced by a positive bias in the sentiment model. Of the 400 candidate records sampled for manual verification, the experts confirmed the assigned aspect in 340, a candidate-level precision of 0.850 (95% bootstrap CI 0.813–0.885); among these 340 confirmed records the three-class sentiment accuracy was 72.1% (95% CI 67.1–76.8), the macro-F1 score was 0.505 (95% CI 0.425–0.573), and the recall rate for the neutral category was only 9.5%. This suggests that subsequent analyses should account for the asymmetry in the model’s classification.

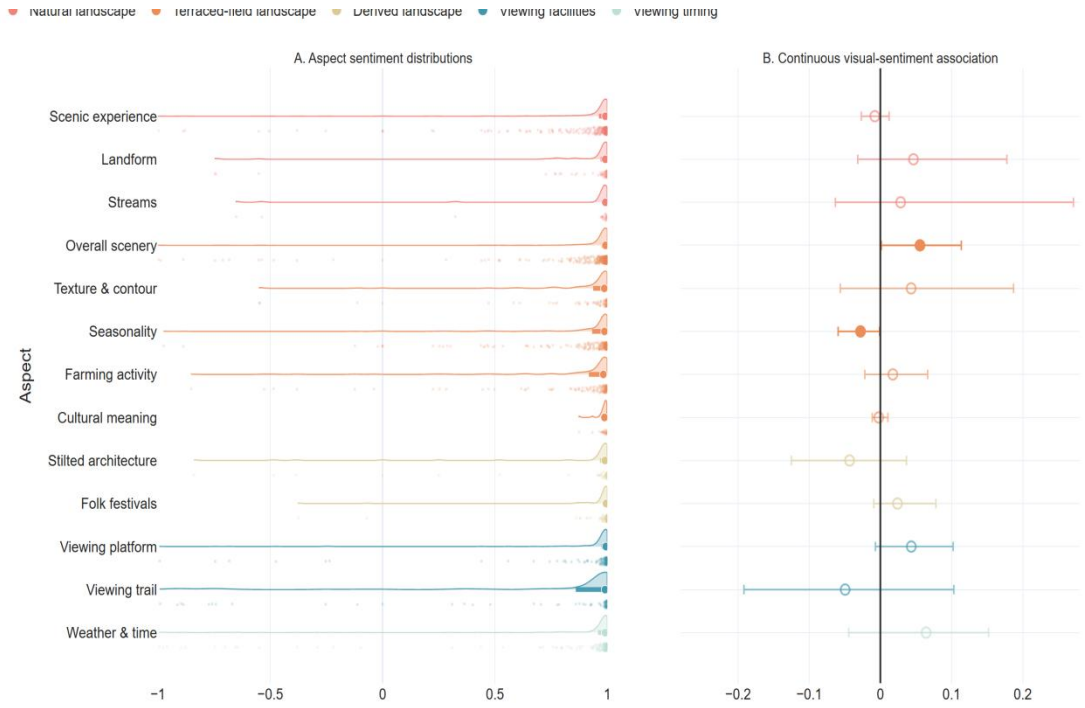


Figure 5. Aspect sentiment distributions and continuous visual-sentiment associations adjusted for log(1+image count). Filled points have intervals excluding zero; open points have intervals crossing zero.

Figure 5 Panel B further examines whether the intensity of visual recordings is correlated with tourists' aspect evaluations. Overall scenery showed a weak positive relationship, while seasonal phenology may exhibit a weak negative relationship because tourists still actively record even when phenological conditions are unfavorable; no stable relationships were found for the remaining 11 aspects.

The continuous visual-sentiment association for overall scenery was weakly positive, with $\beta = 0.056$ and a 95% CI of 0.001–0.114. This implies that for every one standard deviation increase in the average visual salience of the overall scenery, the average emotional rating in the text increases by 0.056. Although the magnitude of this effect is small, it suggests that the overall scenery not only occupies a central position in both textual and visual attention but also that its visual salience is to some extent synchronised with more positive landscape evaluations. The overall scenery may be one of the few aspects capable of simultaneously bridging the three perceptual stages of "being noticed—being photographed—being evaluated."

Seasonal phenology, on the other hand, showed a weak negative correlation, with $\beta = -0.028$ and a 95% CI ranging from −0.059 to −0.001. The more prominent seasonal cues are in an image, the less likely the image is to correspond to a more positive evaluation of phenology. This may be because visitors may actively photograph and comment on the current phenological state even when they have missed the optimal viewing season, or when the rice fields are not yet mature or have already been harvested; therefore, high visual salience may encompass both ideal and less-than-ideal seasonal phases. This result indicates that the "photogenic quality" of seasonal phenology and visitors' "satisfaction" with the seasonal state are distinct attributes, potentially reflecting a decoupling between visual presentation and emotional evaluation.

The 95% Bootstrap confidence intervals for the remaining 11 aspects all span 0; there is currently no evidence that increased visual salience consistently correlates with more positive or negative evaluations of these aspects. For example, although texture and contour command a high share of visual attention, they are rarely mentioned in the text, and their visual salience does not exhibit a consistent emotional association. This suggests that the texture of terraced fields can serve as an important compositional element but is not necessarily translated by visitors into explicit value judgments. Similarly, the act of photographing architecture, facilities, and cultural activities may be influenced by composition, location, the rarity of events, and the convenience of documentation, rather than being driven solely by satisfaction levels.

A dual-expert validation of 100 images revealed significant differences in the reliability of the visual salience model across various aspects (Supplementary Table S7). Experts showed relatively good agreement on scenic experience, overall scenery, texture and contour, and cultural connotations; the model also exhibited relatively strong ranking correlations with expert consensus regarding stilted architecture, textural contours, seasonal phenology, and overall scenery; consistency or model validity was weaker for springs and streams, viewing trails, farming activities, topography and landforms, and weather and time of day. Although folk festivals showed high consistency scores in some areas, estimates were unstable due to sparse

valid samples. Therefore, the visual model effectively supports comparisons in aspects with clear structures and relatively well-defined image boundaries, while exploratory interpretations should be maintained for low-frequency, semantically ambiguous, or context-dependent aspects.

In summary, perceptual attention does not automatically translate into positive evaluations. Overall scenery features are consistently reflected across visual records, linguistic expressions, and positive evaluations; seasonal phenology exhibits “high attention but unstable evaluation trends”; and for most other aspects, no stable correlation has yet been established between visual salience and textual evaluations.

3.4 Overall text sentiment distribution

Overall, tourists’ evaluations of the Longji Rice Terraces tend to reflect “a generally positive emotional bias, but with a small number of concentrated negative experiences.” The model estimates that the overall text sentiment of the 806 image-text pairs is predominantly positive, with a mean of 0.372 and a median of 0.501; positive, neutral, and negative comments account for 65.3%, 25.9%, and 8.8%, respectively (Figure 6). The highest density of sentiment scores fell between 0.5 and 0.8, indicating that most tourists held relatively positive views of their scenic experience at the Longji Rice Terraces; the fact that the median was higher than the mean suggests that a small number of strongly negative comments dragged down the overall average. The distribution extended to a negative long tail approaching −1, further indicating that tourist evaluations were not entirely consistent, and that a small number of tourists still experienced significant dissatisfaction. Both the deviation of the median from the mean and the negative long tail suggest that there may be significant individual differences or specific pain points in landscape perception.

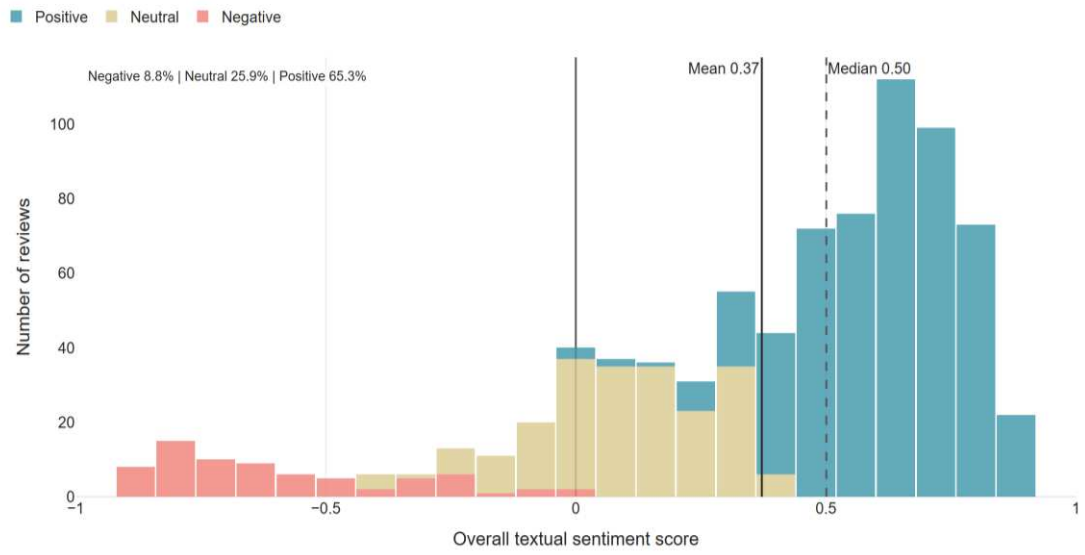


Figure 6. Distribution of overall textual sentiment among 806 image-linked reviews.

Category-specific metrics show: for negative sentiment, precision is 0.389, recall is 0.412, and F1 is 0.400; for neutral sentiment, precision is 0.583, recall is 0.656, and F1 is 0.618; for positive sentiment, precision is 0.891, recall is 0.804, and F1 is 0.845. Misclassifications were concentrated between adjacent categories (9 negative reviews were classified as neutral, and 7 neutral reviews were classified as negative), with a total of 5 misclassifications between positive and negative categories. The weakest category is negative sentiment, for which both precision and recall are close to 0.4, whereas performance for the neutral category is moderate. Because recall is lower than precision for the positive category (0.804 vs. 0.891) and higher than precision for the neutral category (0.656 vs. 0.583), the model, if anything, slightly under-counts positive reviews and over-counts neutral ones, while the small share of negative reviews is estimated with the least certainty; subsequent analyses therefore treat the category distribution with caution.

As a convergence check for the independent target variable, the Spearman correlation between Y_i and the platform star rating across 806 samples was 0.336 ($p < 0.001$); the mean Y_i for 4-star reviews was 0.014, while for 5-star reviews it was 0.453, a statistically significant difference (Mann–Whitney U, $p < 0.001$). However, there is a clear ceiling effect in the star ratings (657 5-star reviews, 149 4-star reviews, and no 3-star or lower reviews); therefore, this test is used only to provide directional support.

Overall, Y_i is consistent with the direction of the independent evaluation metric, but category-specific errors mean that the distribution estimates should be regarded as reference values subject to known errors.

3.5 Predictive association between multimodal features and overall text sentiment

Model evaluation results indicate that, among the candidate algorithms, XGBoost performed best; among the candidate feature combinations, "text-only" matched the best cross-validated performance while using the fewest features. However, to conduct cross-modal SHAP analysis, the "text+image-Smax" combination was selected as the exploratory interpretive combination.

For the selection of candidate algorithms, comparisons were conducted using a fixed "text+image-Smax" feature combination. A 70%/30% holdout test showed that XGBoost's R^2 was 0.134, higher than Random Forest's 0.118 and the mean baseline's value of approximately 0; Its RMSE was 0.383 and its MAE was 0.290, both lower than those of Random Forest and the mean baseline (Figure 7), indicating that, under this holdout split, XGBoost is better suited for handling mixed text and image features than Random Forest. Therefore, XGBoost was selected as the algorithm for subsequent feature combination comparisons and exploratory relationship modelling.

Regarding feature selection, after fixing XGBoost and conducting 10 repetitions of 5-fold cross-validation, the average R^2 for "text-only" and the highest value for "text+image-Smax" were both 0.084 ± 0.046. The "text+image-Smean" combination yielded 0.076 ± 0.043, while the "text+Smax+Smean" combination, which included both Smax and Smean, yielded 0.074 ± 0.044. In contrast, the average R^2 for both "image-only" combinations was below zero (Smax: −0.012 ± 0.049; Smean: −0.024 ± 0.053; Figure 8). This indicates that text features provide the primary predictive information, and the inclusion of image features did not yield a significant increase in predictive power. However, even the best combination achieved an average R^2 of only approximately 0.08, suggesting that the model's overall explanatory power is low. This indicates that overall sentiment is strongly driven by non-landscape factors (such as

transportation and services) as well as individual differences, and that the existing modalities can only capture the landscape-related components of the evaluation. This result corroborates the findings at the aspect level in Section 3.3: since there is generally no stable correlation between the visual salience of individual aspects and their textual evaluations, image features aggregated from these saliencies are unlikely to provide additional explanatory power for overall sentiment. Consistent evidence across both levels indicates that there is a partial decoupling between visual attention and emotional evaluation, rather than a random outcome of a specific model or analytical level.

The mean R^2 values for the "text-only" and "text+image-Smax" models were consistent, but the "text-only" model used fewer features and could not shed light on the roles of image-aspect salience and low-level visual features in cross-modal relationships. Therefore, this study selected the "text+image-Smax" XGBoost model as the exploratory interpretive model and used SHAP to compare the relative contributions of text-aspect evaluations, image-aspect salience, colour richness, visual entropy, and edge density.

Because aspect-level sentiment and overall sentiment were extracted from the same review, the predictive contribution of text features may partly reflect shared method variance and co-sourced measurement. These results should therefore be interpreted as associations and model contributions, not as causal effects of landscape perceptions on overall sentiment.

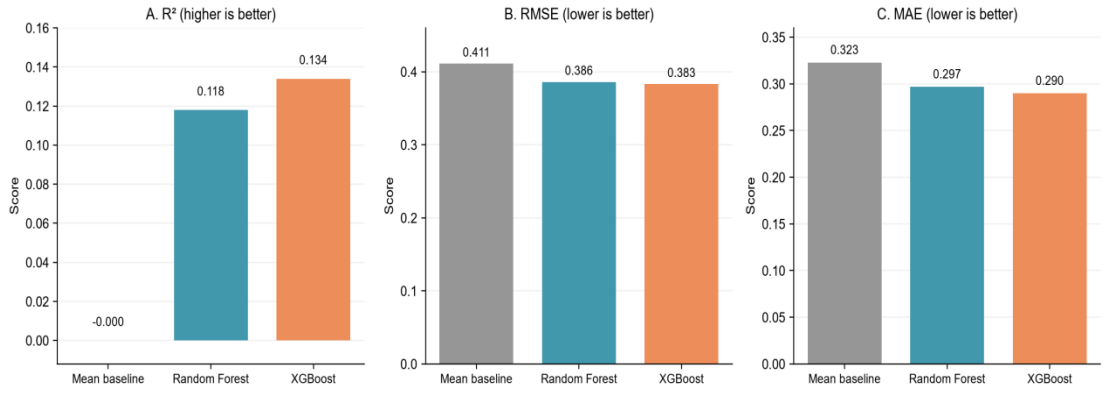


Figure 7. Hold-out comparison of the mean baseline, Random Forest and XGBoost using text+image-Smax, evaluated by R^2, RMSE and MAE.

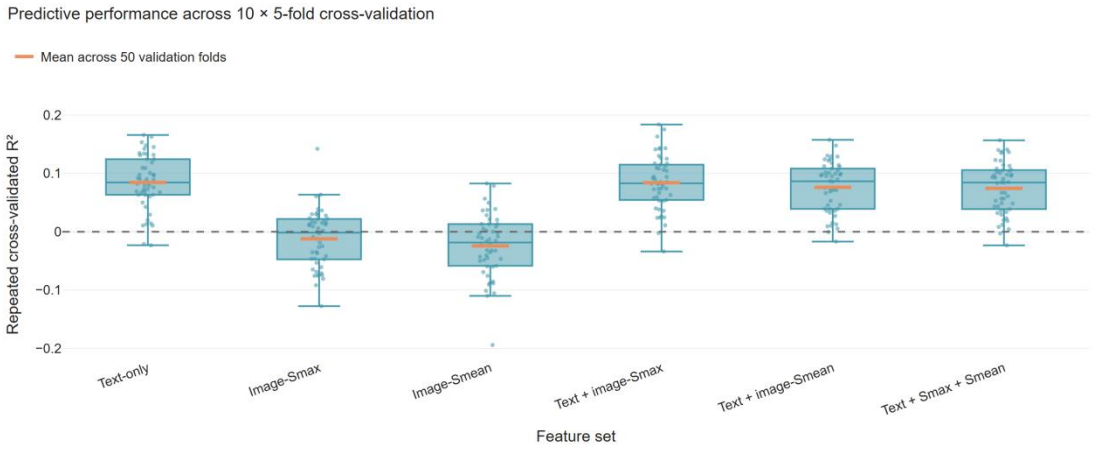


Figure 8. Predictive performance of candidate feature combinations across 10 repetitions of five-fold cross-validation with XGBoost. Orange lines indicate arithmetic means across 50 validation folds.

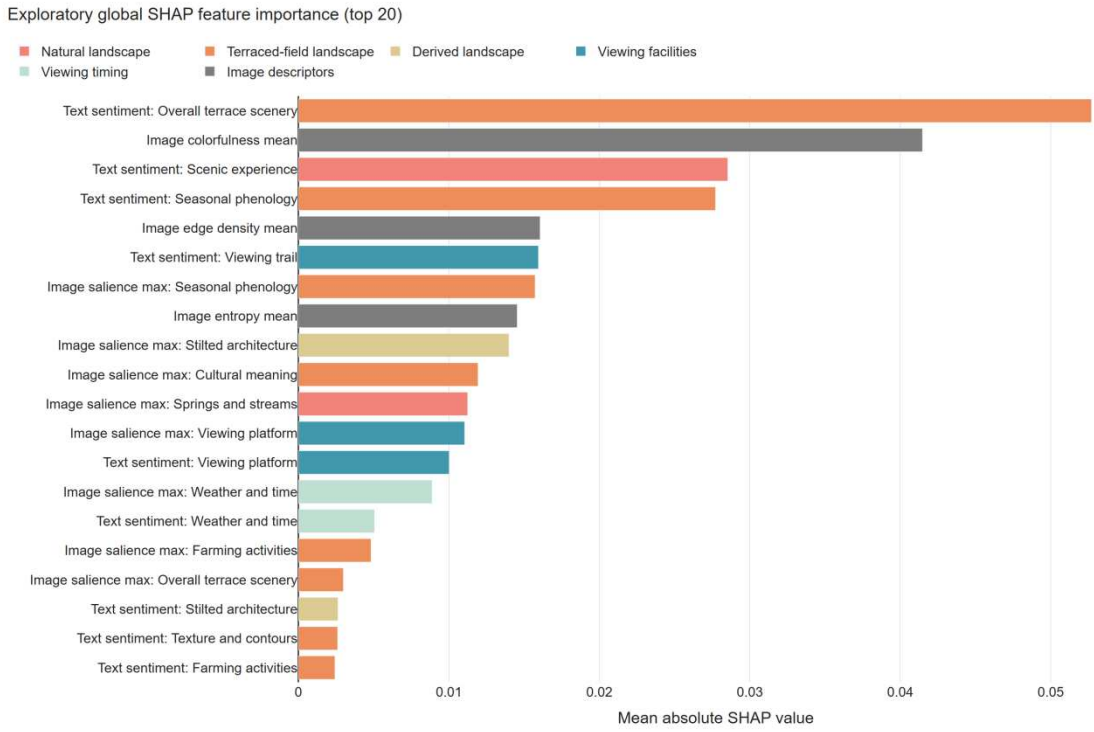


Figure 9. Mean absolute SHAP importance of the 20 highest-ranked features in the exploratory text+image-Smax XGBoost model. SHAP values are derived from an exploratory model and do not indicate incremental predictive validity.

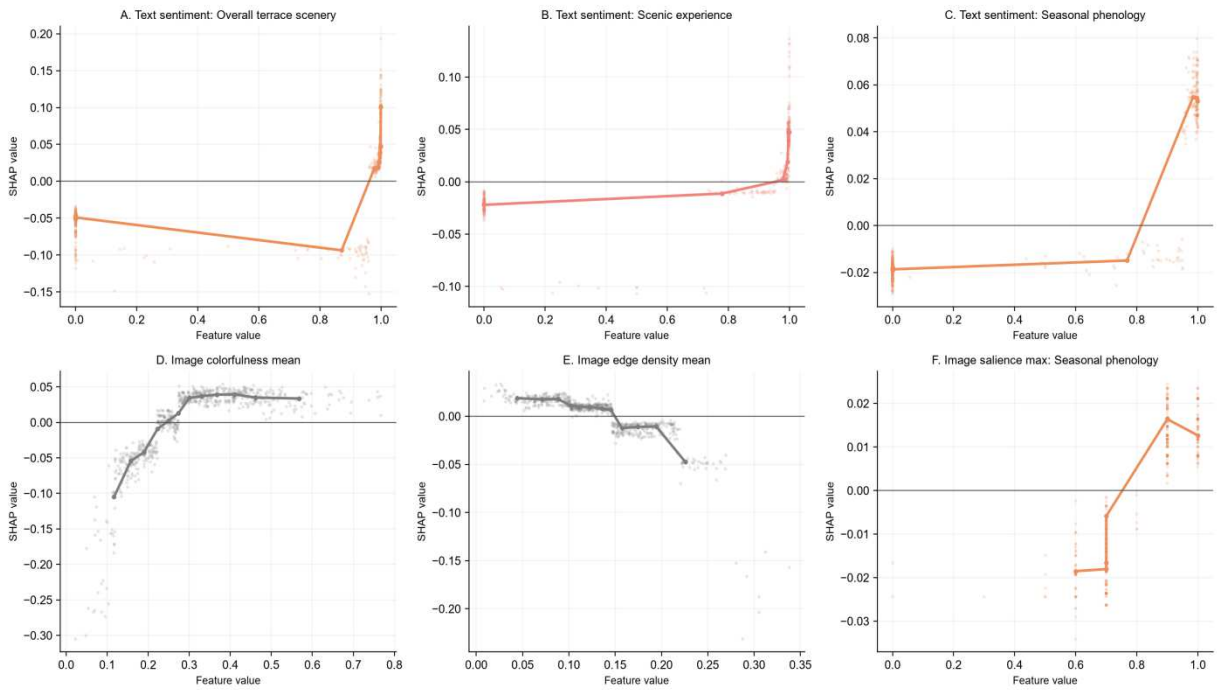


Figure 9 further reveals the hierarchical roles of multimodal features. The top four features collectively account for 48.9% of the total, indicating that the model's predictions are not uniformly derived from all landscape aspects but are concentrated on a few core experiences. Among these, the text sentiment regarding the overall scenery ranks first, suggesting that the overall scenery may serve as a comprehensive evaluation factor linking specific landscape perceptions to the overall experience. It is worth noting that image colour richness ranks second, while the contributions of most visual salience aspects are relatively limited. This suggests that image information is not without value, but its supplementary value is primarily reflected in environmental cues such as colour, structure, and visual atmosphere, rather than semantic information about specific aspects that parallels textual evaluations.

Figure 10. Exploratory SHAP contribution curves for selected predictors. Panels A–C show textual aspect evaluations, D–E low-level visual descriptors, and F aspect-level visual salience. Points are observations and lines are binned medians.

Figure 10, Panels A–C, show high-value thresholds for core landscape evaluations: Overall Scenery, Scenic Experience, and Seasonal phenology only rapidly shift to significantly positive SHAP when emotional ratings approach the highest range, with

Overall Scenery exhibiting the greatest amplitude. Since aspect-level emotions and overall emotions are derived from the same review, this relationship may partly stem from co-sourced textual measurements.

Figure 10, Panels D–E, show the nonlinearity of lower-level visual features: colour richness primarily manifests as a penalty at low values, with its positive contribution tending to plateau after reaching a moderate level; higher edge density is accompanied by negative SHAP. Since edge density cannot distinguish between terraced field textures, building boundaries, and photographic noise, it can only serve as a clue requiring further validation.

Figure 10 Panel F shows that seasonal phenology (Smax) makes a positive contribution in the higher range, but its global importance is lower than that of seasonal phenology text sentiment. The linguistic evaluation of seasonal landscapes is closer to the overall text sentiment than the criterion of "at least one image prominently capturing it"; simultaneously, Smax is subject to a chance effect related to the number of images, so this result remains exploratory.

Combining Figure 10 Panels A–F, SHAP reveals three types of internal model structures: high thresholds for core text evaluations, low colour penalties and saturation, and negative contributions from high edge density. Overall, linguistic evaluations align more closely with the underlying overall text sentiment than visual saliency; images merely supplement visual environmental cues. This is consistent with the finding from repeated cross-validation that no stable incremental value from images was detected. These results corroborate the findings in Section 3.3 regarding the aspect level: the visual salience at the aspect level generally does not consistently align with text evaluations, which serves as the micro-level basis for why image features at the holistic level struggle to provide predictive increment. Together, these findings indicate that images capture objects of attention, while text conveys evaluative judgments; in this case, the two operate along distinct pathways.

4 Discussion

This study indicates that tourists' perceptions of the Longji Rice Terraces do not follow a linear path from "landscape exposure" to "understanding of cultural connotations" and then to "formation of emotional evaluations." Image recording, linguistic expression, and evaluation formation reflect visual selection, meaning recognition, and attitude construction, respectively; while the three are interrelated, they are not interchangeable. As an agricultural cultural heritage site, the Longji Rice Terraces are most readily viewed and disseminated through their overall scenery, the texture of the terraces, and seasonal colours, whereas the farming knowledge, community collaboration, and cultural connotation that form the core of this living heritage are rarely reflected in tourists' explicit expressions. This discrepancy suggests that tourism communication has expanded the landscape visibility of agricultural cultural heritage, but increased visibility does not necessarily imply that heritage value is simultaneously transformed into an understanding of that value.

4.1 From landscape visibility to heritage recognition: perceptual focus and cross-modal discrepancy

Visitors' attention is focused on the overall scenery, scenic experiences, and seasonal phenomena, consistent with the selective viewing emphasised by the "tourist gaze" (Urry and Larsen 2011). The scale, curves, and seasonal colours of the terraces can be quickly recognised and are well-suited for display and dissemination through photographs; in contrast, irrigation systems, farming knowledge, and community collaboration are difficult to convey directly through a single image. Consequently, as a living heritage shaped by the natural environment, agricultural production, and local culture, the Longji Rice Terraces may be reduced in tourist narratives to a few highly visually recognizable scenic symbols. The apparent absence of cultural connotation does not stem from visitors' subjective neglect, but rather from the lack of intuitive

visual carriers for deeper cultural connotations—which require knowledge, narratives, or participatory experiences as intermediaries.

The disconnect between images and text further reveals this process of transformation. The texture of the terraced fields often constitutes the primary visual structure of a photograph, yet it does not necessarily become the subject that tourists name or evaluate on its own; conversely, elements with clear forms or functions—such as stilted architecture and viewing platforms—are more likely to trigger corresponding verbal expressions once they appear in the frame. Therefore, whether a landscape enters the text depends not only on its visual prominence but also on whether it possesses a recognizable name, function, and cultural associations. Abstract values such as cultural connotation and ecological wisdom lack a single, stable visual carrier and rely more heavily on visitors' prior knowledge or on-site interpretations to be articulated in language. This implies that "being photographed" primarily indicates that the landscape has entered a visual record, whereas "being spoken about" indicates that it has, to some extent, entered explicit cognition.

This distinction also helps us reinterpret multimodal data. Images and text are not two parallel channels of measurement used to verify one another: images primarily capture form, colour, and environmental conditions, while text serves the functions of classification, interpretation, and evaluation. Even if a particular landscape appears frequently in both modalities, a stable cross-modal connection does not necessarily exist; conversely, low-frequency elements may form strong image-text correspondences due to their high recognizability. Therefore, judging visitor interest solely based on photography frequency or absolute co-occurrence counts may overestimate the consistency of perceptions regarding agricultural cultural heritage.

Visual salience does not automatically translate into positive evaluations. Seasonal landscapes, in particular, illustrate that when visitors photograph a subject, they may do so out of appreciation, for the purpose of recording information, or because of unmet expectations or dissatisfaction. For example, missing the optimal viewing period may still prompt visitors to document rice paddies that are not yet ripe or have already been

harvested. Photographs depict the objects of attention, not the emotional orientation behind that attention.

Two results therefore need to be read together. Image features added little incremental value for predicting overall sentiment beyond text (Section 3.5), and aspect-level visual salience showed no stable positive association with the corresponding aspect sentiment (Section 3.3). Neither result means that images are irrelevant to heritage perception; rather, images and text perform different functions. Images record what visitors attended to and under what visual conditions—form, colour, season and weather—and thus reveal the objects and structure of attention. Text names, interprets and evaluates what was seen, and therefore carries the judgements that shape overall sentiment. In this sense, images are better suited to answering "what visitors saw", whereas text more closely reflects "how visitors interpret and evaluate what they saw". The SHAP results should be read in the same way: SHAP values describe the relative contribution of each feature to the predictions of one already-specified exploratory model, and do not show that adding that feature significantly improves the model's predictive performance. The high SHAP ranking of colour richness, for example, means that the fitted model relied on this feature, not that image features provide incremental predictive validity, which repeated cross-validation did not detect. The value of multimodal research should therefore not be measured solely by predictive accuracy; it lies more in revealing the selectivity, omissions and reconstruction of meaning that occur as heritage value is translated across media.

4.2 From displaying landscapes to interpreting living systems: implications for heritage management

The above findings imply that the focus of agricultural cultural heritage interpretation should not merely be on increasing the amount of information presented, but rather on helping visitors connect visible forms with the socio-ecological processes that shaped them; the construction of interpretive systems should shift from a simple accumulation of information to the establishment of logical connections. The curves of

terraced fields are not merely aesthetic compositions; they are also the result of adaptation to the terrain, the organisation of water systems, the maintenance of field ridges, and long-term collective labour. Seasonal colours are not merely a spectacle to be admired; they are linked to crop cycles, agricultural production, and community life. Interpretive design can use the overall scenery and seasonal landscapes—which already capture visitors' attention—as entry points, transforming them from "photogenic scenes" into clues for understanding living agricultural systems.

Viewing platforms and trails can serve as key nodes for this shift in meaning, but they should not be limited to providing optimal photo spots. Brief narratives corresponding to the direction of the view can connect "what is seen before one's eyes—the formation process—and cultural connotation."For content that is less visually or textually evident—such as cultural connotations, farming activities, and folk festivals—isolated interpretive signs may have limited effectiveness; instead, such elements are better integrated into the visitor experience through community-led tours, agricultural hands-on activities, oral histories, and festival calendars. Relevant content should be selected and expressed with the participation of local communities, ensuring that tourism interpretation connects to real production and daily life, rather than transforming living traditions once again into performative symbols (MacCannell 1973). Management effectiveness should not be judged solely by the number of photos or online mentions; it should also examine whether visitors have developed an understanding of landscape processes, local knowledge, and the role of the community.

4.3 Limitations and future research

This study still has several limitations. First, photo-accompanied comments represent visitors who are more inclined toward visual documentation and online sharing and cannot be equated with all visitors; the prominence of photos also cannot substitute for on-site visual attention. Second, since aspect-specific emotions and overall emotions are derived from the same comment, the predictive contribution of text features may be influenced by co-source measurement, a positive bias in emotional

distribution, and non-landscape-related evaluations such as transportation, services, and ticket prices. Third, there are differences in the identification validity across visual aspects, and the manually verified sample for low-frequency scenes remains limited. Finally, cross-sectional data, Bootstrap analysis, and SHAP can only reveal associations and model contributions; they cannot prove the causal process from visual exposure to cultural understanding or emotional formation.

Future research could combine online text-and-image data with physiological measurements (such as eye-tracking experiments) and qualitative interviews to directly test and deconstruct the dynamic mechanisms underlying the transition from visual exposure to cognitive processing. At the same time, introducing independent outcome variables—such as heritage knowledge, interpretive effectiveness, and willingness to revisit—could mitigate the impact of co-variance. Comparisons across seasons, scenic areas, and agricultural cultural heritage cases could further determine whether visual focus and the weak manifestation of cultural connotation represent a phenomenon specific to the Longji Rice Terraces or a more universal perceptual mechanism in agricultural heritage tourism. Overall, the most significant value of multimodal user-generated content lies not in providing additional predictive variables, but in enabling researchers to identify how heritage values are respectively highlighted, transformed, or omitted during the processes of visual communication, linguistic expression, and evaluation formation.

References

Bai S, Chen K, Liu X et al (2025) Qwen2.5-VL technical report. https://arxiv.org/abs/2502.13923. Accessed August 25, 2026

Bao Q, Zhong H, Tan Y, Wang Y (2021) Characteristics of the cultural landscape genes of rice-terrace agricultural heritage sites in southern China: a case study of the Longji Terraces, Longsheng, Guangxi. Soc Sci 74–79

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Stat Methodol 57:289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324

Campello RJGB, Moulavi D, Sander J (2013) Density-based clustering based on hierarchical density estimates. In: Pei J, Tseng VS, Cao L et al (eds) Advances in Knowledge Discovery and Data Mining. Lecture Notes in Computer Science, vol. 7819. Springer, Berlin, pp. 160–172. https://doi.org/10.1007/978-3-642-37456-2_14

Cao H, Wu Z, Zheng W (2024) Impact of touristification and landscape pattern on habitat quality in the Longji Rice Terrace ecosystem, southern China, based on geographically weighted regression models. Ecol Indic 166:112259. https://doi.org/10.1016/j.ecolind.2024.112259

Cheng X, Van Damme S, Li L, Uyttenhove P (2019) Evaluation of cultural ecosystem services: a review of methods. Ecosyst Serv 37:100925. https://doi.org/10.1016/j.ecoser.2019.100925

Chen Q, Liu R, Jiang Q, Xu S (2024) Exploring cross-cultural disparities in tourists' perceived images: a text mining and sentiment analysis study using LDA and BERT-BILSTM models. Data Technol Appl 58:669–690. https://doi.org/10.1108/DTA-10-2023-0645

Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, August 13–17, 2016. ACM, New York, pp. 785–794. https://doi.org/10.1145/2939672.2939785

Chen W, Zhou K, Hu B et al (2025) Unlocking visitor experiences in cultural heritage sites with SHAP-interpretable AI and social media sentiment analysis. npj Herit Sci 13:457. https://doi.org/10.1038/s40494-025-02014-0

Chhabra D, Healy R, Sills E (2003) Staged authenticity and heritage tourism. Ann Tour Res 30:702–719. https://doi.org/10.1016/S0160-7383(03)00044-6

Egarter Vigl L, Marsoner T, Giombini V et al (2021) Harnessing artificial intelligence technology and social media data to support cultural ecosystem service assessments. People Nat 3:673–685. https://doi.org/10.1002/pan3.10199

FAO (2023) Guidelines for preparing a GIAHS proposal document. Food and Agriculture Organization of the United Nations, Rome

Fu G (2017) Ethnic habitus, mountain environment and the culture of the Longji Terraces. Guangxi Minzu Yanjiu 85–90

Fuller AM, Min Q, Jiao W, Bai Y (2015) Globally Important Agricultural Heritage Systems (GIAHS) of China: the challenge of complexity in research. Ecosyst Health Sustain 1:1–10. https://doi.org/10.1890/EHS14-0007.1

Gardiner S, Vada S, Yang ECL et al (2022) Recreating history: the evolving negotiation of staged authenticity in tourism experiences. Tour Manag 91:104515. https://doi.org/10.1016/j.tourman.2022.104515

Gosal AS, Ziv G (2020) Landscape aesthetics: spatial modelling and mapping using social media images and machine learning. Ecol Indic 117:106638. https://doi.org/10.1016/j.ecolind.2020.106638

Grootendorst M (2022) BERTopic: neural topic modelling with a class-based TF-IDF procedure. https://arxiv.org/abs/2203.05794. Accessed August 25, 2026

Hua H, Zhou S, Ding Z, Pan Y (2018) The change mechanism of human-environment interactions from the perspective of contextualisation: a case study of the Honghe Hani Rice Terraces as a World Cultural Heritage site. Sustainability 10:2230. https://doi.org/10.3390/su10072230

Jia M, Feng J, Chen Y, Zhao C (2024) Visual analysis of social media data on experiences at a World Heritage tourist destination: Historic Center of Macau. Buildings 14:2188. https://doi.org/10.3390/buildings14072188

Kim H, Jamal T (2007) Touristic quest for existential authenticity. Ann Tour Res 34:181–201. https://doi.org/10.1016/j.annals.2006.07.009

Lin M, Wang M, Lu Y et al (2025) Synergising conservation and tourism utilisation in agricultural heritage sites: a comparative analysis of economic resilience in Wujiang and Longsheng, China. Land 14:796. https://doi.org/10.3390/land14040796

Lin T, Li X, Zhang X et al (2026) A multimodal study of decoding heritage tourists' perception and emotion in historic urban areas of Ganzhou. npj Herit Sci 14:2. https://doi.org/10.1038/s40494-025-02242-4

Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. In: Guyon I, Luxburg U, Bengio S et al (eds) Advances in Neural Information Processing Systems 30. Curran Associates, Red Hook, pp. 4765–4774

Lu Y, Tang X, Min Q, Shao J (2017) The Longji Terraces system of Longsheng, Guangxi. China Agriculture Press, Beijing

MacCannell D (1973) Staged authenticity: arrangements of social space in tourist settings. Am J Sociol 79:589–603. https://doi.org/10.1086/225585

Ma R, Furuya K (2024) Social media images and the application of computer vision methods in landscape studies: a systematic literature review. Land 13:181. https://doi.org/10.3390/land13020181

McGraw KO, Wong SP (1996) Forming inferences about some intraclass correlation coefficients. Psychol Methods 1:30–46. https://doi.org/10.1037/1082-989X.1.1.30

McInnes L, Healy J, Melville J (2018) UMAP: uniform manifold approximation and projection for dimension reduction. https://arxiv.org/abs/1802.03426. Accessed August 25, 2026

Min Q (2021) Agri-cultural heritage: an interdisciplinary field with development prospects. J Resour Ecol 12:437–443. https://doi.org/10.5814/j.issn.1674-764x.2021.04.001

Moore K, Buchmann A, Månsson M, Fisher D (2021) Authenticity in tourism theory and experience. Practically indispensable and theoretically mischievous? Ann Tour Res 89:103208. https://doi.org/10.1016/j.annals.2021.103208

Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830

Reimers N, Gurevych I (2020) Making monolingual sentence embeddings multilingual using knowledge distillation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Stroudsburg, pp. 4512–4525. https://doi.org/10.18653/v1/2020.emnlp-main.365

Ren W, Li Q, Zhan K, Fu W (2024) Research on the landscape perception of the Quanzhou World Cultural Heritage site in Fujian based on machine learning. Landsc Archit 31:120–129

Rössler M, Lin RC-H (2018) Cultural landscape in World Heritage conservation and cultural landscape conservation challenges in Asia. Built Herit 2:3–26. https://doi.org/10.1186/BF03545707

Santoro A, Venturi M, Agnoletti M (2020) Agricultural heritage systems and landscape perception among tourists: The case of Lamole, Chianti (Italy). Sustainability 12:3509. https://doi.org/10.3390/su12093509

Shen F, Li Y (2012) Rice-farming culture and the landscape ecology of terraced fields: a case study of the Longji Terraces, Guangxi. Guangxi Minzu Yanjiu 128–133

Sun Y, Timothy DJ, Wang Y et al (2019) Reflections on agricultural heritage systems and tourism in China. J China Tour Res 15:359–378. https://doi.org/10.1080/19388160.2019.1578713

Tian M, Min Q, Jiao W et al (2016) Agricultural heritage systems tourism: definition, characteristics, and development framework. J Mt Sci 13:440–454. https://doi.org/10.1007/s11629-015-3724-2

Tieskens KF, Van Zanten BT, Schulp CJE, Verburg PH (2018) Aesthetic appreciation of the cultural landscape through social media: an analysis of revealed preference in the Dutch river landscape. Landsc Urban Plan 177:128–137. https://doi.org/10.1016/j.landurbplan.2018.05.002

Urry J, Larsen J (2011) The Tourist Gaze 3.0, 3rd ed. Sage, London

Wang X, Mou N, Zhu S et al (2024) How to perceive tourism destination image? A visual content analysis based on inbound tourists' photos. J Destin Mark Manag 33:100923. https://doi.org/10.1016/j.jdmm.2024.100923

Wu H-C, Li T (2014) A study of experiential quality, perceived value, heritage image, experiential satisfaction and behavioural intentions for heritage tourists. J Hosp Tour Res 42:26–73. https://doi.org/10.1177/1096348014525638

Yang X, Zhao Y, Zhao J et al (2022) Tourists' perceived attitudes toward the famous terraced agricultural cultural heritage landscape in China. Agriculture 12:1394. https://doi.org/10.3390/agriculture12091394

Zhang J, Gan R, Wang J et al (2022) Fengshenbang 1.0: the foundation of Chinese cognitive intelligence. https://arxiv.org/abs/2209.02970. Accessed August 25, 2026

Zhang J, Luo Y, Cao H, Su S (2024) Scrutinising the cultural ecosystem services of Chinese classical gardens: a novel deep learning approach based on online reviews from a multisensory perspective. Appl Geogr 172:103404. https://doi.org/10.1016/j.apgeog.2024.103404

Zhang J, Xu C (2024) A grounded theory-based model of information communication in cultural heritage digital reading on social media from the perspective of

embodied theory. Herit Sci 12:208. https://doi.org/10.1186/s40494-024-01313-2

Zhang K, Chen Y, Li C (2019) Discovering tourists' behaviours and perceptions in a tourism destination by analysing the visual content of photos using a computer deep learning model: the case of Beijing. Tour Manag 75:595–608. https://doi.org/10.1016/j.tourman.2019.07.002

Zheng X, Huang Y, Xie Z, Zheng A (2025) Quantifying tourist perception of cultural landscapes in traditional towns in China using multimodal machine learning. npj Herit Sci 13:320. https://doi.org/10.1038/s40494-025-01907-4

Zhu G, Li X, Zhang Y et al (2021) A multi-stakeholder involvement mechanism in tourism management for preserving terraced landscapes at Important Agricultural Heritage Systems (IAHS) sites: a case study of Dazhai Village in the Longji Rice Terraces, China. Land 10:1146. https://doi.org/10.3390/land10111146

Zube EH, Sell JL, Taylor JG (1982) Landscape perception: research, application, and theory. Landsc Plan 9:1–33. https://doi.org/10.1016/0304-3924(82)90009-0

Use of generative AI

Generative AI was used in two ways while this work was prepared. First, Claude (Anthropic; Claude Opus 5), accessed between August and September 2026, assisted the English translation and language editing of the authors' Chinese manuscript. Second, the same tool assisted in drafting and checking parts of the analysis code. The authors reviewed, corrected and verified all output, including the reported text, references, equations, declarations and analytical results, and take full responsibility for the submitted content. No generative-AI output was treated as an author or as evidence, and no research datum, result or conclusion reported in this article was generated by generative AI. The vision-language model used as a research instrument,

Qwen2.5-VL-7B-Instruct, is a separate matter and is described in Methods and in Supplementary Note 1.

Supplementary information

Supplementary Tables S1–S7, Supplementary Figures S1–S2 and Supplementary Notes 1–3 are provided in a separate file.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 04SupplementaryInformation.docx