Journal of Autonomous Intelligence (2024) Volume 7 Issue 4
doi: 10.32629/jai.v7i4.1471

ORIGINAL RESEARCH ARTICLE

Exploration of the cultural attributes of Chinese character sculpture using machine learning technology

Zhen Luo

International College, Krirk University, Bang Khen District, Bangkok 10220, Thailand; luozhen7979@163.com

ABSTRACT

The article employs machine learning, specifically the CLIP (Contrastive Language-Image Pretraining) model, to analyze Chinese character sculptures' cultural attributes. It overcomes challenges in multi-dimensional data processing and high digitization costs. The process involves normalizing sculpture images, using FastText for vector representations of Chinese characters, and mapping text to the same embedding space as images for word embedding. The CLIP model, through unsupervised training, minimizes the negative logarithmic likelihood loss between image and text embeddings to establish cultural attribute representations. Key findings include the CLIP model's improved performance over the M3 model, with a 5.4% higher average AUC. The model demonstrates high efficiency and accuracy, evident in its low *RMSE* (0.034) and *MAE* (0.025) and fast analysis time of 182 ms. This approach effectively and accurately analyzes the cultural attributes of Chinese character sculptures, addressing existing research gaps.

Keywords: Chinese character sculpture; cultural attribute analysis; machine learning; CLIP model; unsupervised training

ARTICLE INFO

Received: 28 November 2023
Accepted: 9 January 2024
Available online: 5 February 2024

1. Introduction

Chinese character sculpture is a creative form that integrates the shape and structure of Chinese characters into sculpture art. By carving, shaping, and other handicraft techniques, the strokes, lines, and shapes of Chinese characters are given three-dimensional expression, conveying the artistic beauty and cultural connotations of the text. However, there are still some problems in the current research on sculpture. When analyzing its cultural attributes, there are challenges such as strong subjectivity[1,2], low data processing efficiency[3,4], and high digitization costs[5,6]. By introducing machine learning technology, it can effectively improve the objectivity and analysis efficiency of the cultural attributes of Chinese character sculpture, and more comprehensively understand and present this unique art form. The inspiration for this study stems from the intricate blend of linguistic art and sculptural expression in Chinese character sculpture, an artistic form that embodies deep cultural significance. This research is motivated by the need to unravel and digitally encapsulate the profound cultural essence embedded in these sculptures, a domain that remains insufficiently explored and understood in current scholarly discourse. Our investigation confronts several formidable challenges, notably the strong subjectivity inherent in cultural interpretations, the low efficiency in processing complex data sets, and the considerable costs associated with digital conversion of these intricate artworks. Addressing these challenges is critical to developing a nuanced understanding of the cultural

attributes of Chinese character sculpture.

This study introduced the CLIP model in machine learning technology to achieve a comprehensive analysis of the cultural attributes of Chinese character sculptures. Chinese character sculpture images from different regions, historical periods, and artistic genres were normalized, including steps such as size adjustment, grayscale standardization, and brightness equalization. The feature points of the sculpture image were matched to a three-dimensional space for triangulation, generating a three-dimensional point cloud, and reconstructing the surface shape and three-dimensional structure of the Chinese character sculpture. Chinese character sculpture text data was collected, and the FastText model was utilized to semantically correspond text data with images. The semantic correspondence between text and images was established, and the weight matrix was adjusted. The text processed by word embedding was mapped to the space of image embedding and effectively compared in a common semantic space. In the unsupervised training phase of the CLIP model, comparative learning methods were used to minimize the negative logarithmic likelihood loss between image and text embeddings and establish cultural attribute representations. By using the gradient descent method, the negative logarithmic likelihood loss of comparative learning was minimized. By using gradient weighted mapping method, the model's attention to Chinese character sculpture images was visualized, presenting the model's decision-making process in an interpretable manner. The K-means clustering algorithm was used to reveal the internal structure and correlation of cultural attributes of Chinese character sculpture. The clustering center was iteratively optimized to more accurately depict the group structure of cultural attributes of Chinese character sculptures. In the experiment, a leave-one-out cross validation was used to partition the dataset, and precision, recall, *F1* value, accuracy, *AUC* value, *RMSE*, *MAE*, and analysis time were used as evaluation indicators. The results showed that the CLIP model had a calculated mean of 0.98 for precision, recall, *F1* value, and accuracy in the analysis of Chinese character sculptures with 10 different cultural attributes, which was significantly better than models such as M3, Dual-Attention Network, Image-Text Embedding Model, and ViLBERT (Vision-and-Language Bidirectional Encoder Representations from Transformers). The average *AUC* of CLIP model for classifying Chinese character sculptures under 10 different cultural attributes was 0.98, which was about 5.4% higher than the M3 model. The model used achieved the lowest *RMSE* (0.034) and *MAE* (0.025) while also achieving the fastest analysis speed of only 182 milliseconds. This study's principal contribution lies in harnessing advanced machine learning technology, specifically the CLIP model, to significantly enhance the objectivity and analytical efficiency in exploring the cultural attributes of Chinese character sculpture. Through our innovative approach, we offer a more comprehensive and nuanced understanding of this unique art form, thereby filling a critical gap in existing research.

2. Related work

The in-depth research on the attributes of sculpture is a key topic at present. The existing research on sculpture mainly focuses on aspects such as history[7,8], art[9,10], and linguistics[11,12]. From the perspective of philosophical argument in art, Scott[13] discussed various methods for constructing marble sculptures. Distinguishing art based on the spatial and temporal arrangement of semiotic elements, Fraser[14] traced his research on sculpture back to the late 19th century in artistic writing. After conducting continuous research on black sculptures in Brazil, Barata[15] raised the issue of artistic development related to the changes in African Brazilian religion. Facing the imaginative, creative, and technical challenges involved in contemporary sculpture, Mihai Ionut[16] explored the application of new media and modern photography in contemporary sculpture art works. In the process of studying the history of sculpture, scholars such as Heginbotham et al.[17] successfully used Energy Dispersive X-Ray Florescence (ED-XRF) spectroscopy to measure the production date of the sculpture. In order to protect the history and tradition of folk applied art, Jabbarov[18] conducted a detailed analysis of the types, forms, and origins of patterns in Uzbekistan's sculpture art. Although relevant research has been conducted on the historical evolution of sculpture in the past, it has often overlooked the in-

depth exploration of its cultural attributes[19,20]. Due to limitations in data processing and method selection, there are still shortcomings in the in-depth analysis of sculpture cultural attributes.

In recent years, multimodal learning methods in machine learning have provided an innovative research approach for deeply understanding the artistic features and cultural connotations of sculpture. CLIP is a multimodal learning model developed by OpenAI[21,22], which combines training from image and text data to enable the model to understand the relationships between images and text, and achieve cross modal tasks[23,24]. The main idea is to learn by comparing images and text, map relevant content to a common embedding space[25], and achieve semantic alignment of images and text without task specific labels[26,27]. Previous studies have shown that using the CLIP model can better capture the semantic relationships between images and text, and analyze object attributes more comprehensively[28,29]. Utilizing the semantic capabilities of CLIP models in large-scale data scenarios, scholars such as Gal et al.[30] proposed a new text driven approach that allows for the transfer of generated models to new domains without the need to collect individual images. Using CLIP technology, scholars including Huang[31] developed a multimodal artificial intelligence based on Pathology Language-Image Pretraining (PLIP) for processing publicly annotated medical images and achieving comprehensive understanding of images and texts. To improve the efficiency of local text driven editing tasks for general images, scholars such as Avrahami[32] adopted the Latent Diffusion Model (LDM). This model run in a low dimensional latent space, achieving accelerated diffusion by eliminating the need for resource intensive CLIP gradient calculation in each diffusion step. Inspired by the latest progress in natural language processing (NLP) real-time learning research, scholars including Zhou[33] proposed Context Optimization (CoOp), a method specifically designed to adjust CLIP models for downstream image recognition. Facing the problems of long training time, high storage requirements, and identity loss in existing personalized methods, Gal et al.[34] adjusted the encoder domain based on the CLIP model to achieve rapid personalization from text to image models. Based on the CLIP model idea, scholars such as Zheng et al.[35] embedded images and text into a shared visual text space through instance loss, and treated each image/text group as a category using unsupervised assumptions. By utilizing the CLIP model to combine the advantages of language and images, it is possible to analyze the cultural attributes of Chinese character sculptures more comprehensively and objectively, making up for the shortcomings of existing research.

3. Chinese character sculpture image and text data processing

By extensively searching the databases of well-known digital art museums, online platforms of cultural institutions, and related art research institutions, Chinese character sculpture images from different regions, historical periods, and art genres are obtained to ensure cultural diversity of the data. Through the digital platform of the museum, Chinese character sculpture images covering multiple cultural aspects such as historical tradition and contemporary innovation are obtained. **Table 1** shows the number of Chinese character sculpture images obtained from different regions, historical periods, artistic genres, and cultural levels in the study.

Table 1. Statistics on the quantity of obtained Chinese character sculpture images of various types.

Region	Historical period	Art movement	Cultural aspect	Number of Chinese character sculptures
China	Ancient Times	Traditional Culture Revival	Historical Tradition	302
China	Modern Times	Abstract Art	Contemporary Innovation	204
Taiwan	Contemporary Times	Contemporary Art	Cultural Diversity	153
Hong Kong	End of 20th Century	Experimental Art	Urban Culture	128
Japan	Middle Ages	Traditional Japanese Art	Traditional Culture	189
United States	Contemporary Times	Digital Art	Fusion of Technology and Art	257

Table 1. (*Continued*).

Region	Historical period	Art movement	Cultural aspect	Number of Chinese character sculptures
France	Renaissance	Modernism	Art and Philosophy Integration	172
South Korea	Contemporary Times	Korean Traditional Sculpture	Contemporary Aesthetics	196
Singapore	Contemporary Times	Avant-garde Art	Cross-cultural Communication	138
Australia	Contemporary Times	Indigenous Art	Environmental Protection	111

Through steps such as size adjustment, grayscale standardization, and brightness equalization, the obtained Chinese character sculpture images are normalized to ensure the consistency of the data in subsequent machine learning models:

The image pixels are weighted and summed to achieve size adjustment of Chinese character images. The linear interpolation process is shown in Equation (1):

$$I_a(x,y) = \sum_i \sum_j I_o(i,j) \cdot K(x-i, y-j) \tag{1}$$

Here, I_o represents the original image; I_a represents the resized image; K represents the interpolation kernel function. In the grayscale standardization stage, Z-score standardization is used to process each pixel in the image according to Equation (2):

$$I_s(x,y) = \frac{I_a(x,y) - \mu}{\sigma} \tag{2}$$

Among them, μ is the average grayscale value of the image, and σ is the standard deviation of the image. Through this step, it is ensured that the distribution of image grayscale values has zero mean and unit variance. Histogram equalization is further utilized to achieve brightness equalization, and the cumulative distribution function of the image is mapped to a uniform distribution, as shown in Equation (3):

$$I_b(x,y) = HistEqualize(I_s(x,y)) \tag{3}$$

Here, the mapping relationship of pixel values is adjusted through histogram equalization to improve the overall image contrast. The partial Chinese character sculpture images processed through the above steps are shown in **Figure 1**.

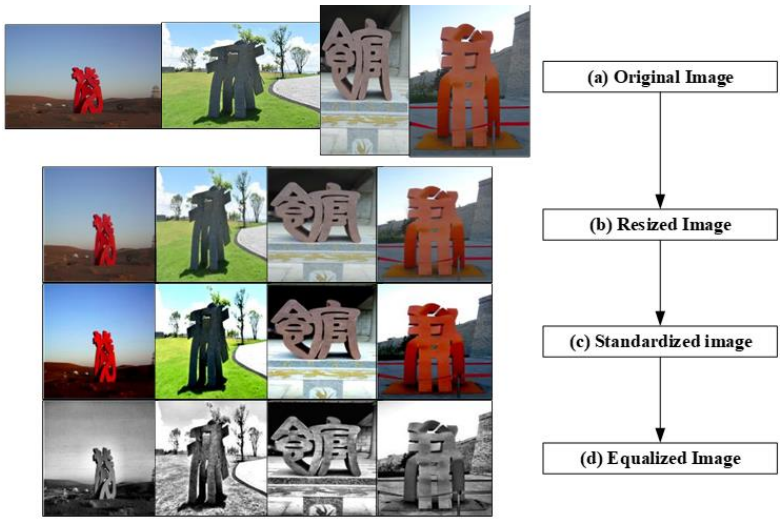

Figure 1. Normalization processing of Chinese character sculpture images.

Figure 1 shows the processing process for Chinese character sculpture images, including original images, size adjustment, standardization, and equalization. According to **Figure 1**, the normalization of the image is successfully achieved while eliminating heterogeneity, providing a stable data foundation for subsequent

machine learning models. The Chinese character sculpture image is generated into a multidimensional model in space, and the image based 3D reconstruction method SfM (Structure from Motion) is utilized to reconstruct the 3D structure from the Chinese character sculpture image, as shown in **Figure 2**.

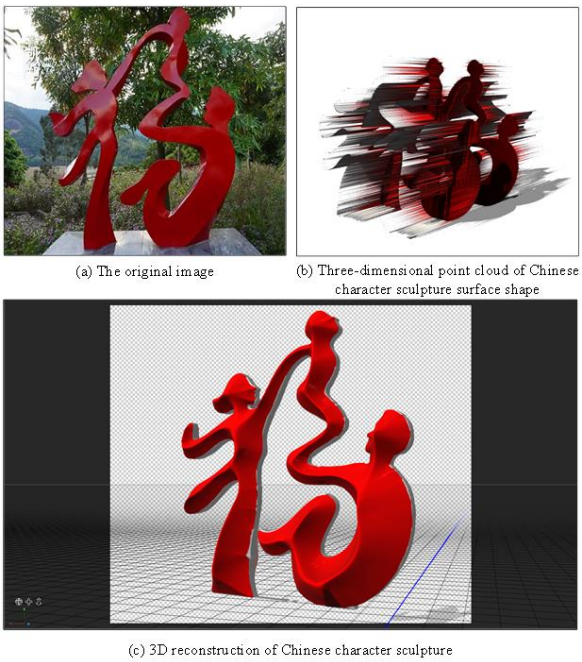

Figure 2. Reconstruction process of Chinese character sculpture images in three-dimensional space.

Figure 2 shows the reconstruction process of Chinese character sculpture images in three-dimensional space. The feature points of the original sculpture image in **Figure 2a** are matched to three-dimensional space, and the surface shape of the Chinese character sculpture is reconstructed through triangulation, generating a three-dimensional point cloud composed of a large number of points in **Figure 2b**. Each point corresponds to a position on the surface of the object, resulting in the Chinese character sculpture model shown in **Figure 2c**.

By visiting art magazines, exhibition reviews, and online art platforms, this study collects expert analysis and evaluations of different Chinese character sculptures, and obtains professional perspectives on the aesthetic characteristics of Chinese character sculptures. Referring to the actual exhibition guides of multiple museums, the interactive experience descriptions of the audience in the exhibition are extracted. The text content mainly comes from the exhibition guide manual, visitor message book, and actual feedback from visitors officially released by the museum, in order to capture the audience's intuitive feelings and understanding of Chinese character sculptures. Multiple experts and scholars have collected academic descriptions of Chinese character sculpture. This study covers multiple fields such as art history, cultural studies, and sculpture theory, providing deeper cultural background and theoretical support for research.

The FastText model is adopted to semantically correspond the collected text data with Chinese character sculpture images, with a focus on considering the n-grams information of Chinese characters. The Chinese text is segmented to obtain the vocabulary sequence of the Chinese text; the word embedding processing is carried out. If dictionary D contains embedded representations of all characters, then for a Chinese character w, its word embedding representation is calculated as Equation (4):

$$Embedding(w) = \frac{1}{|w|}\sum_{n=1}^{|w|} D[w_n] \tag{4}$$

Here, w_n represents the nth character of the Chinese character w, and |w| is the length of the word. The semantic correspondence between text and images is further established, and the weight matrix is adjusted. The text processed by word embedding is mapped into the space of image embedding, enabling effective

comparison between text and image in a common semantic space. Some text mapping results are shown in **Table 2**.

Table 2. Text mapping results.

Text ID	Chinese text	Word embedding visualization location	Image embedding similarity score	Equivalent vocabulary size	Text length
001	Classical style sculpture	(2.5, 3.8)	0.89	300	20
002	Modern innovative design Chinese sculpture	(−1.2, 4.5)	0.75	400	25
003	Traditional cultural expression	(0.9, 1.2)	0.92	250	15
004	Abstract artwork	(−3.0, −2.5)	0.81	350	18
005	Unique Creation of contemporary artists	(4.2, 0.5)	0.88	280	22

Table 2 shows some of the text mapping results. Among them, the text ID provides a unique identifier for each text, used to track and reference each text. The Chinese character text describes the content related to each text and Chinese character sculpture. The visual position of word embedding displays the coordinates of the word embedding in the two-dimensional space obtained through dimensionality reduction technology, reflecting the relative position of each text in the embedding space. The image embedding similarity score displays the similarity score between each text and the corresponding image embedding, indicating the degree of semantic correspondence between the text and the image in the shared embedding space. The equivalent vocabulary reflects the number of words used to describe the content of Chinese character sculptures, and a larger vocabulary represents a richer description. The text length provides the number of characters per Chinese character text, used to evaluate text complexity and information density.

4. Unsupervised training of CLIP model

The unsupervised training phase of the CLIP model uses a contrastive learning method to minimize the negative logarithmic likelihood loss between image and text embeddings, forcing the model to learn to embed related images and text closely, while separating irrelevant images and text embeddings, and establishing cultural attribute representations in the embedding space. In this study, the CLIP model was enhanced to analyze the cultural attributes of Chinese character sculptures more effectively. Key improvements include a strengthened dual-pathway encoder architecture, with the text encoder expanded from 15 to 30 layers, each featuring 76 self-attention heads and a hidden layer dimension of 1280, specializing in processing Chinese text data. The image encoder, based on an enhanced Vision Transformer design, has 40 self-attention heads per layer and an expanded hidden layer size of 2560, processing 20 images per second at a resolution of 1000 × 1000 pixels. A fusion layer with 512 neurons effectively combines high-dimensional text and image feature vectors. The model's downstream segment incorporates contrastive learning with complex loss function optimization, a learning rate of 0.0005, and a batch size of 512, reducing overfitting risk. The output layer, comprising 200 neurons, generates the final categorization results. These improvements significantly boost the model's performance in processing and understanding the cultural attributes of Chinese character sculptures.

For each sample, a positive sample pair (I, T) is constructed, where I and T are related image and text embeddings, respectively. At the same time, negative sample pair (I', T') is constructed by selecting samples of different categories from the dataset, where I' and T' are image and text embeddings that are not related to I and T, respectively. The distance metric in the embedded space is defined as cosine similarity, as shown in Equation (5):

$$d(I,T) = \frac{I \cdot T}{|I| \cdot |T|} \tag{5}$$

Among them, $I \cdot T$ represents the dot product of vectors I and T, and $|I|$ and $|T|$ represent the norm

of vectors *I* and *T*, respectively. In comparative learning, the cosine similarity of positive sample pairs should be close to 1, while the cosine similarity of negative sample pairs should be close to −1. The loss function of the comparison function adopts negative logarithmic likelihood loss, as shown in Equation (6):

$$\mathcal{L} = -\log\left(\frac{exp(d(I,T)/\tau)}{exp\left(\frac{d(I,T)}{\tau}\right) + \sum_{i=1}^{N} exp(d(I',T_i)/\tau)}\right) \tag{6}$$

Here, N represents the number of negative samples, and τ is a temperature parameter used to balance the distance between positive and negative samples. Using the gradient descent method, the negative logarithmic likelihood loss of contrastive learning is minimized by adjusting model parameters to minimize the distance between positive sample pairs and maximize the distance between negative sample pairs, as shown in **Figure 3**.

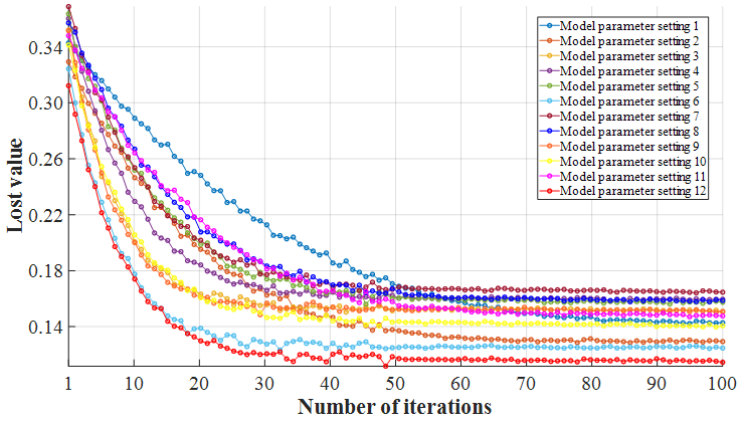

Figure 3. Gradient descent process.

Figure 3 shows the gradient descent process of negative logarithmic likelihood loss. The horizontal axis represents the number of iterations (1–100), and the vertical axis represents the loss value. Different curves represent different model parameters. As the number of iterations increases, the model loss under each parameter gradually decreases. The 12th parameter setting performs best and is used in the study. The gradient calculation of the loss function on the model parameters is shown in Equation (7):

$$\frac{\partial \mathcal{L}}{\partial \theta} = -\frac{1}{\tau}\left(T - \frac{exp(I \cdot T/\tau)}{exp\left(I \cdot \frac{T}{\tau}\right) + \sum_{i=1}^{N} exp(I' \cdot T_i/\tau)}\right) \cdot \frac{\partial (I \cdot T)}{\partial \theta} \tag{7}$$

Among them, θ represents the model parameters. By extracting multi-layer embeddings in the trained CLIP model, the cultural attribute representation C of Chinese character sculptures is obtained. Among them, $C = \{C_1, C_2, \cdots C_M\}$. M is the number of embedded layers. By completing unsupervised training of comparative learning, the CLIP model successfully establishes a cultural attribute representation of Chinese character sculptures in the embedding space.

5. Extraction and aggregation of cultural attribute features of Chinese character sculpture

The gradient weighted mapping method is used to visualize the model's attention to Chinese character sculpture images and present the model's decision-making process in an interpretable manner.

The feature map $A^{(k)}$ of the last layer in the CLIP model is considered, where k represents the channel index. For the category score y_c output by the model, directional propagation is used to calculate the gradient relative to the category score, as shown in Equation (8):

$$\frac{\partial y_c}{\partial A^{(k)}} \tag{8}$$

This gradient represents the direction in which the category score increases by y_c, which is the degree of attention the model places on that category. The global average pooling on the gradient of feature map $A^{(k)}$ is performed to obtain weight α_k on channel k, as shown in Equation (9):

$$\alpha_k = \frac{1}{Z}\sum_i \sum_j \frac{\partial y_c}{\partial A^{(k)}}(i,j) \tag{9}$$

Here, Z is the spatial size of the gradient. The obtained weight α_k is used to weight and sum the feature maps to generate a thermal map L_c, as shown in Equation (10):

$$L_c = ReLU(\sum_k \alpha_k A^{(k)}) \tag{10}$$

Among them, $ReLU$ represents correcting the linear unit to ensure that the result is non negative. The process concentrates the heat map L_c on the key areas identified by the model as category c, thereby providing a visual display of model decisions. To map the thermal map back to the original image space, the thermal map is normalized according to Equation (11):

$$\hat{L}_c(i,j) = \frac{L_c(i,j)}{\max(L_c)} \tag{11}$$

Here, $\max(L_c)$ represents the maximum value of the thermodynamic diagram. The regions of interest for CLIP model in Chinese character sculpture images are shown in **Figure 4**.

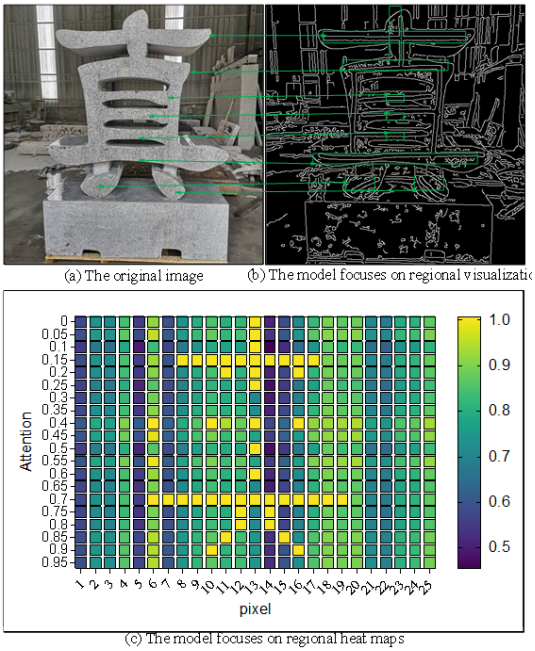

Figure 4. CLIP model focus area visualization.

Figure 4 respectively shows the visualization results of the original image, CLIP model's focus area, and the model's focus area heat map. In **Figure 4b**, green wireframes are used to mark the key focus areas of the model, while green lines are used to correspond to the content in the original image. In the thermal diagram shown in **Figure 4c**, the horizontal axis represents different pixels, and the vertical axis represents the degree of attention to the region. K-means clustering algorithm is used to reveal the internal structure and correlation of cultural attributes of Chinese character sculpture. By iteratively optimizing the clustering center, samples are allocated to the nearest center to more accurately depict the cultural attribute group structure of Chinese character sculptures. The clustering process is shown in **Figure 5**.

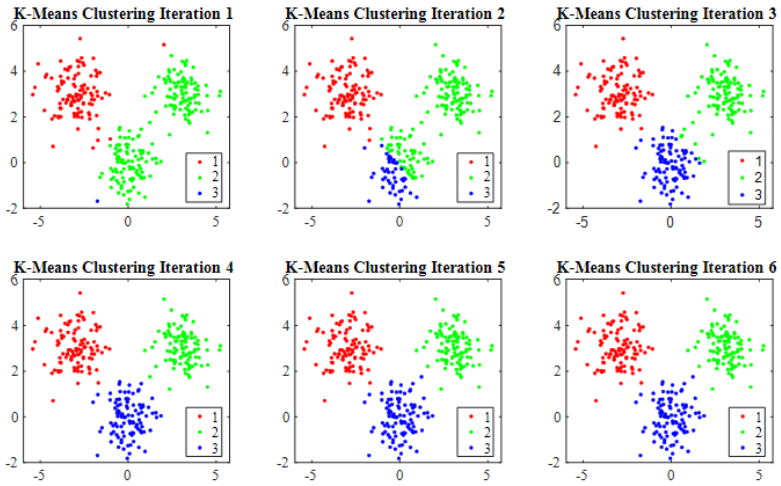

Figure 5. Clustering process of cultural attributes of Chinese character sculptures.

Figure 5 shows the process of clustering the cultural attributes of Chinese character sculpture "Zhen" six times using the K-means clustering algorithm. After reducing the cultural characteristics of Chinese character sculptures to 2D and normalizing them, the X and Y axes in **Figure 5** are formed. It can be seen that the data is divided into three clusters, corresponding to the three structural regions that the CLIP model focuses on for the "Zhen" of Chinese character sculptures. Initially, the cluster center is randomly selected, and through 6 iterations, the algorithm gradually adjusts the cluster center to better capture the data structure. Finally, the data points of each cluster are assigned to the nearest cluster center, forming a clustering division of the data. The CLIP model is used to analyze the cultural attributes of Chinese character sculptures with the character "Zhen" as an example. The results are shown in **Figure 6**.

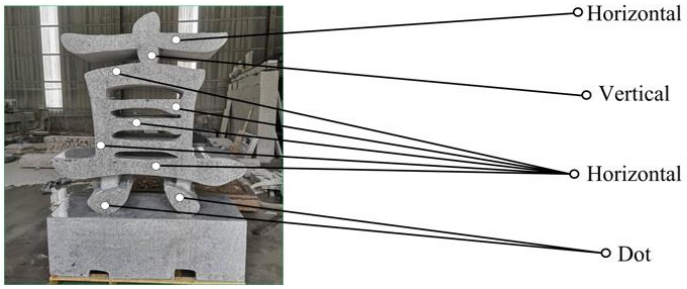

Figure 6. Cultural attribute analysis of Chinese character sculpture with the character "Zhen" under the CLIP model.

Figure 6 shows the cultural attribute analysis of the Chinese character sculpture with the "Zhen" character under the CLIP model. It can be seen that the machine learning model can effectively analyze the font structure of "Zhen" character sculptures, and relevant cultural attributes are provided to solve the challenges of existing research in processing multi-dimensional Chinese character sculpture spatial data. The following are the results of the information fields analyzed by the model:

Horizontal: In Chinese characters, the horizontal stroke represents the horizon and carries the symbolic significance of balance and stability. In the character "真" (zhēn), the upper horizontal stroke maintains stability in its structure, conveying the cultural connotations of balance and stability, reflecting the traditional Chinese cultural pursuit of social order stability.

Vertical: Commonly used to represent a vertical line, the vertical stroke has directional qualities pointing upwards or downwards. In the character "真" (zhēn), the vertical stroke supports the entire character, giving a sense of integrity and firmness, embodying traditional Chinese cultural values of loyalty, integrity, and steadfastness.

Horizontal: The lower horizontal stroke in the character "具" (jù) forms a small "口" character, symbolizing speech and language. In the character "真" (zhēn), the lower horizontal stroke of the "具" character associates "真" with the concepts of truth and sincerity, emphasizing the importance of truthfulness and sincerity, reflecting the moral pursuit of these values in traditional Chinese culture.

Dot: The dot at the lower end, together with the horizontal stroke, forms the lower part of the "具" character, representing the meaning of "具" and symbolizing speech and language. In the character "真" (zhēn), the dot connects "真" with the expression of truth and sincerity, emphasizing the importance of truthfulness and sincerity, reflecting the moral pursuit of these values in traditional Chinese culture.

6. Model evaluation

The leave-one-out cross validation is used as the dataset partitioning method. In each validation round, one sample in the dataset is used as the validation set, while the remaining samples are used for model training. The process is repeated until each sample is used as a validation set. To comprehensively evaluate the performance of the model, four indicators are adopted: precision, recall, *F1* value, and accuracy. The precision calculation is shown in Equation (12):

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{12}$$

Here, True Positives (TP) represents the number of samples correctly predicted as positive categories by the model, while False Positives (FP) represents the number of samples incorrectly predicted as positive categories by the model. The precision range is between 0 and 1, with higher values indicating that the model can more accurately identify positive categories. The proportion of successfully predicted positive categories in all actual positive category samples is calculated through recall rate, as shown in Equation (13):

$$Recall = \frac{True\ Positives}{True\ Positoves + False\ Negatives} \tag{13}$$

Among them, False Negatives (FN) represents the number of samples that the model incorrectly predicts as negative categories. The *F1* value is further utilized by taking into account the precision and recall of the model, as calculated by Equation (14):

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{14}$$

Accuracy is used as an overall performance indicator to evaluate the predictive accuracy of the model for all cultural attributes, as shown in Equation (15):

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total\ Samples} \tag{15}$$

Here, True Negative (TN) is the number of samples correctly predicted by the model as negative categories, and Total Samples is the number of all samples. The precision, recall, *F1* value, and accuracy calculation results of CLIP and M3, Dual-Attention Network, Image-Text Embedding Model, and ViLBERT models for analyzing Chinese character sculptures with 10 different cultural attributes are shown in **Figure 7**.

Figure 7 shows the calculation results of the precision, recall, *F1* value, and accuracy of the CLIP model used in this article, along with models such as M3, Dual-Attention Network, Image-Text Embedding Model, and ViLBERT, for analyzing Chinese character sculptures with 10 different cultural attributes, including historical tradition, contemporary innovation, cultural diversity, urban culture, traditional culture, technology and art integration, art and philosophy integration, contemporary aesthetics, cross-cultural communication, and environmental protection. The horizontal axis in **Figure 7** represents different cultural attributes, while the vertical axis represents calculated values. The average precision, recall, *F1* value, and accuracy of the other four models for analyzing Chinese character sculptures with 10 different cultural attributes were 0.92, 0.92,

0.92, 0.91; 0.93, 0.92, 0.92, 0.90; 0.95, 0.93, 0.94, 0.94; 0.92, 0.88, 0.89, 0.93, respectively. When the CLIP model achieved higher performance, the calculated average results under all four different indicators reached 0.98. The machine model used in this study has better accuracy in analyzing the cultural attributes of Chinese character sculptures.

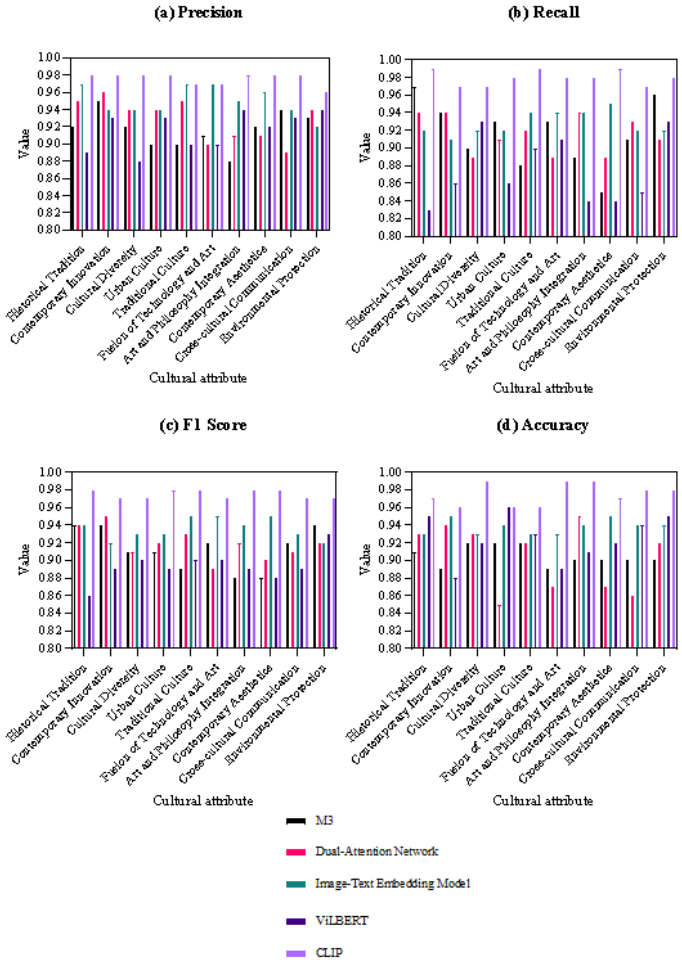

Figure 7. Comparison results between CLIP and four other models.

The AUC was used to evaluate the adaptability of the model to Chinese character sculptures with different cultural attributes, and ROC (Receiver Operating Characteristic) curves were drawn to visualize the performance of different models under different thresholds, as shown in **Figure 8**.

Figure 8 show the ROC curves of CLIP and other four models for the classification of Chinese character sculptures with 10 different cultural attributes. The horizontal axis represents FPR (False Positive Rate), and the vertical axis represents TPR (True Positive Rate). It can be seen that the ROC curve corresponding to the CLIP model was closer to the upper left corner, indicating high performance. When the curves of the other four models for classifying Chinese character sculptures with 10 different cultural attributes were more dispersed, the CLIP model can effectively distinguish and recognize different cultural attributes of Chinese character sculptures. The AUC mean values of M3, Dual-Attention Network, Image-Text Embedding Model, ViLBERT, and CLIP models for Chinese character sculpture classification under 10 different cultural attributes were 0.93, 0.92, 0.91, 0.87, and 0.98, respectively. The CLIP model had an improvement of approximately 5.4% compared to the best performing M3 model among other models.

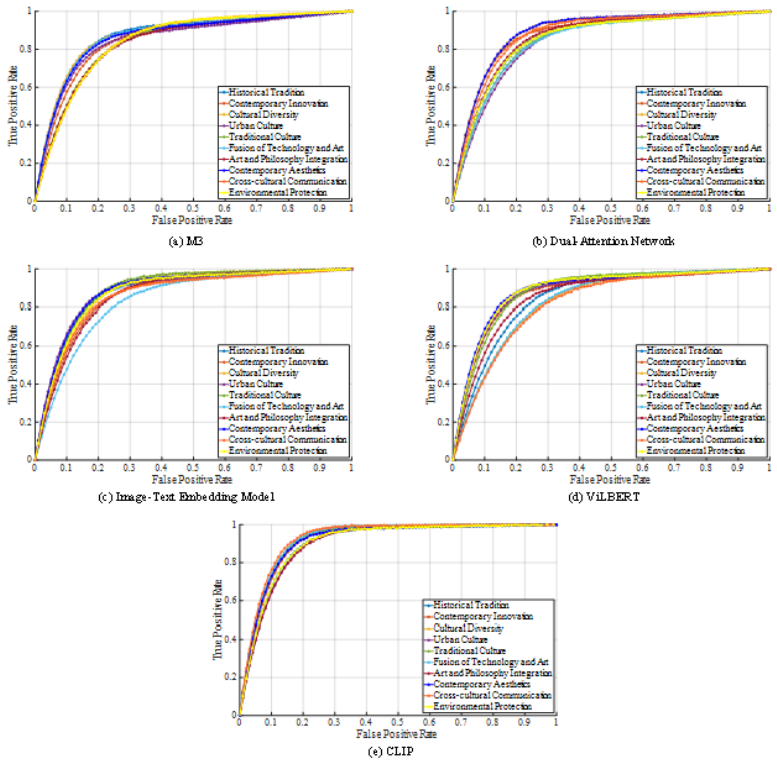

Figure 8. ROC curves of different models for classifying Chinese character sculptures with different cultural attributes.

The difference between the predicted value of the model and the actual value was measured using RMSE and MAE, respectively, and was calculated using Equations (16) and (17):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{16}$$

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n} \tag{17}$$

Here, n is the number of samples; y_i is the actual value; $\hat{y}_i$ is the predicted value of the model. The results of *RMSE*, *MAE*, and analysis duration are shown in **Table 3**.

Table 3. *RMSE*, *MAE*, and analysis time of different models for predicting the cultural attributes of Chinese character sculptures.

Model	***RMSE***	***MAE***	**Analysis time**
CLIP	0.034	0.025	182 ms
M3	0.042	0.032	231 ms
Dual-Attention Network	0.039	0.028	573 ms
Image-Text Embedding Model	0.036	0.026	459 ms
ViLBERT	0.038	0.030	864 ms

According to **Table 3**, compared to the M3, Dual-Attention Network, Image-Text Embedding Model, and ViLBERT models, the CLIP used in the study achieved the fastest analysis speed (182 ms) with the lowest *RMSE* (0.034) and *MAE* (0.025), enabling efficient and accurate analysis of the cultural attributes of Chinese character sculptures.

The significant advantages of the CLIP model in the task of analyzing the cultural attributes of Chinese character sculptures stem from a series of innovations in its design and implementation. The adopted leave-one-out cross-validation method ensures that each sample is fully utilized, thereby enabling the model to receive balanced training and validation across the entire dataset, enhancing the model's generalization ability to new data.

The CLIP model's outstanding performance on multiple core performance indicators, such as precision, recall, F1 score, and accuracy, reflects its efficiency and reliability in handling complex cultural attribute analysis tasks. In terms of precision and recall, these two indicators respectively measure the accuracy and completeness of the model in identifying positive categories. High precision means a low rate of false positives, while high recall ensures that the model captures most of the true positive samples, demonstrating the CLIP model's effective balance between minimizing errors and maximizing coverage.

The high performance of the CLIP model in AUC assessment further proves its ability to maintain consistent performance across different thresholds. The ROC curve being close to the top left corner indicates its ability to achieve a high true positive rate while maintaining a low rate of false positives. Such characteristics are particularly important in handling the analysis of Chinese character sculptures with diverse and complex cultural attributes.

In terms of prediction accuracy, the low *RMSE* and *MAE* values exhibited by the CLIP model indicate a smaller deviation between its predictions and the actual values, further emphasizing the model's ability to precisely capture and interpret data. Additionally, its rapid analysis speed indicates the efficiency of the CLIP model in processing large-scale data, which is particularly important for modern application scenarios that require quick responses and the processing of large amounts of data.

Overall, the superior performance of the CLIP model in the task of analyzing the cultural attributes of Chinese character sculptures is derived from its innovative approaches in data processing, model structure, and training methods. These advantages enable the CLIP model to excel not only in accuracy and efficiency but also in handling complex and diverse cultural attribute analysis tasks with exceptional adaptability and reliability.

7. Conclusions

By introducing a machine learning model, this article successfully achieved a deep analysis of the cultural attributes of Chinese character sculptures. Innovative methods such as normalization, 3D reconstruction, and semantic correspondence were adopted to establish effective connections between images and text. The experimental results showed that the machine learning model performed excellently in analyzing Chinese character sculptures with diverse cultural attributes. At the same time, the model achieved a balance between precision and speed, demonstrating excellent performance. However, research still needs to focus on the model's generalization ability and robustness to different samples. Future research can focus on optimizing algorithms and expanding datasets to more comprehensively address the analytical challenges of complex cultural attributes.

Conflict of interest

The author declares no conflict of interest.

References

1. Koutsabasis P, Vosinakis S. Kinesthetic interactions in museums: conveying cultural heritage by making use of ancient tools and (re-) constructing artworks. Virtual Reality. 2017, 22(2): 103-118. doi: 10.1007/s10055-017-0325-0
2. Guo X. Analysis on the Style and Evolution of the Sculpture Art in Shanxi Merchants Courtyard. Highlights in Art and Design. 2023, 3(2): 57-59. doi: 10.54097/hiaad.v3i2.10229
3. Liang W, Tadesse GA, Ho D, et al. Advances, challenges and opportunities in creating data for trustworthy AI. Nature Machine Intelligence. 2022, 4(8): 669-677. doi: 10.1038/s42256-022-00516-1
4. Li H, Wang W, Li Q, et al. A novel minimum-time feedrate schedule method for five-axis sculpture surface machining with kinematic and geometric constraints. Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture. 2018, 233(5): 1483-1499. doi: 10.1177/0954405418780167
5. Barreau JB, Jouneau J, Charlet C, et al. Digitization, Virtual Reality and Robotic Sculpture for the Preservation

and Enhancement of the Public Heritage of the Sculpted Rocks of Rothéneuf. Journal on Computing and Cultural Heritage. 2022, 15(4): 1-21. doi: 10.1145/3522595
6. García-Molina DF, López-Lago S, Hidalgo-Fernandez RE, et al. Digitalization and 3D Documentation Techniques Applied to Two Pieces of Visigothic Sculptural Heritage in Merida Through Structured Light Scanning. Journal on Computing and Cultural Heritage. 2021, 14(4): 1-19. doi: 10.1145/3427381
7. Bovcon N. Virtual museums: interpreting and recreating digital cultural content. Neohelicon. 2021, 48(1): 23-38. doi: 10.1007/s11059-021-00582-1
8. Rani A. Digital Technology: It's Role in Art Creativity. Journal of Commerce & Trade. 2018, 13(2): 61. doi: 10.26703/jct.v13i2-9
9. Christidou D, Pierroux P. Art, touch and meaning making: an analysis of multisensory interpretation in the museum. Museum Management and Curatorship. 2018, 34(1): 96-115. doi: 10.1080/09647775.2018.1516561
10. Zhilin M, Savchenko S, Hansen S, et al. Early art in the Urals: new research on the wooden sculpture from Shigir. Antiquity. 2018, 92(362): 334-350. doi: 10.15184/aqy.2018.48
11. Cialone C, Tenbrink T, Spiers HJ. Sculptors, Architects, and Painters Conceive of Depicted Spaces Differently. Cognitive Science. 2017, 42(2): 524-553. doi: 10.1111/cogs.12510
12. Yang J. Analysis of the Clever Application of Sculpture Language in Modern Sculpture. Tiangong. 2018, (1): 16-17.
13. Scott DA. Ancient Marbles: Philosophical Reflections on the Restoration of Sculpture. Studies in Conservation. 2022, 68(4): 388-406. doi: 10.1080/00393630.2022.2049032
14. Fraser H. Grief encounter: the language of mourning in fin-de-siècle sculpture. Word & Image. 2018, 34(1): 40-54. doi: 10.1080/02666286.2017.1333880
15. Barata M. The Sculpture of Black Origin in Brazil. Art in Translation. 2022, 14(1): 96-102. doi: 10.1080/17561310.2022.2046532
16. Mihai Ionut R. Technology and imagination in contemporary art. aspects of modern sculptural object. Limba Si Literatura–Repere Identitare in Context European. 2018, 22(22): 276-284.
17. Heginbotham A, Erdmann R, Hayek LAC. The dating of French gilt bronzes with ED-XRF analysis and machine learning. Journal of the American Institute for Conservation. 2018, 57(4): 149-168. doi: 10.1080/01971360.2018.1515389
18. Jabbarov RR. Patterns in applied art of the Uzbek folk. European Journal of Arts. 2023, (1): 11-14. doi: 10.29013/eja-23-1-11-14
19. Pulham P. The Sculptural Body in Victorian Literature. Edinburgh University Press, 2020. doi: 10.1515/9780748693436
20. Ullah I, Soomro MA, Mudassar Zulfiqar. A Review of Archaeological Reports and Literature on the Gandhara Sculpture Collection of the Royal Ontario Museum. Academic Journal of Social Sciences (AJSS). 2020, 4(3): 377-403. doi: 10.54692/ajss.2020.04031212
21. Zhao J, Liu X, Luo W, et al. Research on multimodal search tools for military image resources based on CLIP model. Chinese Journal of Medical Library and Information. 2022, 31(8): 14-20.
22. Baldrati A, Bertini M, Uricchio T, et al. Composed Image Retrieval using Contrastive Learning and Task-oriented CLIP-based Features. ACM Transactions on Multimedia Computing, Communications, and Applications. 2023, 20(3): 1-24. doi: 10.1145/3617597
23. Chen Z, Du H, Wu Y, et al. Cross modal video clip retrieval based on visual text relationship alignment. Chinese Science: Information Science. 2020, 50(6): 862-876. doi: 10.1360/SSI-2019-0292
24. Du P, Li X, Gao Y. A Review of Research on Multimodal Visual Language Representation Learning. Journal of Software Science. 2020, 32(2): 327-348.
25. Zhang L, Zhang L, Yuan Q. Remote sensing big models: progress and prospects. Journal of Wuhan University (Information Science Edition). 2023, 48(10): 1574-1581.
26. Wang C. Artificial Intelligence Driven Digital Image Art Creation: Methods and Case Analysis. Journal of Intelligent Science and Technology. 2023, 5(3): 406-414.
27. Chefer H, Alaluf Y, Vinker Y, et al. Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models. ACM Transactions on Graphics. 2023, 42(4): 1-10. doi: 10.1145/3592116
28. Jiang D, Ye M. Transformer network for cross modal text to image pedestrian recognition [J]. Chinese Journal of Image and Graphics. 2023, 28(5): 1384-1395.
29. Liu T, Wu Z, Chen J, Jiang Y. A multimodal pre training method for visual language understanding and generation. Journal of Software. 2022, 34(5): 1-11. doi: 10.13328/j.cnki.jos.006770
30. Gal R, Patashnik O, Maron H, et al. StyleGAN-NADA. ACM Transactions on Graphics. 2022, 41(4): 1-13. doi: 10.1145/3528223.3530164
31. Huang Z, Bianchi F, Yuksekgonul M, et al. A visual–language foundation model for pathology image analysis using medical Twitter. Nature Medicine. 2023, 29(9): 2307-2316. doi: 10.1038/s41591-023-02504-3
32. Avrahami O, Fried O, Lischinski D. Blended Latent Diffusion. ACM Transactions on Graphics. 2023, 42(4): 1-11. doi: 10.1145/3592450
33. Zhou K, Yang J, Loy CC, et al. Learning to Prompt for Vision-Language Models. International Journal of Computer Vision. 2022, 130(9): 2337-2348. doi: 10.1007/s11263-022-01653-1

34. Gal R, Arar M, Atzmon Y, et al. Encoder-based Domain Tuning for Fast Personalization of Text-to-Image Models. ACM Transactions on Graphics. 2023, 42(4): 1-13. doi: 10.1145/3592133
35. Zheng Z, Zheng L, Garrett M, et al. Dual-path Convolutional Image-Text Embeddings with Instance Loss. ACM Transactions on Multimedia Computing, Communications, and Applications. 2020, 16(2): 1-23. doi: 10.1145/3383184