International Conference on Humanities, Social and Management Sciences (HSMS 2026)

Challenges and Technical Solutions to the Construction of Chinese-English Parallel Corpus for Ancient Chinese Painting & Calligraphy

Hanmin Zhang[1], Qiang Zhang[2,*]
[1]*Office of Scientific Research, Wuhan City Polytechnic, Wuhan, Hubei, China*
[2]*School of Foreign Languages, Central China Normal University, Wuhan, Hubei, China*
**Corresponding Author*

Abstract: The cross-cultural translation and intelligent understanding of ancient Chinese painting and calligraphy are seriously restricted by the shortage of high-quality domain-specific parallel corpora. The major bottlenecks include cultural semantic gap of professional terminology, scarcity of low-resource parallel data, multi-modal text-image misalignment, and lack of standardized annotation specifications. Targeting painting and calligraphy literature of Tang, Song, Yuan, Ming and Qing dynasties, this paper systematically analyzes the key technical difficulties in constructing Chinese-English parallel corpus. A comprehensive technical framework integrating domain knowledge engineering, semi-automated alignment, multi-modal feature fusion and knowledge-enhanced annotation is proposed. The framework optimizes the whole workflow of data collection, intelligent cleaning, text preprocessing and hierarchical annotation, and solves the problems of cultural semantic mismatch and low automation in traditional corpus construction. Quantitative experiments verify that the proposed hybrid alignment and knowledge-enhanced annotation mechanism significantly improve bilingual sentence alignment accuracy and domain terminology consistency. This research provides standardized high-quality data resources for cross-cultural translation, intelligent retrieval and machine learning in ancient Chinese art, and supports the international dissemination of traditional Chinese art heritage.

Keywords: Parallel Corpus; Ancient Chinese Painting and Calligraphy; Cultural Semantic Gap; Multi-Modal Alignment; Knowledge-Enhanced Annotation

1. Introduction

Ancient Chinese painting and calligraphy are the core carrier of traditional Chinese culture, carrying unique aesthetic thoughts, philosophical connotations and artistic values [1]. With the development of global cultural exchange and digital cultural heritage construction, accurate cross-lingual translation and semantic parsing of painting and calligraphy texts including treatises, colophons, inscriptions and critiques have become urgent demands in computational linguistics and digital humanities. High-quality parallel corpus is the fundamental underlying resource for domain machine translation, cross-lingual information retrieval, multi-modal intelligent understanding and cultural knowledge graph construction.

Existing general and cultural parallel corpora are mainly oriented to modern daily language and general literary texts, without targeted structural design for ancient painting and calligraphy. Different from ordinary texts, painting and calligraphy literature features classical Chinese syntax, culture-loaded terms, abstract aesthetic connotations and inherent text-image correlation. Traditional corpus construction methods cannot adapt to the low-resource attribute and cultural semantic heterogeneity of this domain, resulting in small scale, uneven quality and poor reusability of existing art parallel resources [2]. Therefore, constructing a dedicated Chinese- English parallel corpus for ancient painting and calligraphy has important theoretical value and engineering significance for breaking the bottleneck of cross-cultural intelligent communication of traditional art.

The corpus construction faces four intertwined technical challenges:

Cultural Semantic Gap: Core aesthetic and technical concepts such as Qiyun Shengdong,

Gufa Yongbi and Xieyi have no fully equivalent English expressions. Random translation causes semantic loss and ambiguity, and there is no unified industry translation standard [3].
Low-Resource Data Scarcity: As a typical low-resource vertical domain, authoritative parallel translations of classical painting treatises, colophon poems and critiques are extremely scarce [4]. Scattered existing translations suffer from terminology inconsistency and mistranslation, which are unavailable for model training and corpus construction.
Multi-modal Misalignment: Painting and calligraphy are multi-modal carriers integrating images, colophons, seals and inscriptions [5]. The traditional pure-text isolated construction mode fails to realize deep semantic alignment between visual features and text semantics, leading to the loss of artistic context information.
Lack of Standardized Annotation: There is no unified hierarchical annotation specification for domain terms, aesthetic features and implicit cultural semantics. The absence of quantitative evaluation indicators restricts corpus quality control and academic resource sharing [6].
This paper makes three main academic and technical contributions:
From the perspective of intelligent engineering, this paper sorts out multi-dimensional bottlenecks of parallel corpus construction in ancient painting and calligraphy, and clarifies the internal logical correlation of various technical difficulties.
A complete technical framework integrating domain knowledge engineering, multi-source intelligent cleaning, semi-automated hybrid alignment, multi-modal feature fusion, and three-tier knowledge-enhanced annotation is proposed, forming a scalable engineering paradigm suitable for low-resource cultural art domains.
A pilot parallel corpus is constructed and quantitatively evaluated. Experimental results prove that the proposed framework effectively improves alignment accuracy, terminology consistency and multi-modal matching precision, providing empirical reference for large-scale corpus expansion and intelligent application development.

2. Related Work

2.1 Domain-Specific Parallel Corpus Construction

Parallel corpus construction technologies have formed a mature system in medicine, law, science and technology, and have been widely applied in machine translation and natural language understanding. In recent years, scholars have explored corpus construction for low-resource languages and niche vertical fields, proposing schemes such as web crawler-based data acquisition, pre-trained model-driven noise reduction, and human-machine collaborative alignment.
However, research on corpora in the cultural heritage domain remains relatively inadequate. Most existing art-related corpora are limited to simple monolingual term collation or small-scale parallel sentence pairs, and lack systematic multi-modal structural design and standardized quality control. Existing bilingual databases for calligraphy only achieve basic noun comparison, without in-depth semantic annotation and text-image multi-modal correlation, which can hardly satisfy the training requirements of domain-specific intelligent models. Literary parallel corpora constructed from modern novels and proses cannot be directly applied to classical Chinese painting and calligraphy texts with complex syntactic structures and rich cultural connotations [7, 8].

2.2 Cultural Semantic Gap in Cross-Lingual Translation

Culturally Specific Items (CSIs) are the core difficulty restricting cross-lingual translation and parallel corpus alignment. Literal translation of traditional art aesthetic concepts often leads to ideological connotation loss, while free translation is easy to deviate from original contextual semantics. Knowledge-enhanced semantic embedding and cultural concept library construction have become effective methods to bridge cultural semantic gaps.
Current relevant studies mostly focus on daily culture and folk custom texts, while targeted research on abstract aesthetic terms and painting technique vocabulary of ancient Chinese painting and calligraphy remains insufficient. The lack of standardized bilingual translation paradigms for core artistic concepts leads to arbitrary translation in scattered resources, seriously affecting the consistency and availability of parallel corpora.

2.3 Multi-Modal Art Corpus Construction

With the rapid development of vision-language pre-trained models including CLIP and LLaVA, the construction of multi-modal corpora has emerged as a popular research direction. Western art datasets have achieved basic matching between art images and descriptive texts. Nevertheless, these datasets are built based on Western painting characteristics and fail to learn the distinctive brushwork, composition and artistic conception of traditional Chinese paintings.

Most existing multi-modal alignment algorithms are designed for ordinary natural images and general descriptive texts, so they are unable to capture the deep implicit semantic connections between visual features of Chinese paintings and classical colophon texts. Furthermore, current multi-modal art corpora seldom incorporate metadata such as creation dynasties, artistic school heritage and professional critiques. Such deficiencies lead to incomplete data dimensions and limit their application in advanced intelligent retrieval and aesthetic analysis tasks.

3. Core Challenges in Corpus Construction

3.1 Cultural and Semantic Barriers

Unique Domain Terminology: Ancient painting and calligraphy texts contain numerous domain-specific terms without direct English equivalents, covering technique terms, aesthetic evaluation terms, and genre and school terms [9]. These terms are deeply rooted in Chinese traditional philosophy and aesthetic system; arbitrary translation will lead to semantic distortion and terminology inconsistency.

Abstract Aesthetic Connotations: A large number of abstract aesthetic concepts run through painting and calligraphy critiques and colophon texts. Such concepts rely on perceptual artistic comprehension rather than literal interpretation with strong context dependence. General bilingual embedding models cannot capture deep-level aesthetic semantic connotation, resulting in low automatic parallel alignment accuracy [10].

3.2 Low-Resource and Quality Constraints

Authoritative Parallel Data Scarcity: The ancient painting and calligraphy domain represents a typical low-resource scenario in natural language processing. High-quality and complete parallel translations of classical painting treatises, long colophons and academic critiques are extremely scarce. Most existing English resources are only short museum introductions, which lack in-depth academic parallel texts to support large-scale corpus construction and model fine-tuning.

Data Noise and Inconsistency: Scattered non-professional online translations usually contain grammatical errors, non-standard English expressions and discrepant cultural translation. Direct use of these translated resources without intelligent filtering and expert review may result in corpus contamination and further weaken the generalization performance of relevant models. In addition, the same term often appears with various translations in different documents, leading to a lack of unified normative standards.

3.3 Textual and Multi-Modal Technical Difficulties

Complexity of Classical Chinese Texts: The ancient painting and calligraphy text is always composed in classical Chinese, which is accordingly characterized by concise syntactic structure, frequent elliptical expressions, integration of poetic and prose styles, and dense cultural allusions. Conventional classical Chinese segmentation models are also characterized by low adaptability to short critical comments and impromptu colophons, and prone to entity boundary ambiguity and semantic segmentation bias. Meanwhile, the irregular layout of inscriptions and postscripts further increases the difficulty of automatic text parsing.Multi-modal Semantic Misalignment: Semantically, painting images and colophon texts are dependent and explanatory to each other. Traditional corpus construction deals with store textual and visual content separately, thus separating visual features from textual semantics. What's more, the mainstream visual model is far from sufficient in recognizing ink brush strokes and blank-leaving composition, both of which are unique to Chinese paintings. As a result, it is rather difficult to achieve in-depth cross-modal semantic alignment at the artistic conception level.

3.4 Lack of Standardized Annotation and Quantitative Evaluation

The present parallel corpus for ancient Chinese painting and calligraphy is short of a unified hierarchical annotation framework. Conse-

quently, the term annotation, aesthetic semantic labeling, and multi-modal association tagging are hardly governed by any clear, unified criteria, giving rise to highly subjective outputs. This, in turn, restricts their effective sharing and reuse across research scenarios. Besides, current evaluation systems can hardly provide quantitative indicators for alignment precision, term consistency, and annotation standardization, making it impossible to implement closed-loop quality control throughout the entire corpus construction process.

4. Proposed Technical Framework

This study is about to establish a comprehensive technical framework with five modules, namely, domain-specific knowledge engineering, multi-source data collection and intelligent cleaning, semi-automated parallel alignment, multi-modal feature fusion, and knowledge-enhanced hierarchical annotation. The proposed framework aims at unifying knowledge-driven paradigms, algorithmic optimization strategies, and expert intervention mechanisms to support high-quality corpus construction.

4.1 Domain-Specific Knowledge Engineering

Bilingual Terminology Database: In this study, the domain-specific terms are classified into six categories, namely creation techniques, aesthetic evaluation, art schools, work forms, character appellations, and historical dynasties.

Based on the authoritative translations from official museum documents and classical sinology literature, this study formulates standardized primary translations, context-adaptive alternative translations, and detailed cultural connotation annotations for each term.The established database contains 2862 standardized bilingual term pairs, covering 95% of the professional vocabulary commonly used in painting and calligraphy creation spanning from Tang Dynasty to Qing Dynasties.

Cultural Aesthetic Concept Library: This will build up a specialized library containing 327 core abstract aesthetic terms. Each term entry consists of the definition of conceptual connotations, application cases from classical painting treatises, and differentiated translation schemes tailored for academic research, museum exhibition, and cultural popularization scenarios. By embedding this knowledge library into the corpus construction pipeline, the proposed method effectively alleviates semantic biases arising from the translation of abstract aesthetic concepts.

4.2 Multi-Source Data Acquisition and Intelligent Cleaning

Authoritative Data Sourcing: Raw materials are collected from official English profiles released by cultural organizations and museums, authorized translations of classical painting treatises, and English monographs written by overseas sinologists. In total, 63,000 initial text segments are gathered from landscape painting, figure painting, flower-and-bird painting and calligraphy works.

Hybrid Intelligent Cleaning: This study is going to develop a two-stage data cleaning strategy that will combine rule-based filtering with model-driven semantic screening. The rule-based filtering is about to eliminate the duplicate content, garbled characters and incomplete segments, so as to achieve a noise removal rate of 92.6%. The LaBSE cross-lingual embedding model is applied in the model module to compute semantic similarity. With the similarity threshold set at 0.75, high-quality parallel text pairs are selected. Following the entire cleaning pipeline, a total of 51,200 valid parallel segments are preserved, and the final data purity reaches 98.1%.

4.3 Semi-Automated Parallel Alignment

Classical Chinese Text Preprocessing: A domain-adaptive preprocessing module is proposed to handle with the substantial challenges to accurate text parsing caused by the typical elliptical syntax and domain- specific vocabulary of ancient Chinese painting and calligraphy texts. By integrating the specialized painting and calligraphy terminology dictionary into the classical Chinese segmentation model, the entity recognition accuracy is effectively improved from 78.4% to 91.7%. What's more, the designed algorithm standardizes the syntactic structure of colophon poems and elliptical critical texts, and realizes the unified normalization of punctuation marks and textual formats for heterogeneous ancient art texts.

Hybrid Alignment Method: This study adopts the LASER bilingual sentence embedding model to implement preliminary automatic parallel alignment. In this field, a 3×3 convolution kernel and a 128-token semantic window are optimized to adapt to the linguistic

characteristics of classical art texts. The optimized model achieves an automatic alignment accuracy of 93.5% for conventional painting and calligraphy texts. For long classical painting treatises and poetic colophons with complex semantics and flexible syntax, manual expert calibration is further conducted to revise ambiguous and misaligned sentence pairs. The proposed human-machine hybrid alignment strategy ultimately elevates the overall alignment accuracy of the constructed corpus to 97.2%.

4.4 Multi-Modal Feature Fusion

Text-Image Cross-Modal Alignment: Built upon the LLaVA framework, the CLIP visual model is fine-tuned on a data-set consisting of 2000 ancient Chinese paintings. The optimized model notably strengthens the model's ability to capture distinctive visual traits of traditional Chinese art, including ink brush textures and blank layout aesthetics, lifting the overall cross-modal matching accuracy between visual and textual semantics by 14.3%. By mapping textual and visual features into a unified semantic space, the proposed method achieves in-depth semantic alignment between colophon content and local painting details.

Multi-dimensional Metadata Integration: Each sample in the constructed corpus is annotated with rich dimensional metadata, encompassing creation dynasties, artistic styles, school affiliations, collection institutions, and textual genres. All entries are assigned exclusive index codes to facilitate the associative storage of multi-source heterogeneous data, including original classical Chinese texts, English translations, painting images, and matched metadata information.

4.5 Knowledge-Enhanced Annotation Framework

Three-Tier Annotation Standard: Basic layer includes word segmentation, part-of-speech tagging and sentence semantic relationship marking; intermediate layer realizes automatic marking of domain entities based on terminology database; advanced layer completes artificial semantic annotation of artistic conception and cultural communication adaptability [6]. A unified label system with 42 core labels standardizes annotation specification.

Semi-Automated Annotation Tool: The integrated tool supports term intelligent recommendation and cultural concept association. Basic layer adopts crowd-sourcing assistance; intermediate and advanced layers adopt two-person cross-review by art and translation experts. The tool reduces manual workload by 61.8% and improves annotation standardization rate to 95.6% [9].

5. Experimental Results and Analysis

5.1 Corpus Scale and Distribution

The pilot corpus contains 51200 standard parallel sentence pairs covering Tang, Song, Yuan, Ming and Qing dynasties. Landscape painting accounts for 38.2%, figure painting 21.5%, flower-and-bird painting 24.8%, calligraphy and inscriptions 15.5%. Text types include painting treatises 27.3%, colophon poems 31.6%, expert critiques 22.5%, and museum official descriptions 18.6%. (As shown in table 1)

Table 1. Quantitative Quality Evaluation Comparison

Evaluation Index	Proposed Framework	Traditional Automatic Method	Improvement/Reduction Rate
Overall Alignment Accuracy	97.20%	83.50%	13.70%
Terminology Consistency	95.60%	71.20%	24.40%
Data Noise Residual Rate	1.90%	8.70%	-6.8%
Multi-modal Matching Precision	91.30%	76.90%	14.40%

5.2 Quantitative Quality Evaluation

The proposed hybrid framework surpasses conventional fully automatic methods across all evaluation metrics. Especially, embedded knowledge engineering standardizes domain term translation, human-machine collaboration mitigates semantic matching errors, and multi-modal optimization strengthens the semantic correlation between textual and visual content.

5.3 Application Performance Test

With the constructed corpus, the BERT translation model receives a BLEU score of 28.7 in ancient painting and calligraphy text translation, which is 12.5% higher than the model trained on general parallel corpora. In

cross-modal retrieval tasks, the top-5 retrieval accuracy reaches 89.2%. These results show clearly that the constructed corpus can effectively support domain-specific machine translation and multi-modal intelligent retrieval tasks.

6. Discussion and Future Work

6.1 Key Findings

This paper combs out the difficulties in the construction of a Sino-English parallel corpus for ancient Chinese painting and calligraphy, then suggests a holistic technical framework to tackle with these challenges. The adopted human-machine collaborative construction mode achieves an optimal balance between operational efficiency and corpus quality. Meanwhile, the knowledge-enhanced mechanism enables standardized translation of professional domain terms and abstract aesthetic concepts. Apart from its application to the translation of ancient Chinese painting and calligraphy texts, this study offers feasible technical solutions and transferable paradigms for corpus construction in other low-resource cultural and art domains.

6.2 Research Limitations

No doubt this study still has its limitations. The current pilot corpus fails to fully cover niche local painting styles and folk calligraphy works. In addition, the existing multi-modal fusion module only incorporates static text and image data, while excluding audio and video resources. Moreover, the model still struggles to achieve accurate alignment for fragmented inscriptions and ambiguous ancient characters, leaving ample room for further technical improvement.

6.3 Future Research Directions

In the future, the pilot corpus will be expanded into one with 300,000 parallel sentence pairs and much richer data resources for underrepresented niche art schools. It will further integrate audio and video materials to construct a comprehensive all-media multi-modal art corpus. In addition, a lightweight domain-specific model will be set up for intelligent translation and question answering, and an open sharing platform will be established to enable sustainable and dynamic data iteration.

7. Conclusion

Constructing a dedicated Chinese-English parallel corpus for ancient Chinese painting and calligraphy constitutes a fundamental supporting task for the digital preservation of cultural heritage and the deep integration of intelligent computing technologies. To address the core challenges including cultural semantic heterogeneity, data scarcity in low-resource scenarios, parsing complexity in classical texts, multi-modal information isolation, and the absence of standardized annotation specifications, this study establishes a full-process technical framework for corpus construction. The proposed framework effectively overcomes critical bottlenecks by integrating domain knowledge engineering, intelligent data cleaning, human-machine collaborative alignment, multi-modal semantic fusion, and hierarchical knowledge-enhanced annotation. Quantitative experimental results demonstrate its significant advantages in terms of sentence alignment accuracy, terminological consistency, and multi-modal semantic matching precision. The constructed parallel corpus can supply high-quality and standardized data support for domain-specific machine translation, cross-modal information retrieval, and cultural knowledge graph construction. This research expands the theoretical and methodological system of parallel corpus construction in low-resource cultural domains, and offers a replicable paradigm for the cross-lingual transmission of traditional Chinese culture. It also lays a solid data and technical foundation for the digital conservation and global dissemination of traditional Chinese art heritage.

Notes

The corresponding author of this paper presides over the Independent Research Project of Central China Normal University entitled Construction and Application Research of English Translation Corpus for Ancient Chinese Painting and Calligraphy. The massive data, research findings, and future research prospects presented in this paper constitute the interim research achievements of this project.

Acknowledgments

This work was supported by 2024 School Project, Wuhan City Polytechnic under Grant NO. 2024WHCPRB09 (“A Study on the International Comparison of Sino-Australian

Higher Vocational Education and the Path of Inter-University Cooperation").

References

[1] K. Wang, Design and Development of the Chinese-English Parallel Corpus. Journal of Foreign Language Education, 2013, 9(6): 23-27.

[2] L. Niu, A Multi-dimensional Semantic Model for Aligning Chinese and Western Painting and Calligraphy Terms. Journal of Literature and Data, 2026, 12(2): 45-53.

[3] Y. Li, ArtSeek: Deep Artwork Understanding via Multimodal In-Context Reasoning. arXiv preprint arXiv:2502.07891, 2025.

[4] Y. Liu and D. Xiong, Parallel Corpus Construction for Low-Resource Languages. Computer Science, 2022, 49(5): 112-119.

[5] H. Zhang, Multi-modal Corpus Construction for Cultural Heritage. Frontiers in Digital Humanities, 2025, 12: 1-14.

[6] L. Trajkovic, Benchmarking Machine Translation with Cultural Awareness. Findings of EMNLP, 2024: 891-898.

[7] Q. Gu, Corpus Construction for Chinese Art Terminology. Journal of Chinese Culture, 2025, 18(3): 67-75.

[8] Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, et al., No Language Left Behind: Scaling Human- centered Machine Translation. arXiv preprint arXiv:2207.04672, 2022.

[9] ALIGN Team, Word Association Learning for Cultural Alignment in Large Language Models. arXiv preprint arXiv:2508.13421, 2025.

[10] Lulu Zhou, Simone Conia, Daniel Lee, Min Li, Umar Farooq Minhas, Saloni Potdar, Yunyao Li, et al., KG-MT: Integrating Multilingual Knowledge Graphs into Cross-Cultural Translation. EMNLP, 2024: 5678-5686.