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.

Multimodal contrastive learning traces the early dissemination of Romance of the Three Kingdoms in Thailand

Yancong Su
Xiamen University of Technology
Liang Fang
Minnan Normal University
Fei Yu
Minnan Normal University
Wenjie Yang
Minnan Normal University
Ying Lin
Minnan Normal University
Baozhu Lin
Minnan Normal University
Chou Yong Tan
University of Malaya
Shunxing Li
Minnan Normal University
Hongrun Wu
wuhongrun@whu.edu.cn

Minnan Normal University

Article

Keywords:

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11432789/v1

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

Additional Declarations: No competing interests reported.

Multimodal contrastive learning traces the early dissemination of *Romance of the Three Kingdoms* in Thailand

Yancong Su[1,2], Liang Fang[2], Fei Yu[2], Wenjie Yang[2], Ying Lin[2,3*], Baozhu Lin[4], Chou Yong Tan[5], Shunxing Li[2,6], Hongrun Wu[2,5*]

[1]School of Film, Television and Communication, Xiamen University of Technology, 600 Ligong Road, Xiamen, 361024, Fujian, China.
[2]Key Laboratory of Intelligent Optimization and Information Processing, Minnan Normal University, 36 Xianqianzhi Street, Zhangzhou, 363000, Fujian, China.
[3]School of Humanities, Tongji University, 1239 Siping Road, Shanghai, 200092, China.
[4]School of Foreign Studies, Minnan Normal University, 36 Xianqianzhi Street, Zhangzhou, 363000, Fujian, China.
[5]Faculty of Engineering, University of Malaya, Kuala Lumpur, 50603, Malaysia.
[6]Institute of Minnan Culture Research, Minnan Normal University, 36 Xianqianzhi Street, Zhangzhou, 363000, Fujian, China.

*Corresponding author(s). E-mail(s): fjzzlinying@126.com; wuhongrun@whu.edu.cn;

Abstract

The Romance of the Three Kingdoms (*RTK*, 三國演義) is a global literary canon, yet how and when it first reached Thailand remains contested because few documents survive. Using the 1838 Wat Prasert Sutthawat murals, the earliest large-scale Thai *RTK* visual corpus, we build the first expert-verified multimodal dataset for mural–print comparison. We introduce the Mural–Illustration Similarity Prediction (MISP) model, which couples self-supervised multimodal contrastive embedding with supervised cross-modal attention fusion to produce interpretable similarity scores. MISP attains lower predictive error than single-modal and multimodal baselines and agrees closely with independent expert

attributions. The murals most closely resemble the Zhou Yuejiao (周曰校) illustrated edition, identifying likely source editions and constraining when they reached Thailand. Our work extends the evidence for early Thai reception and shows how AI can advance stylistic analysis and provenance inference.

1 Introduction

Understanding how classical narratives traverse linguistic and cultural frontiers and become selectively reconfigured in new cultural ecologies is a central problem for literary history. Among such works, the *Romance of the Three Kingdoms* (*RTK*), a 14th-century historical novel attributed to Luo Guanzhong(羅貫中) with its earliest surviving versions appearing in the 15th century, stands as a cornerstone of Chinese literature and a key vector of cultural exchange globally. This monumental narrative unfolds the turbulent years during the late Eastern Han dynasty (AD 184–220) and the subsequent tripartite division of power among the Wei, Shu, and Wu states, until the reunification of China under the Western Jin dynasty in 280 AD. Its literary influence across the region has been likened to the impact of Shakespeare's works on English literature [1, 2]. While its dissemination across Europe (e.g., France, Britain, Germany) and the Sinographic Sphere (e.g., Japan, Korea, Vietnam) has been substantially documented (scholarship has mapped transmission routes, textual lineages, and performative adaptations of *RTK*), its transmission trajectory into Thailand remains remarkably unclear. Basic critical questions, such as which specific Chinese version arrived and when, lack supporting archival evidence.

Scholars hypothesized that the *RTK* might have been introduced to Thai-Chinese communities in the mid-17th century (late Ming Dynasty), primarily through merchant networks and Chinese immigration[3]. Adverse preservation conditions (tropical humidity, episodic warfare, limited archival consciousness) and the literacy profile of local Thai-Chinese communities have resulted in an extremely sparse corpus of extant *RTK* novels in Thailand. The National Library reportedly holds only thirty-nine Chinese novels from the late Ming-Qing period. The earliest surviving Chinese edition of *RTK* presently preserved in Thailand is the 1883 (the ninth year of the Guangxu reign during the Qing dynasty) *Shanmeitang* (善美堂) issue of the *Xiuxiang Diyi Caizishu*, appearing nearly eight decades after the 1805 Thai translation by Chao Phraya Phra Klang, the first vernacular rendering produced through oral Chinese literacy [4]. Owing to the translator's lack of direct Chinese literacy, this mediation introduced structural pruning, shifts in nomenclature, and narrative compression, severely complicating any attempt to infer the original Chinese source editions from the Thai text alone. In other words, any investigation into the early transmission of *RTK* in Thailand must necessarily rely on non-textual materials.

Among the scarce visual survivals, particularly temple murals have survived as crucial sources for reconstructing the early transmission of *RTK* in Thailand. Specifically, the mural cycle Wat Prasert Sutthawat (known as "Wat Klang") in Bangkok is of singular significance. Unlike the severely deteriorated murals at Wat Nanong and Wat Chetupon, or the numerically limited RTK scenes at Wat Bowonniwet,

Wat Prasert Sutthawat preserves a large, internally coherent corpus of 364 narrative murals. Epigraphic and contextual records date the temple's construction to 1838 (late Qing dynasty) and document the significant involvement of Chinese artisans, implying that the murals predate both the earliest extant Chinese *RTK* edition locally (1883 *Shanmeitang* edition of *Xiuxiang Diyi Caizishu* (繡像第一才子書)) and an illustrated Thai-language adaptation published in 1865. Although Chao Phraya Phra Klang's 1805 Thai translation is three decades earlier, its derivative nature—being based primarily on oral narratives rather than a direct written source—limits its value for tracing the specific Chinese edition used or detailing the novel's precise transmission path. This temporal precedence positions the Wat Prasert Sutthawat murals as potentially the earliest surviving visual witness to *RTK* reception in Thailand and, by extension, as a proxy dataset for establishing a *terminus ante quem*—the date by which the underlying Chinese editions must already have reached Thai contexts.

Despite this significance, prior research on these murals has proceeded mainly through qualitative, case-based description, and has seldom been connected with edition-critical or image-text transmission studies[5–9]. No systematic, large-scale visual comparative analysis has yet been undertaken to align individual mural scenes with candidate Chinese *RTK* novel illustrations. Consequently, questions of source dependency, selective motif retention, and localization (such as iconographic indigenization and spatial recontextualization) remain unresolved. While traditional art-historical analysis and manual iconographic comparison offer detailed qualitative insights, they are time-consuming, labor-intensive, subjective, and difficult to scale across hundreds of *RTK* related images or multiple potential edition illustrations.

Recent advances in artificial intelligence have transformed workflows in cultural heritage, particularly through deep learning (DL) approaches in computer vision, natural language processing, multimodal representation, and knowledge organization. Knowledge graph-enhanced ResNet-34 convolutional neural network has been used to classify Miao batik patterns [10]; graph neural network-based models for sentiment classification of cultural heritage such as nursery rhymes [11]; self-supervised model based on autoencoder (AE) and generative adversarial network (GAN) architectures achieve real-time automatic reconstruction for Longmen Grottoes' ancient stone steles [12]; YOLO object detection model fused with MobileNet backbones accelerate large-scale oracle bone inscriptions identification and recognition [13], thereby facilitating the study and preservation of ancient scripts; a conditional diffusion model designed to conduct three-dimensional reconstructions of damaged cultural artifacts, restoring their original forms and enhancing their representation within digital museum environments [14]. CLIP-based contrastive models in the art domain learn a unified vision–language space enabling classification and retrieval while domain-adapted, style-aware augmentations boost generalisation and fine-grained zero-shot transfer[15, 16]. Collectively, these studies demonstrate the methodological versatility of deep learning models for pattern extraction, semantic enrichment, and virtual restoration within the cultural heritage domain, and they underscore the technology's transformative contribution to the data-driven preservation, interpretation, and reuse of traditional cultural resources.

Inspired by the above mentioned cultural-heritage computational models, this paper aims to study the trace and characterization of the early dissemination of *RTK* in Thailand through calculated visual–semantic similarity relationships between the Wat Prasert Sutthawat murals and candidate Chinese *RTK* novel illustration editions. The core challenges inherent in this task include: extracting robust and rich features from a low-resource mural corpus with constrained manual effort; aligning these mural representations with a historically curated reference set of Chinese illustrated editions predating the construction of Wat Prasert Sutthawat so as to enable temporally plausible source attribution; and producing ranked candidate source linkages that can be subjected to critical evaluation by humanities scholars. To address these challenges, we construct a corpus of 727 images: 88 selected *RTK*–themed mural scenes from Wat Prasert Sutthawat and 639 novel illustrations drawn from 16 representative Chinese editions pre-construction of Wat Prasert Sutthawat. Both the mural panels and the printed illustrations contain embedded Chinese textual material (e.g., Chinese couplet, dramatis personae labels, and topic descriptors).

To jointly harness the dataset's visual features and in-image textual descriptors, we propose a Mural–Illustration Similarity Prediction (abbreviated as MISP) model with two main components: (i) a self-supervised Multimodal Contrastive Learning (MCL) module that derives rich iconographic-semantic embeddings via both intra-modal (image–image, text–text) and inter-modal (image–text, text-image) contrastive objectives; and (ii) a supervised Cross-Modal Multi-Head Attention Fusion (CMAF) based ranking module that fuses aligned embeddings through a multi-head cross-attention mechanism to produce calibrated mural-illustration similarity scores. The effectiveness of our proposed model is validated against expert-annotated mural-illustration pairings, and comparative experiments against single-modal contrastive baselines and contemporary multimodal contrastive architectures show consistent reductions in predictive error across all data splits. These results underscore the proposed model's potential to advance domain-adaptive contrastive learning for cross-source cultural heritage image similarity analysis. More broadly, by making quantifiable the iconographic transformations accompanying transregional literary transmission, this paper offers a generalizable methodological template for historical attribution, reception studies, and conservation prioritization in other multi-version narrative corpora.

This study bridges traditional comparative iconography and AI-based pattern discovery, proceeding from the hypothesis that systematically extracting and aligning visual and textual features can expose differential affinities between mural scenes and specific Chinese editions. The study makes three contributions. First, we assemble the first expert-verified *multimodal* dataset pairing Thai mural scenes with curated Chinese *RTK* illustrations, adding the in-image textual descriptors that single-modal image matching ignores. Second, we propose MISP, a hybrid pipeline coupling self-supervised multimodal contrastive learning with supervised cross-modal multi-head attention fusion to generate similarity scores and infer probable edition-family attributions, and we corroborate the resulting attributions with non-parametric bootstrap resampling. Third, we show that the model's transparent similarity rationales augment, rather than replace, expert visual analysis, yielding reusable digital-heritage infrastructure that is transferable to analogous source-attribution problems.

2 Methods

2.1 Research Data

We created a multimodal corpus of 727 images comprising 88 selected *RTK* mural panels from Wat Prasert Sutthawat and 639 illustrations drawn from 16 representative historical Chinese illustrated novel editions of the *RTK*, all published prior to the construction of Wat Prasert Sutthawat. From the temple's 364 *RTK*-themed murals we retained the 88 panels whose narrative episodes can be unambiguously identified and matched to counterpart illustrations across the candidate editions and which preserve legible in-image text, so that every panel supports a well-posed cross-modal comparison. Most of these mural panels are accompanied by in-image textual annotations (e.g., character names and topic headings), while they likewise include paratextual elements such as Chinese poetic couplets, captions, and topical descriptors. This curated dataset thus forms a multimodal corpus of visual and semantic richness. The number of illustrations of the novel collected from each *RTK* edition varies (Table 1). IDs 1–4 correspond to Jiangnan-style editions, whose illustration repertoire features double-page continuous spreads and half-page vertical formats; IDs 5–14 are in the Jianyang-style, which favors a top-image and bottom-text layout; 15–16 display a Jiangnan–Jianyang hybrid style. "Jiangnan"(江南) denotes the lower Yangtze (Yangzi) cultural region; "Jianyang"(建陽) means the publishing center in northern Fujian. Representative examples of both murals and novel illustrations are shown in Figure 1.

To evaluate the performance of our proposed model in predicting the similarity between mural depictions and novel illustrations of the same *RTK* narrative episodes, we conducted an expert annotation study involving five domain specialists with backgrounds in literature and art history. Each mural–illustration pair was assessed according to an overall coherence criterion, defined as an integrative judgment of whether the combined visual and textual evidence in a given pair consistently conveys the same narrative situation and thematic interpretation. This annotation protocol enables fine-grained evaluation of cross-modal alignment across heterogeneous visual and textual signals. Each expert independently assigned a binary integer score — 0 (indicating no meaningful correspondence) or 1 (indicating high correspondence) — and the final similarity score for each pair was computed as the arithmetic mean of the five expert ratings.

Table 1: Edition Information for Illustrated Novels

Serial ID	Novel Illustration Edition	Edition Information	Abbreviation	Quantity (pieces)
1	Zhou Yuejiao (周曰校) Edition	This study uses the copy of *Fully Illustrated Romance of the Three Kingdoms* (全像三國志傳演義) produced at the Jinling Wanjuanlou studio in the 21st year of the Wanli reign (1593), collated by Zhou Yuejiao, preserved in the Naikaku Bunko (Cabinet Library), Japan.	Zhou (周)	72
2	Li Zhuowu (李卓吾) Edition	This study uses the Nanjing Library copy of the late-Ming Wu Guanming edition of *Li Zhuowu's Annotated Romance of the Three Kingdoms* (李卓吾先生批評三國志), accessed via the facsimile in Chen Xianghua's *Continuation Series of the Collection of Ancient Editions of the Romance of the Three Kingdoms* in vols. 11–12.	Li (李)	72
3	Li Liweng (李笠翁) Edition	This study uses the National Library of China copy of the early-Qing edition *Li Liweng's Annotated Romance of the Three Kingdoms* (李笠翁批閱三國志).	Weng(翁)	69
4	Yixiangtang (遺香堂) Edition	This study uses the Library of Congress copy of the Qing-dynasty Yixiangtang studio *Illustrated Romance of the Three Kingdoms* (繪像三國志).	Yi (遺)	15

Serial ID	Novel Illustration Book	Edition Information	Abbreviation	Quantity (pieces)
5	Ye Fengchun (葉逢春) Edition	This study uses the 1548 (27th year of the Jiajing reign of the Ming dynasty) Jianyang Ye Fengchun edition *Newly Printed Popular Romance of the Three Kingdoms* (新刊通俗演義三國史傳), produced within the Jianyang printing tradition and consulted via the facsimile in Chen Xianghua's Continuation Series of *The Collection of Ancient Editions of the Romance of the Three Kingdoms* (三國志演義古版叢刊續輯).	Ye (葉)	50
6	Yu Xiangdou (余象斗) Edition	This study uses the 1592 (20th year of the Wanli region of Ming dynasty) Shuangfengtang studio edition *The Annotated and Illustrated Romance of the Three Kingdoms* (音釋補遺按鑒演義全像批評三國志傳), printed by Yu Xiangdou and consulted via the facsimile in the Series of Ancient Editions of Fiction.	Yu (余)	56
7	Zheng Shaoyuan (鄭少垣) Edition	This study uses the 1605 (Wanli 33, Ming dynasty) edition titled *Newly Printed the Popular Romance of the Three Kingdoms, Collated by the Capital Standard* (新鍥京本校正通俗演義按鑒三國志傳), printed by Zheng Shaoyuan (鄭少垣) of Lianhuitang studio, consulted via the facsimile in the *Series of Ancient Editions of Fiction* (古本小說叢刊).	Zheng (鄭)	65

Serial ID	Novel Illustration Book	Edition Information	Abbreviation	Quantity (pieces)
8	Zhongdetang (种德堂) Edition (Incomplete)	This study uses the edition titled *New Illustrated Edition of the Romance of the Three Kingdoms, Collated by the Capital Standard* (新鐫京本校正按鑒演義全像三國志傳), printed by Xiong Chongyu (熊冲宇), consulted from the National Library of China exemplar.	De(德)	6
9	Zhu Dingchen (朱鼎臣) Edition	Compiled by Zhu Dingchen and printed by Wang Siyuan: *Newly Printed and Critically Annotated Edition of the Romance of the Three Kingdoms* (新刻音釋旁訓評林演義三國志史傳),consulted via the facsimile in Chen Xianghua's*Five Ancient Editions of the Romance of the Three Kingdoms*(三國志演義古版叢刊五種).	Zhu(朱)	61
10	Liu Longtian (劉龍田) Edition	This study uses the Wanli reign-period edition *Newly Printed Fully Illustrated Large-Character Popular Romance of the Three Kingdoms*(新鋟全像大字通俗演義三國志傳), consulted via the University of California, Berkeley Library copy.	Liu(劉)	66
11	Liu Rongwu (劉榮吾) Edition(Incomplete)	This study uses the Liguangtang studio edition printed by Liu Rongwu, titled *The Exquisitely Printed and Illustrated Romance of the Three Kingdoms* (精鐫按鑒全像鼎峙三國志傳), and the facsimile consulted is in the Series of *Ancient Editions of Fiction* .	Rong(荣)	62

Serial ID	Novel Illustration Book	Edition Information	Abbreviation	Quantity (pieces)
12	The Wei (魏氏) Edition(Incomplete)	This study uses the Wei family's edition titled *Second Edition of the Fully Illustrated Romance of the Three Kingdoms: A Saga of Heroes* (二刻按鑒演義全像三國英雄志傳), consulted via the National Library of China copy.	Wei(魏)	8
13	Beijing (北京) Library (which is now known as the National Library of China) Edition(Incomplete)	This study uses the edition titled *Newly Printed Fully Illustrated Romance of the Three Kingdoms* (新刻全像演義三國志傳), consulted via the National Library of China copy. Following prior practice, this edition is referred to as the "Beijing pictorial edition."	Bei(北)	11
14	Chengdetang (誠德堂) Edition	This study uses the edition printed by Xiong Qingbo, titled *A New Expanded Edition of the Popular Complete Romance of the Three Kingdoms, Printed by the Capital Standard* (新刻京本補遺通俗演義三國全傳), in 1596 (Wanli 24, Ming dynasty), produced by the Jianyang Chengdetang studio, consulted via the Tohoku University Library copy (Japan).	Cheng(誠)	4

Serial ID	Novel Illustration Book	Edition Information	Abbreviation	Quantity (pieces)
15	Zhongzhengtang (忠正堂) Edition	This study uses the edition printed by Xiong Fogui, titled *Newly Printed Annotated Romance of the Three Kingdoms: A Combined Edition with Phonetics and Commentary* (新鍥音釋評林演義合刻三國志傳), produced by the Jianyang Zhongzhengtang studio in 1603 (Wanli 31, Ming dynasty), and consulted via the facsimile in Chen Xianghua's Continuation Series of *The Collection of Ancient Editions of the Romance of the Three Kingdoms*.	Zhong(忠)	20
16	Kangxi Fragmented (康熙殘葉) Edition(Incomplete)	This study uses illustrations from the Kangxi-period (Qing dynasty) reprint of the Ming Jianyang edition *Newly Printed, Fully Illustrated Romance of the Three Kingdoms* (新刻全像演義三國志傳), consulted the facsimile in Fu Xihua's *Selection of Woodblock Illustrations from Classical Chinese Literature* (中國古典文學版畫選集).	Kang(康)	2

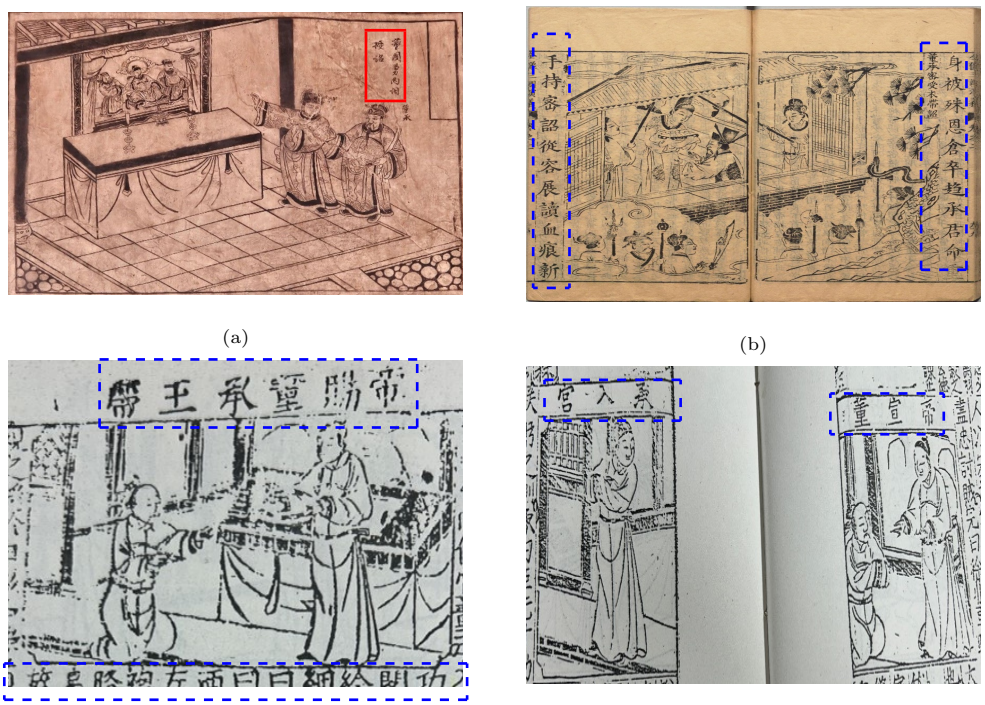


Fig. 1 Mural and illustrations from the "Zhou", "Zhu", and "Zhong" editions of *Romance of the Three Kingdoms* depicting "Dong Cheng Secretly Receiving the Belt Edict." The red solid box in (a) marks the text region in the mural; blue dashed boxes in (b)–(d) mark the corresponding text regions in the novel illustrations.

2.2 Research Framework

In the field of self-supervised learning, contrastive learning has proven to be a powerful method for extracting meaningful data representations without the need for labeled samples. Its primary strength lies in its capacity to capture subtle similarities and distinctions among samples by constructing positive and negative pairs and optimizing a contrastive loss function. This makes it especially suitable for tasks where manual annotation is both expensive and time-consuming. Building on these advantages, we propose a framework incorporating Multimodal Contrastive Learning (MCL) with a Cross-Modal Multi-Head Attention Fusion (CMAF) mechanism to jointly extract visual and textual features from both murals and illustrations, thereby predicting mural–illustration similarity. The overall research framework, as shown in Figure 2, consists of data preprocessing, multimodal contrastive learning, cross-modal multi-head attention fusion based mural-illustration similarity prediction, and model performance evaluation.

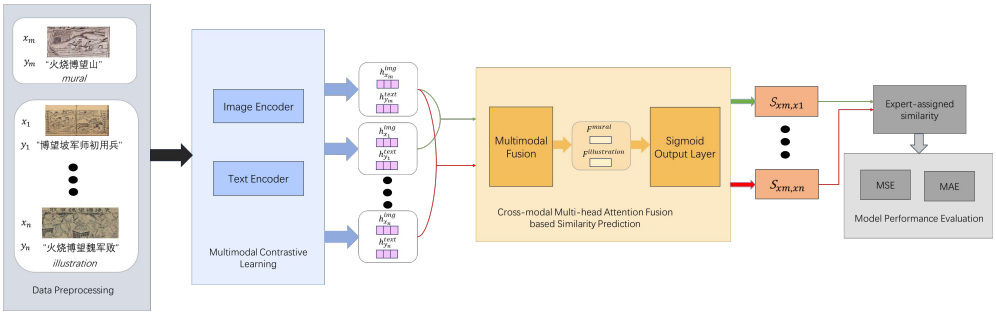


Fig. 2 Overview of the proposed Mural–Illustration Similarity Prediction (MISP) framework, comprising four stages: data preprocessing, multimodal contrastive learning, cross-modal multi-head attention fusion for similarity prediction, and model-performance evaluation against expert annotations.

2.2.1 Data Preprocessing

In the preprocessing stage, mural and illustration images are rescaled to meet the model's input dimension requirements, followed by normalization and standardization to ensure consistency and stability across samples. Text materials extracted from murals and novel illustrations undergo OCR transcription and manual proofreading.

2.2.2 Multimodal Contrastive Learning

The preprocessed mural images and their textual descriptions are fed into a trained multimodal contrastive learning (MCL) component, generating embeddings that are deeply semantically rich and highly compositional[17, 18] . The MCL powerfully leverages contrastive learning across diverse modalities to align visual and textual data within a shared embedding space, ensuring that semantically corresponding instances drawn from heterogeneous modalities are closely aligned, whereas non-corresponding instances remain distinctly separated [19, 20]. This process fundamentally enhances representational quality and enables robust zero-shot capabilities for complex tasks such as fine-grained attribute recognition in artwork. The core architectures and training details of the MCL are further demonstrated in Section 2.3.

2.2.3 Cross-modal Multi-head Attention Fusion based Similarity Prediction

We introduce a dual-branch cross-modal multi-head attention fusion (CMAF) mechanism to integrate extracted features from mural images, novel illustrations, and their aligned textual narratives via MCL into a unified semantic space. This design addresses pronounced cross-domain discrepancies—spanning visual style, compositional structure, and narrative granularity—between mural imagery and novel illustration, thereby improving the fidelity of downstream semantic alignment. In the mural branch, the mural image embedding h^{img} (derived by MCL in section 2.3) acts as the query and attends over its corresponding aligned textual embedding h^{text}, producing a fused

mural representation F^{mural} that encodes image-anchored semantic grounding. In the novel illustration branch, the textual embedding queries the corresponding illustration image features to generate fused representations $F^{illustration}$, enabling concept-driven linguistic guidance to highlight salient visual regions and decrease stylistic noise. The resulting fused branch representations are subsequently concatenated to form a joint multimodal vector. This joint vector is then passed to a lightweight multilayer perceptron with nonlinear activation and a final normalization layer, from which a bounded similarity value S_{x_m,x_n} between a pair of mural and novel illustration (abbreviated as mural–illustration pair (x_m, x_n)) is obtained.

2.2.4 Model Performance Evaluation

To evaluate the mural–illustration similarity prediction, we adopt two widely used error metrics, Mean Squared Error (MSE) and Mean Absolute Error (MAE), which quantify the agreement between the model's predicted similarity scores and the expert-annotated ground truth and thereby reflect the model's alignment with human judgment.

Mean Squared Error (MSE): Mean Squared Error is particularly sensitive to large discrepancies, making it a valuable indicator for identifying substantial misalignments between model predictions and expert judgments. The MSE is calculated as follows,

$$\text{MSE} = \frac{1}{M} \sum_{(x_m,x_n)\in\mathscr{M}} (G_{x_m,x_n} - S_{x_m,x_n})^2 \tag{1}$$

where $\mathscr{M}$ is the set of all mural–illustration pairs and $|\mathscr{M}| = M$. G_{x_m,x_n} is the expert-assigned similarity for mural x_m and novel illustration x_n, and S_{x_m,x_n} is our model predicted similarity score. Lower MSE values indicate closer agreement with expert judgments.

Mean Absolute Error (MAE): To complement MSE, we also employ MAE, which is less sensitive to outliers and thus provides a more robust evaluation of overall model accuracy. A lower MAE reflects a closer overall correspondence between the model's outputs and expert assessments, especially in the presence of annotation noise or subjectivity. It is defined as:

$$\text{MAE} = \frac{1}{M} \sum_{(x_m,x_n)\in\mathscr{M}} |G_{x_m,x_n} - S_{x_m,x_n}| \tag{2}$$

2.3 Training of Multimodal Contrastive Learning Model

The designed multimodal contrastive learning model in this study leverages both inter-modal and intra-modal self-supervision to provide complementary benefits in representation learning, the key idea of which aims to facilitate modeling better joint multimodal features in the fusion encoder. As illustrated in Figure 3, the training procedure comprises the following stages.

2.3.1 Data Augmentation

To enhance the MCL's generalization capability and robustness to input variability, the data augmentation transforms any given data example randomly resulting in two correlated views of the same example, which we consider as a positive pair. For the visual modality, two separate data augmentation operators are sampled from the same family of augmentations (such as random cropping, color distortion, noise introduction, and horizontal flipping etc.) and applied to an image example x to obtain two correlated views x_i and x_j, which is also called a positive pair. Following [21], all other augmented views in the current mini-batch that do not originate from image x are treated as negative samples. For the textual modality, with given textual example y, we generate two augmentations y_i and y_j via a principled combination of word-level manipulations, such as random token deletion and context-aware synonym substitution. This augmentation protocol encourages the model to learn invariant, robust representations across diverse textual surface forms that preserve the same core semantics [22].

2.3.2 Feature Extraction

During feature extraction, a dual-encoder architecture independently maps the augmented image and text inputs into their respective feature spaces. For the image modality, EfficientNet serves as the visual encoder f_{img}, mapping the augmented input x_i into a high-dimensional feature space. Leveraging compound scaling, this encoder captures both local and global visual characteristics, yielding the intermediate representation defined in Equation 3 [23]. Subsequently, this representation is passed through a two-layer multilayer perceptron (MLP) projection head g_{img} with ReLU activations, which maps it into a lower-dimensional embedding space to produce the image embedding $z_{x_j}^{\text{img}}$ (Equation 4) used for the contrastive loss. Similarly, we obtain $h_{x_j}^{\text{img}}$ and $z_{x_j}^{\text{img}}$ for the other augmentation x_j using Equation 3 and 4. After contrastive pretraining, we discard the projection head $g_{\text{img}}(\cdot)$ and retain the encoder $f_{\text{img}}(\cdot)$ and its output representation $h_{x_i}^{\text{img}}$ and $h_{x_j}^{\text{img}}$ for downstream tasks.

$$h_{x_i}^{\text{img}} = f_{\text{img}}(x_i) \tag{3}$$

$$z_{x_i}^{\text{img}} = g_{\text{img}}(h_{x_i}^{\text{img}}) \tag{4}$$

For the text modality, a SimCSE encoder built upon the BERT architecture, denoted f_{text}, encodes one augmented textual input y_i [24]. A sentence-level pooling operation produces the intermediate representation $h_{y_i}^{\text{text}}$, which is projected into a lower-dimensional space via a linear projection head g_{text}, yielding the projected text embedding $z_{y_i}^{\text{text}}$.

$$h_{y_i}^{\text{text}} = f_{\text{text}}(y_i) \tag{5}$$

$$z_{y_i}^{\text{text}} = g_{\text{text}}(h_{y_i}^{\text{text}}) \tag{6}$$

Finally, The representations $z_{x_i}^{\text{img}}$, $z_{x_j}^{\text{img}}$, $z_{y_i}^{\text{text}}$, and $z_{y_i}^{\text{text}}$ then serve as the basis for subsequent intra-modal and inter-modal contrastive optimization.

2.3.3 Contrastive Loss

The alignment module of the MCL explicitly incorporates both inter-modal and intra-modal contrastive objectives, consistent with recent advancements [25, 26]. We next detail each objective and highlight their complementary contributions to representation alignment and robust multimodal fusion.

Inter-Modal Contrastive Loss: The inter-modal alignment loss, often referred as cross-modal contrastive loss, is designed to align representations of corresponding samples from different modalities in a shared embedding space [26, 27]. Its primary objective is to bring representations of matched image-text pairs closer while pushing those of unmatched pairs farther apart. Following common practice [25], the contrastive objective typically utilizes a cross-entropy formulation (InfoNCE). To facilitate cross-modal semantic alignment, we define the bidirectional contrastive objective between the image and text modalities. For a mini-batch of N image–text pairs, the total inter-modal loss $\mathcal{L}_{\text{inter}}$ is defined in Eq. 7 and comprises the image-to-text loss $\mathcal{L}_{\text{image-text}}$ and the text-to-image loss $\mathcal{L}_{\text{text-image}}$.

$$\mathcal{L}_{inter} = \frac{1}{2}[\mathcal{L}_{image-text} + \mathcal{L}_{text-image}] \tag{7}$$

The image-to-text loss is given in Equation 8, where $z_{x_i}^{\text{img}}$ and $z_{y_i}^{\text{text}}$ denote the embeddings of the i-th postive image-text pair, respectively. The operator $sim(,)$ computes similarity scores between features of example pairs. In Equation 8, $\mathscr{N}$ denotes the index set of positive image-text samples, and $\widehat{\mathscr{N}^{(i)}}$ denotes the corresponding negative indices (i.e., not paired with the positive image-text sample (x_i, y_i)). The temperature $\tau > 0$ controls the sharpness of the softmax, amplifying the contribution of hard negatives. Similarly, the text-to-image loss is defined in Equation 9.

$$\mathcal{L}_{image-text} = -\frac{1}{N}\sum_{i\in\mathscr{P}} \log \frac{\exp(\text{sim}(z_{x_i}^{\text{img}}, z_{y_i}^{\text{text}})/\tau)}{\exp(\text{sim}(z_{x_i}^{\text{img}}, z_{y_i}^{\text{text}})/\tau) + \sum_{k\in\widehat{\mathscr{N}^{(i)}}} \exp(\text{sim}(z_{x_i}^{\text{img}}, z_{y_k}^{\text{text}})/\tau)} \tag{8}$$

$$\mathcal{L}_{text-image} = -\frac{1}{N}\sum_{i\in\mathscr{Q}} \log \frac{\exp(\text{sim}(z_{y_i}^{\text{text}}, z_{x_i}^{\text{img}})/\tau)}{\exp(\text{sim}(z_{y_i}^{\text{text}}, z_{x_i}^{\text{img}})/\tau) + \sum_{k\in\widehat{\mathscr{Q}^{(i)}}} \exp(\text{sim}(z_{y_i}^{\text{text}}, z_{x_k}^{\text{img}})/\tau)} \tag{9}$$

Intra-modal Contrastive Loss: Intra-modal contrastive learning seeks to discriminate positive pairs from negatives within the same modality, reinforcing modality-specific invariances. Therefore, the intra-modal contrastive loss $\mathcal{L}_{Intra}$, defined as Eq.10, comprises the visual loss $\mathcal{L}_{image-image}$ and texutal loss $\mathcal{L}_{text-text}$.

$$\mathcal{L}_{intra} = \frac{1}{2}[\mathcal{L}_{image-image} + \mathcal{L}_{text-text}] \tag{10}$$

The InfoNCE objective for the visual intra-modal contrastive loss is defined in Eq. 11, where $z_{x_i}^{\text{img}}$ and $z_{x_j}^{\text{img}}$ are the image representations produced by the image encoder described in Section 2.3.2. Similarly, the contrastive loss for the textual modality is

given in Eq. 12, where the feature vectors z_i^{text} and z_j^{text} are generated by the textual encoder in Section 2.3.2. In the contrastive losses of Equations 11 and 12, $\mathscr{X}$ and $\mathscr{Y}$ denote the sets of image and text positives paired with anchor i, while the negative pools $\widehat{\mathscr{X}(i)}$ and $\widehat{\mathscr{Y}(i)}$ comprise all remaining in-batch indices.

$$\mathcal{L}_{image-image} = -\frac{1}{N} \sum_{(x_i,x_j)\in\mathscr{X}} \log \frac{\exp(\text{sim}(z_{x_i}^{\text{img}}, z_{x_j}^{\text{img}})/\tau)}{\exp(\text{sim}(z_{x_i}^{\text{img}}, z_{x_j}^{\text{img}})/\tau) + \sum_{k\in\widehat{\mathscr{X}(i)}} \exp(\text{sim}(z_{x_i}^{\text{img}}, z_{x_k}^{\text{img}})/\tau)} \tag{11}$$

$$\mathcal{L}_{text-text} = -\frac{1}{N} \sum_{(y_i,y_j)\in\mathscr{Y}} \log \frac{\exp(\text{sim}(z_{y_i}^{\text{text}}, z_{y_j}^{\text{text}})/\tau)}{\exp(\text{sim}(z_{y_i}^{\text{text}}, z_{y_j}^{\text{text}})/\tau) + \sum_{k\in\widehat{\mathscr{Y}(i)}} \exp(\text{sim}(z_{y_i}^{\text{text}}, z_{y_k}^{\text{text}})/\tau)} \tag{12}$$

Finally, the total loss function integrates both components: the inter-modal contrastive loss L_{inter}, the intra-modal loss L_{intra}. The unified training objective is expressed as:

$$\mathcal{L}_{\text{total}} = \lambda\mathcal{L}_{\text{inter}} + (1-\lambda)\mathcal{L}_{\text{intra}} \tag{13}$$

where $\lambda \in [0,1]$ is a weighting hyperparameter that scales the inter-modal contrastive loss, with $1-\lambda$ applied to the intra-modal contrastive loss.

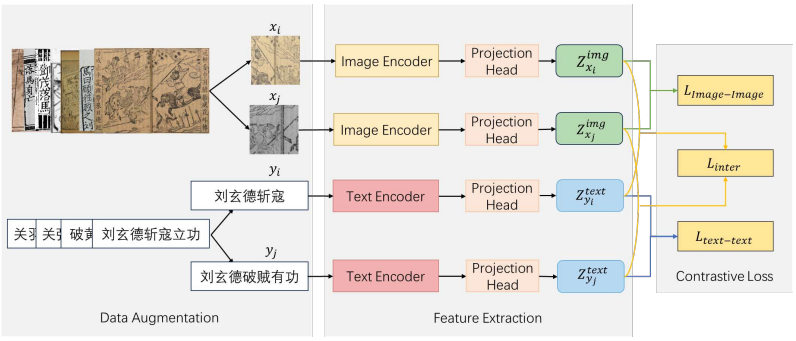


Fig. 3 Pretraining pipeline of the multimodal contrastive learning module. Augmented image and text views are mapped by modality-specific encoders and projection heads into a shared space and optimized with combined inter-modal (image–text) and intra-modal (image–image, text–text) contrastive objectives.

3 Results

We first report quantitative evaluations of the proposed model against strong unimodal and multimodal baselines in Section 3.1, then apply the trained framework to analyze visual similarity patterns between the mural corpus and multiple illustrated editions of *RTK* in section 3.2.

3.1 Model Evaluation

3.1.1 Experimental Setup

All experiments are conducted using the PyTorch deep learning framework and executed on an NVIDIA RTX 4060 GPU. The software environment is configured with Python 3.8 and CUDA 11.6. During the pretraining phase, the Adam optimizer is employed with an initial learning rate of 3×10^{-5}. The batch size is fixed at 16, and input images are uniformly resized to a resolution of 224×224. The model is trained for 30 epochs. A temperature coefficient of 0.07 is set to scale the contrastive loss. For the fine-tuning phase in CMAF module, with the learning rate reduced to 1×10^{-5} to facilitate more stable convergence. The batch size and input image resolution remain consistent with the pretraining configuration. Fine-tuning is performed for an additional 20 epochs. The dataset was split into two equal halves: one used for model evaluation (Section 3.1.3) and the other for automated similarity analysis between murals and illustrated editions (Section 3.2). Of the expert assigned mural-illustration similarity scores, 63.68% fell in the 0–0.50 range, 15.08% in 0.50–0.60, 10.61% in 0.60–0.70, 5.58% in 0.70–0.80, 4.47% in 0.80–0.90, and 0.56% in 0.90–1.00.

3.1.2 Baseline Models

Our proposed approach is rigorously benchmarked against five representative baselines, carefully selected to span two critical categories: unimodal contrastive self-supervision (SimCLR [28], MoCo [29]) and multimodal representation learning models (BLIP [30], CoCa [31], ALBEF [32]). These baselines collectively span the dominant design axes of contrastive learning, including instance discrimination versus cross-modal alignment, pure contrastive versus hybrid generative objectives, and queue/momentum mechanisms versus dual-encoder or encoder–decoder fusion.

SimCLR: It employs a ResNet backbone followed by a two-layer MLP projection head and maximizes agreement between two stochastic augmentations of the same image via an NT-Xent contrastive loss with in-batch negatives. Its strength lies in leveraging large effective negative sets without momentum encoders or memory banks. In our experiments, we extract the image embeddings of murals and illustrations from the projection head, concatenate them, pass the result through a multilayer perceptron, and normalize it to obtain the mural-illustration similarity score.

MoCo: It introduces a momentum-updated encoder and a first-in–first-out queue to maintain a large and consistent pool of negative keys, stabilizing contrastive optimization under smaller batch sizes. In this paper, MoCo is employed to extract features from murals and illustrations. The resulting feature vectors are concatenated and passed through a linear transformation layer, followed by normalization, to compute the final similarity score.

BLIP: It integrates dual unimodal encoders with a multimodal Transformer and leverages an interplay of image–text contrastive, image–text matching, and captioning objectives, where bootstrapped captions mitigate dataset noise. In our experiments, multimodal embeddings are first obtained via dual unimodal encoders, after which a multimodal Transformer integrates visual features into the textual embeddings.

The resulting fused representations are then concatenated and fed into a multilayer perceptron with a normalization layer to compute the similarity score.

CoCa: It adopts a dual-encoder plus decoder (encoder–decoder) architecture trained jointly with contrastive loss (for alignment) and a generative captioning loss (for semantic enrichment). Its decoder-based generative supervision encourages richer alignment beyond surface-level co-occurrence. In our experiments, we employ CoCa's dual encoders to obtain image and text embeddings. The normalized embeddings are concatenated and fed into a multilayer perceptron followed by a normalization layer to compute the similarity score.

ALBEF: This model follows an "align-then-fuse" paradigm, i.e., an initial dual-encoder contrastive stage aligns modalities before a multimodal Transformer refines fused representations and masked language modeling. A momentum distillation mechanism improves sample efficiency. In our experiments, we obtain aligned dual-encoder embeddings using ALBEF and fuse the visual and textual features via cross-modal attention to generate multimodal embeddings. The resulting embeddings of murals and illustrations are then concatenated and passed through a normalized multilayer perceptron to compute a bounded similarity score.

3.1.3 Model Performance Analysis

We rigorously evaluate the proposed Mural–Illustration Similarity Prediction (MISP) method using mean absolute error (MAE) and mean squared error (MSE), comparing it against five established encoder baselines: two self-supervised unimodal encoders (SimCLR and MoCo) and three cross-modal frameworks (BLIP, CoCa, and ALBEF). Lower MAE and MSE indicate stronger multimodal semantic alignment and closer agreement between the model's predicted similarity scores and expert annotations of mural–illustration similarity. We report MAE and MSE across three training/validation/test splits—8:1:1, 7:1:2, and 6:2:2 —to assess both predictive accuracy and robustness under reduced training data. These complementary metrics jointly reflect average deviation (MAE) and sensitivity to larger residuals (MSE).

As shown in Table 2, our proposed model MISP attains the best performance across all data partitions, achieving the lowest MSE of 0.0419 under the 8:1:1 split and the lowest MAE of 0.1741 under the 6:2:2 split. While unimodal methods demonstrate stable behavior, their deficiency in textual semantic modeling inherently limits their upper-bound performance in cross-medium semantic alignment. The MSE of SimCLR and MoCo ranges from 0.0475 to 0.0528 and their MAE from 0.1813 to 0.1954. Cross-modal baselines show marked divergence. BLIP underperforms with MSE 0.0729–0.0781 and MAE 0.2393–0.2479, highlighting challenges in modeling fine-grained pairwise semantics. CoCa achieves the strongest performance among the baselines but shows pronounced sensitivity to training data size. It attains its lowest MAE and MSE under the 8:1:1 split, yet degrades sharply at 6:2:2 with reduced training data, with MSE rising by 63.2% and MAE by 28.2%. This volatility contrasts with the smoother performance trajectory of our proposed MISP. ALBEF maintains moderate accuracy and relative stability, but is consistently outperformed by MISP in MSE and MAE by 7.2–18.8% and 8.9–12.4%, respectively, across different splits. Overall, our model not only leads on all data splits but also shows minimal drift (max

relative MSE increase 10.5% vs. CoCa 76.6%), highlighting superior robustness to the size of training dataset.

These empirical findings strongly support the effectiveness of our proposed MISP for multimodal semantic alignment and further demonstrate its suitability for evaluating similarity between murals and different novel illustrations to identify the specific novel edition referenced in mural creation.

Table 2 Impact of Encoder Architecture on Transfer Performance (MSE and MAE) across Data Partition Strategies

Model	**8:1:1**		**7:1:2**		**6:2:2**	
	MSE	MAE	MSE	MAE	MSE	MAE
SimCLR	0.0501	0.1878	0.0475	0.1870	0.0495	0.1813
MoCo	0.0528	0.1954	0.0504	0.1905	0.0518	0.1866
BLIP	0.0742	0.2393	0.0781	0.2479	0.0729	0.2422
CoCa	0.0424	0.1822	0.0749	0.2327	0.0692	0.2336
ALBEF	0.0498	0.1971	0.0550	0.2010	0.0475	0.1940
MISP	**0.0419**	**0.1810**	**0.0463**	**0.1788**	**0.0443**	**0.1741**

3.1.4 Ablation Study on Loss Weighting

To assess the respective contributions of the inter-modal (cross-modal) and the intra-modal (image–image, text–text) contrastive learning in our proposed model MISP, we conducted a series of experiments by varying the weighting parameter $\lambda \in [0.0, 1.0]$. This schedule traces a continuum from purely intra-modal contrastive learning at $\lambda = 0$ to purely inter-modal contrastive learning at $\lambda = 1.0$.

As shown in Figure 4, the error curves follow a clear non-linear pattern with respect to the inter-modal weight λ. When λ increases from 0 to 0.6 the model demonstrates consistent performance improvement, reflected in a steady decrease in both MAE and MSE. This suggests that introducing a moderate degree of inter-modality alignment improves the MISP's prediction accuracy. The peak performance is achieved at $\lambda = 0.6$, where the lowest MAE and MSE are recorded. However, further increasing the inter-modal weight degrades performance, and assigning full weight to the inter-modal term ($\lambda = 1$) yields the largest errors. This implies that overemphasis on inter-modal consistency may severely hinder the model's capacity to capture essential intra-modal structural patterns and semantic granularity. Overall, an appropriate balance between inter-modal and intra-modal contrastive objectives allows the MISP to leverage both shared cross-modal structure and modality-specific representations.

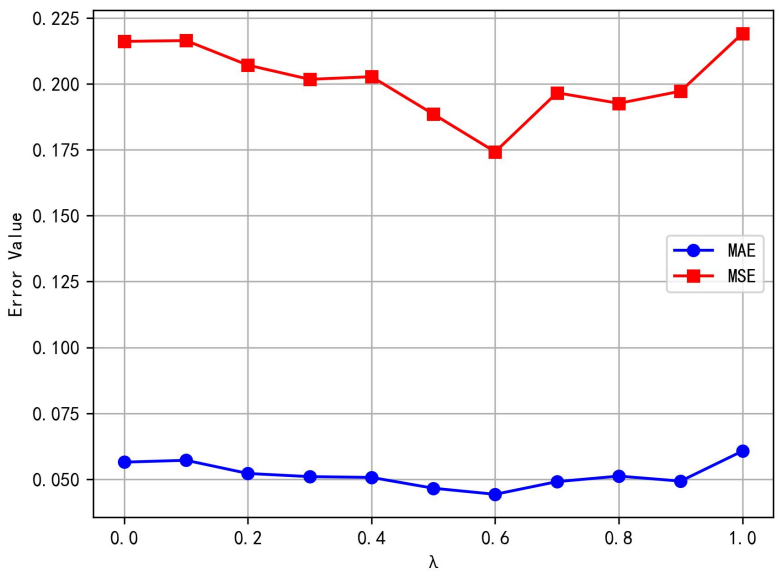


Fig. 4 Error variation under different inter- and intra-modality loss ratios.

3.2 Similarity Analysis Between Murals and Illustrated Editions

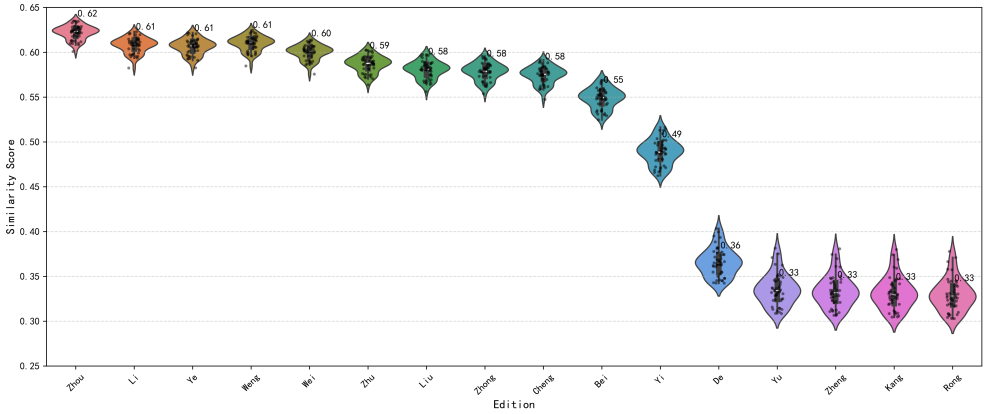


Fig. 5 Similarity distributions between illustrations and murals across different editions.

To systematically examine the visual similarity between the mural corpus and multiple illustrated editions of *RTK*, we employed the trained MISP to compute pairwise similarity scores between each mural and the illustrations of each edition. The distributional characteristics of these scores are presented using violin plots in Figure 5. In

each violin plot, each black dot represents the similarity score predicted by the MISP; the width of the violin represents the density of scores across the similarity value range; the inner shaded band denotes the interquartile range, and the white horizontal line marks the median value of the similarity scores. Notably, the similarity scores fall below 0.7 in Figure 5, consistent with the fact that 10.6% of mural–illustration similarity values are assigned under 0.7 by experts.

As illustrated in Figure 5, the violin plots of illustration–mural similarity scores separate into two clear clusters when using 0.50 as a practical demarcation. The editions "Yu", "Zheng", "Rong", "De", and "Kang" form a lower-similarity cluster characterized by broader dispersion and median values below 0.50, indicating comparatively weak visual correspondence to the murals. The "Yi" edition occupies an intermediate, borderline position: although some scores approach 0.50, the majority fall just below this level. In contrast, a higher-similarity cluster comprising "Liu", "Bei", "Ye", "Zhou", "Zhong", "Zhu", "Li", "Weng", "Cheng", and "Wei" exhibits compact distributions with most similarity scores exceeding 0.50, evidencing stronger visual coherence with the mural corpus. Within this cluster, the median similarity values of editions "Ye", "Li", "Weng", and "Zhou" surpass 0.60. Notably, the "Zhou" edition not only attains the highest median similarity score of 0.62, but its entire score distribution is concentrated in the uppermost range of the similarity scale. Both the minimum and maximum similarity values of the "Zhou" edition exceed the corresponding bounds of all other top-tier editions, indicating that its score distribution is entirely elevated relative to those of "Li", "Weng", and "Ye", whose distributions extend to comparatively lower similarity values. This distributional elevation, rather than merely a higher median, suggests that the "Zhou" edition maintains consistently superior visual correspondence to the mural corpus across the full range of mural panels.

The quantitative similarity scores in Figure 5 accord with the edition attributions previously established by social-historical scholarship. As shown in Figure 6, the mural panel "The Three Heroes Battle Lü Bu" aligns closely with the "Zhou" and "Li" editions, with the closest correspondence to the "Zhou" edition. Specifically, the mural in Figure 6(a) and the illustration in the Zhou edition Figure 6(e) both divide the scene into left and right sections. On the right section, Liu Bei, Guan Yu, and Zhang Fei each ride a swift horse, weapons in hand, aligned in a single vertical register from top to bottom. The tips of their weapons converge on Lü Bu, who, still resisting while withdrawing his mount. Behind Lü Bu, a city gate occupies the upper-left portion of the pictorial field. For the "Li" edition, according to current scholarly consensus, it was largely derived from the "Zhou" edition [33], so a relationship reflected in its second-highest similarity score to the murals. Moreover, the "Li" edition in Figure 6(c) is rendered in one section with a vertical format, the composition develops longitudinally. Although Liu Bei, Guan Yu, and Zhang Fei are likewise arrayed from top to bottom, the direction of their weapons and their spatial relation to Lü Bu diverge from the treatment in the "Zhou" edition and the mural. Like the "Li" and "Zhou" editions, the higher-scoring "Weng" edition also belongs to the Jiangnan tradition. As shown in Figure 6(d), the Jiangnan layout facilitates coherent narrative sequencing and vertically accentuated figural proportion—features likewise observable in the murals. The

"Ye" edition derives from the Jianyang lineage; the scene related with "The Three Heroes Battle Lü Bu" shown in Figure 6(b), its high similarity with murals suggests stylistic convergence, potentially mediated by the transregional circulation of shared model sheets in Thailand. Overall, the murals register a composite influence, with a dominant Jiangnan stylistic imprint and a subordinate Jianyang lineage contribution; within the Jiangnan cluster they show their strongest affinity to the "Zhou" edition. Our proposed quantitative model MISP enables probabilistic provenance hypotheses (i.e., mural panels more likely derived from "Zhou" edition).

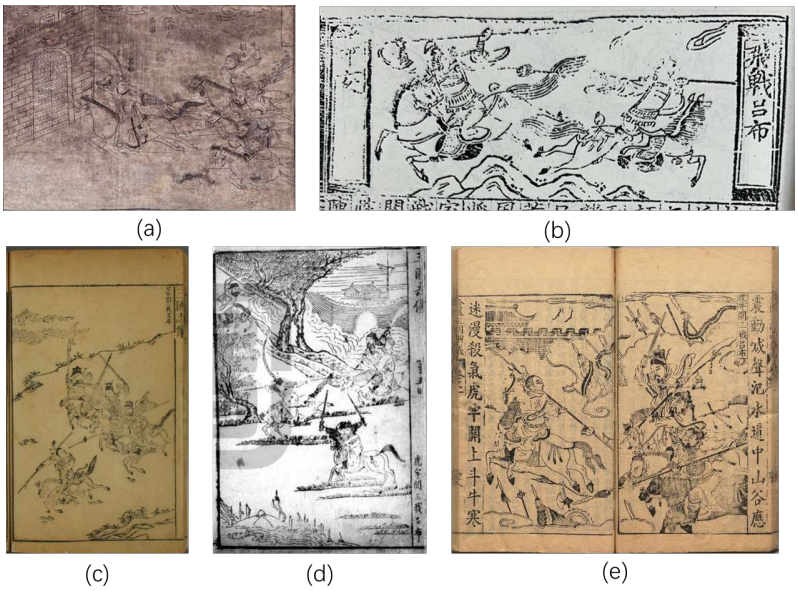


Fig. 6 Murals and printed edition illustrations depicting the episode "The Three Heroes Fight Lü Bu". (a) Mural panel 36: "Breaking the Pass—The Three Heroes Fight Lü Bu" (破關兵三英戰呂佈). (b) "Ye" edition: "Zhang Fei Fights Lü Bu" (張飛戰呂佈). (c) "Li" edition: "Third Bout with Lü Bu at Hulao Pass" (虎牢關三戰呂佈).(d) "Weng" edition: "Third Bout with Lü Bu at Hulao Pass" (虎牢關三戰呂佈). (e) "Zhou" edition: "Third Bout with Lü Bu at Hulao Pass" (虎牢關三戰呂佈).

3.3 Bootstrap Validation of Mural–Illustration Visual Correspondence

To further validate the statistical robustness of the similarity comparisons observed in Section 3.2, we apply a non-parametric bootstrap resampling procedure to each edition's similarity score distribution. Specifically, for each of the sixteen editions, 10,000 bootstrap resamples are drawn with replacement from the observed similarity scores, and the median similarity score is computed for each resample. The vertical axis in Figure 7 therefore represents the average of the median statistics obtained across all 10,000 resampling operations, providing a stable and bias-corrected estimate of each edition's central tendency. The 95% confidence interval for each edition was constructed from the 2.5th and 97.5th percentiles of the resulting bootstrap distribution,

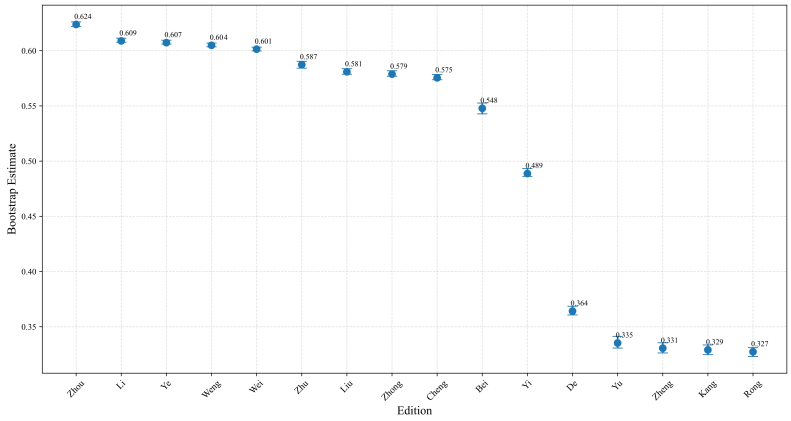


Fig. 7 Bootstrap point estimates and 95% confidence intervals of the median similarity score for each edition.

reflecting the range within which the true median similarity score is expected to fall with 95% probability.

Similar to figure 5, the bootstrap analysis yields two clearly separated performance clusters. Critically, the "Zhou" edition achieves the highest bootstrap point estimate of all sixteen editions (0.624). The confidence intervals in Figure 7 are notably narrow across all editions, indicating that the 10,000 resampling operations produced highly stable median estimates with low sampling variability. This precision allows for reliable differentiation between editions. The lower bound of the "Zhou" edition's 95% CI visibly exceeds the point estimates of editions — "Bei" (0.548), "Yi" (0.489), "De" (0.364), "Yu" (0.335), "Zheng" (0.331), "Kang" (0.329), and "Rong" (0.327) — confirming that the "Zhou" edition's lead over these editions is statistically robust under resampling. Furthermore, the "Zhou" edition's point estimate of 0.624 surpasses editions, namely "Li" (0.609), "Ye" (0.607), "Weng" (0.604), "Wei" (0.601), "Zhu" (0.587), "Liu" (0.581), "Zhong" (0.579), and "Cheng" (0.575), establishing it as the edition with the strongest and most reliable visual correspondence to the mural corpus across all resampled draws.

Taken together, the bootstrap confidence interval analysis corroborates and strengthens the distributional evidence from the violin plots, i.e., the "Zhou" edition consistently and robustly achieves the highest bootstrap-estimated median similarity to the mural corpus. These convergent quantitative findings support the conclusion that the "Zhou" edition most closely approximates the pictorial source from which the Thai mural painters drew their compositional and iconographic conventions.

4 Discussion

By quantifying the visual correspondence between the Wat Prasert Sutthawat murals and sixteen pre-1838 illustrated editions, MISP identifies the Zhou Yuejiao edition as the closest pictorial source, with a subordinate Jianyang contribution. This attribution carries a concrete chronological implication. Because the murals were executed when

the temple was built in 1838, the editions on which they drew must already have been available in Thailand by that date. The murals therefore furnish a *terminus ante quem* of 1838 for the arrival of these Chinese illustrated editions—advancing the documented horizon of *RTK* reception in Thailand by some four decades relative to the earliest surviving local copy, the 1883 *Shanmeitang* issue, and complementing the 1805 Thai translation, which, being mediated through oral retelling, cannot itself identify a source edition.

This edition-level result refines, and renders quantitative, claims that prior scholarship could advance only through qualitative connoisseurship. Where earlier studies of *RTK* transmission to Thailand rested on isolated case comparisons [7–9], our analysis shows that the murals register a composite influence dominated by the Jiangnan tradition—of which the Zhou edition is the strongest representative—with a secondary Jianyang lineage. Two patterns reinforce the reading. The high similarity of the nominally Jianyang "Ye" edition is consistent with the transregional circulation of shared model sheets, and the second-ranked "Li" edition's documented derivation from the "Zhou" edition [33] is recovered by the model rather than imposed a priori, supporting the credibility of the ranking.

More broadly, the study shows how domain-adapted multimodal contrastive learning can serve as a humanities-enabling instrument for provenance inference. By making the iconographic transformations of transregional transmission quantifiable, and by surfacing transparent, inspectable similarity rationales, MISP augments rather than supplants expert judgment and provides a reusable, scalable workflow for source attribution in other multi-version narrative corpora and for prioritising the conservation of fragile mural cycles.

Several limitations temper these conclusions. The corpus, though the largest yet assembled for this problem, remains modest, and the expert labels—binary judgments averaged over five annotators—carry an inherent subjectivity that we did not quantify with a formal inter-annotator agreement statistic; both factors bound the generalisability of the learned representations. Narrative-sequencing cues such as reading order, panel adjacency, and coherence transitions are likewise weakly represented in the current feature space. Future work will therefore enlarge the dataset with additional editions and mural cycles; add a spatio-temporal, narrative-aware layer that encodes ordered reading paths and panel sequencing; incorporate other Southeast Asian *RTK* visual corpora to trace stylistic evolution and quantify regional adaptation; and integrate textual paratexts, spectral pigment data, and conservation records to enrich the multimodal embeddings. These directions should sharpen lineage attribution and further improve interpretability.

Data availability

Part of the original dataset is publicly available at https://github.com/MiyaWu/MISP-Romance-of-the-Three-Kingdoms/tree/main/dataset. The remaining data and materials supporting the findings of this study are available from the corresponding author upon reasonable request.

Code availability

The code of this study is publicly available at https://github.com/MiyaWu/MISP-Romance-of-the-Three-Kingdoms.

Acknowledgements

This study was supported by the National Natural Science Foundation of China (Nos. 62106092 and 22074058) and the National Social Science Fund of China Post-Funding Project (No. 20FZWB038); the Natural Science Foundation of Fujian Province (Nos. 2025J01981 and 2024J01822); the Natural Science Foundation of Zhangzhou City (No. ZZ2024J28); and the Industry–University–Research Collaboration Project of the Xiamen Science and Technology Plan (No. 2023CXY0414) and the Graduate Education Reform Project of Minnan Normal University (No. YJG202305).

Author contributions

Y.L., H.W. and S.L. conceived and designed the study. Y.S., H.W., L.F. and F.Y. designed the MISP model, and H.W., L.F., F.Y. and W.Y. implemented the software. Y.S., W.Y. and H.W. performed the experiments and formal analysis. Y.S. assembled and curated the mural and illustration corpus. Y.L. led the art-historical interpretation, to which Y.S. contributed a computational perspective, while B.L. contributed the philological and in-image textual interpretation; Y.S., Y.L. and B.L. designed and supervised the expert annotation. Y.S. and F.Y. produced the visualizations. C.Y.T. and H.W. validated the technical results. Y.S., H.W. and Y.L. wrote the original draft. F.Y., B.L., C.Y.T. and S.L. reviewed and edited the manuscript. Y.S., Y.L., S.L., H.W. and F.Y. acquired funding. Y.L., H.W. and S.L. supervised the project. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

References

[1] Plaks, A. H. *Chinese Narrative: Critical and Theoretical Essays* Princeton legacy library edn (Princeton University Press, 1977).

[2] Plaks, A. H. *The Four Masterworks of the Ming Novel* (Princeton University Press, 1987).

[3] Pan, J. *A Study on the Catalogue of Ancient Chinese Novels* (Shanghai Ancient Books Publishing House, Shanghai, China, 2005).

[4] Ratchanupap, D. *History of the Three Kingdoms* (Flower and Grass Press, Bangkok, 2001).

[5] Uachoowong, C. *Samkok: The Three Kingdoms Murals in the Ubosot of Wat Prasert Sutthawat*. Master's thesis, Silpakorn University, Thailand (2005).

[6] Chen, L. *The Influence of Chinese Art on Buddhist Art in Thailand during the Reigns of Rama II and Rama III*. Ph.D. thesis, Peking University, Beijing, China (2012).

[7] Jin, Y. *Form Like and Spirit Divine: The Past and Present Transmission of Romance of the Three Kingdoms in Thailand* (Peking University Press, Beijing, China, 2018).

[8] Wang, S. On bangkok-era murals in thailand: Focusing on chinese elements in the wat prasert sutthawat murals. *Studies in Sino–Foreign Art* 120–127 (2020).

[9] Hu, C. An exploration of the romance of the three kingdoms murals in kuan an keng shrine, bangkok. *Journal of Hubei Institute of Fine Arts* 11–16 (2019).

[10] Quan, H., Li, Y., Liu, D. & Zhou, Y. Protection of guizhou miao batik culture based on knowledge graph and deep learning. *Heritage Science* **12**, 202 (2024).

[11] Wu, H. *et al.* Network analysis and sentiment classification of minnan nursery rhymes. *npj Heritage Science* **13** (2025).

[12] Liu, Y., Wang, Y. & Liu, C. A deep-learning method using auto-encoder and generative adversarial network for anomaly detection on ancient stone stele surfaces. *arXiv preprint arXiv:2308.04426* (2023).

[13] Fujikawa, Y. *et al.* Recognition of oracle bone inscriptions by using two deep learning models. *International Journal of Digital Humanities* **5**, 65–79 (2023).

[14] Jaramillo, P. & Sipiran, I. Del Bue, A., Canton, C., Pont-Tuset, J. & Tommasi, T. (eds) *Cultural heritage 3d reconstruction with diffusion networks*. (eds Del Bue, A., Canton, C., Pont-Tuset, J. & Tommasi, T.) *Computer Vision – ECCV 2024 Workshops*, 104–117 (Springer Nature Switzerland, 2025).

[15] Baldrati, A., Bertini, M., Uricchio, T. & Del Bimbo, A. Furferi, R. *et al.* (eds) *Exploiting clip-based multi-modal approach for artwork classification and retrieval*. (eds Furferi, R. *et al.*) *The Future of Heritage Science and Technologies: ICT and Digital Heritage*, 140–149 (Springer International Publishing, 2022).

[16] Conde, M. V. & Turgutlu, K. Clip-art: Contrastive pre-training for fine-grained art classification. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* 3951–3955 (2021).

[17] Radford, A. *et al.* Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning* 8748–8763 (2021). URL https://proceedings.mlr.press/v139/radford21a.html.

[18] Daunhawer, I., Bizeul, A., Palumbo, E., Marx, A. & Vogt, J. E. Identifiability results for multimodal contrastive learning. *The Eleventh International Conference on Learning Representations* (2023).

[19] Kim, W., Son, B. & Kim, I. Vilt: Vision-and-language transformer without convolution or region supervision. *Proceedings of the 38th International Conference on Machine Learning* (2021).

[20] Wang, Z. *et al.* Connecting multi-modal contrastive representations. *Thirty-seventh Conference on Neural Information Processing Systems* (2023).

[21] Shorten, C. & Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *Journal of Big Data* **6**, 60 (2019).

[22] Luo, D. *et al.* Unsupervised document embedding via contrastive augmentation. *arXiv preprint arXiv:2103.14542* (2021).

[23] Tan, M. & Le, Q. Chaudhuri, K. & Salakhutdinov, R. (eds) *Efficientnet: Rethinking model scaling for convolutional neural networks.* (eds Chaudhuri, K. & Salakhutdinov, R.) *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, Vol. 97, 6105–6114 (PMLR, 2019).

[24] He, H., Zhang, J., Lan, Z. & Zhang, Y. Instance smoothed contrastive learning for unsupervised sentence embedding. *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence* (2023).

[25] Liu, Y. *et al.* Contrastive multimodal fusion with tupleinfonce. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* 734–743 (2021).

[26] Zolfaghari, M., Zhu, Y., Gehler, P. & Brox, T. Crossclr: Cross-modal contrastive learning for multi-modal video representations. *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)* 1430–1439 (2021).

[27] Franceschini, R. *et al.* Multimodal emotion recognition with modality-pairwise unsupervised contrastive loss. *2022 26th International Conference on Pattern Recognition (ICPR)* 2589–2596 (2022).

[28] Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. III, H. D. & Singh, A. (eds) *A simple framework for contrastive learning of visual representations.* (eds III, H. D. & Singh, A.) *Proceedings of the 37th International Conference on Machine Learning*, Vol. 119 of *Proceedings of Machine Learning Research*, 1597–1607 (PMLR, 2020).

[29] He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. B. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 9726–9735 (2019).

[30] Li, J., Li, D., Xiong, C. & Hoi, S. Chaudhuri, K. *et al.* (eds) *BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation.* (eds Chaudhuri, K. *et al.*) *Proceedings of the 39th International Conference on Machine Learning*, Vol. 162, 12888–12900 (PMLR, 2022).

[31] Yu, J. *et al.* Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research* (2022).

[32] Li, J. *et al.* Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. & Vaughan, J. W. (eds) *Align before fuse: Vision and language representation learning with momentum distillation.* (eds Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. & Vaughan, J. W.) *Advances in Neural Information Processing Systems*, Vol. 34, 9694–9705 (Curran Associates, Inc., 2021).

[33] Hu, X. *A Study of the Relationship Between Illustrations and Text in Ming Printed Editions of the Romance of the Three Kingdoms.* Phd dissertation, Fujian Normal University, Fuzhou, China (2015).