ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Multimodal AI-Based Semantic Annotation and Knowledge Graph Construction for the Motif of Withered Trees, Bamboo, and Rocks in Literati Painting

Jikang Cao[1,*]

[1] School of Art and Design, Jiangxi Institute of Technology, Nanchang 330098, Jiangxi, China

SUMMARY: *This research puts forward a multi-modal artificial intelligence frame for semantic mark and knowledge graph building of the theme of dead trees, bamboos and stones in Chinese scholar painting. A motif-directional dataset called WTBR-LP is built from open museum pictures, metadata files, names, carved words, description words, and expert labels. The marks explanation system includes three layers: vision things, pen skill shape, and culture meaning. The put-forward model incorporates open-vocabulary visual location-finding, area cutting, vision-language expression, metadata embedding, and graph-neighbor restrictions to produce region-level multi-label notes. The annotation outcomes are further changed into WTBR-KG, which is a source-conscious knowledge graph that connects art works, people, visual motifs, brush stroke characteristics, inscriptions, collection origins, and cultural concepts. In the early-stage experiment configuration, our put-forward method obtains a Macro-F1 of 0.864, mAP@0.5 of 0.803, Hits@1 of 0.912, and triple accuracy of 0.923. The work flow that is assisted by AI cuts down the time of expert checking from 8.70 minutes to 2.05 minutes on each single artwork. The obtained results show that multimodal fusing and graph restriction can promote the accuracy, interpretability, and traceability of semantic marking for scholar painting.*

KEYWORDS: *literati painting; withered trees bamboo and rocks; multimodal AI; semantic annotation; knowledge graph*

1 Introduction

The object confronted by the digitized collation of literati paintings is not a single image itself, but a composite of images, titles, inscriptions, seals, collection records, and art historical contexts. Dead wood, bamboo, and stone are among the most suitable matrices to enter the semantic calculation. Dead wood is often associated with antiquity, decay, and stoicism; bamboo often points to temperance, purity, and writing; and stone assumes the meaning of compositional support, strange interest, and hidden space. The collection system is usually able to record the title, author, date, material and collection number of the work, but it is more difficult to organize the dead wood branches, bamboo leaves, chapped stones, inscription semantics and personality metaphors in the picture into retrievable structured knowledge. For the researcher, the question is not only "whether a certain work is painted with bamboo and stone", but also which areas are painted with bamboo and stone, what kinds of ink and brushwork are used in these areas, how the inscriptions explain these objects, and whether the relevant semantics can form a traceable matriarchal genealogy across the works in the collection.

*15664903038@163.com
https://doi.org/10.65102/is2026766
Published: 30 April, 2026

Studies surrounding Su Shi's Dead Wood and Strange Stones and related traditions have illustrated the strong inscription-dependent and viewing-context-dependent nature of dead wood, bamboo, and stone in literati paintings from the Song dynasty onward.Sturman's study of Song literati inscription practices suggests that traces of writing outside of the image alter the mode of transmission and structure of meaning in a work [1].Ahn's discussion of Su Shi's theory of bamboo painting further suggests that bamboo and stone motifs cannot be categorized only by natural objects, but their meanings are often generated through bodily experience, writing actions, and literati subjective consciousness [2]. These studies have imposed a constraint on the digitization process: if the model only recognizes the three visual objects of "tree", ‘bamboo', and "stone", the most crucial semantic layer of literati painting will be lost.

Computational methods have begun to enter the study of Chinese paintings. Zhang et al. sorted out the computational research paths of traditional Chinese paintings from the perspective of "six methods", pointing out that image understanding, style analysis and semantic modeling need to be combined with the internal formal logic of Chinese paintings [3].Liong et al. carried out a benchmark study on the categorization of traditional Chinese paintings and proved that deep learning can support automatic image classification. deep learning can support automatic image categorization, but its task objectives are still dominated by painting families, styles, or coarse-grained categories [4]. This type of work provides a foundation for visual analysis, but less access to localized matrices like dead wood, bamboo, and stone, brush and ink patterns, and cultural semantic layers. The difficulty with literati paintings is that withered branches may be formed by only a few dry strokes, sloping rocks and foothill chalk lines are close in grayscale images, and bamboo leaves and grasses have similarities in local textures. Coarse-grained classification models have difficulty in stably distinguishing these boundaries.

Knowledge mapping provides another path for this problem.WuMKG has constructed a multimodal knowledge map around Chinese painting and calligraphy, dealing with the structural relationships between works, characters, seals, themes and image resources [5].CP-MNER further proposes a multimodal named entity recognition dataset and knowledge enhancement fusion framework for Chinese painting, illustrating that there is an obvious semantic gap between images and description texts. External knowledge helps entity recognition and downstream knowledge construction [6]. A multimodal knowledge mapping review also pointed out that co-modeling of images, text and structural relationships can enhance retrieval, reasoning and question-answering of complex objects [7]. In the field of cultural heritage metadata complementation, Rei et al. used image, text, and tabular metadata for multimodal attribute prediction, demonstrating that late fusion and knowledge graph storage are suitable for dealing with heterogeneous information about cultural heritage objects [8]. These studies suggest that the digitization of deadwood and bamboo stones should move from single image recognition to multimodal evidence organization.The information of the paper by WuMKG and CP-MNER has been verified by the publication page.

Existing methods still have three shortcomings. First, the generic tags of the Literature and Museum platform are mostly used to express visual objects with words such as "tree", "bamboo", ‘rock', "landscape", and so on. The first is that the generic label of the Literature and Museum Platform mostly expresses the visual objects with the words "tree", "bamboo", "rock", ‘landscape', etc., and lacks the words "dead wood", "strange stone", "sloping stone", "bamboo and stone together", "dry brush", and "dry brush", "dry brush chapping" and other internal terms of literati painting. Second, the inscriptions and collection descriptions contain a large number of semantic clues, such as "ancient meaning," "escape," "clear and open," and "dispersed." but these words do not always map directly to the image area. Thirdly, the knowledge map is often based on works, authors, eras, and collections, and the evidence chain between local image

areas and cultural semantics is insufficient, which makes it difficult to trace back the judgment of "the work has an ancient meaning" to the image evidence or textual evidence.

In this paper, we construct a multimodal semantic annotation and knowledge map generation method for the parent topic of literati paintings of dead wood, bamboo and stone. Based on the open collection images, titles, inscription texts, description texts and expert review labels, the research refines dead wood, bamboo and stone from the generic object category into object layer, morphological layer and cultural semantic layer. The model side combines open-vocabulary visual detection, region segmentation, visual linguistic representation, metadata embedding and atlas neighborhood constraints to achieve region-level multi-label prediction. The atlas end transforms model outputs into triples with source, confidence and ontology constraints for cross-collection search, parent genealogy analysis and case interpretation.

The contribution of this paper focuses on three aspects. First, the caliber of the WTBR-LP dataset and the three-layer labeling system of Kukkiu, Bamboo, and Stone are proposed to enable the local matrices of literati paintings to enter into a reviewable semantic annotation process. Second, designing the annotation model with joint multimodal fusion and graphical constraints to solve the alignment problem between local ink regions, inscription text and abstract semantics. Third, we construct the WTBR-KG knowledge graph, which organizes works, figures, motifs, ink and brush forms, inscription entities, cultural semantics, and curatorial sources into a queryable structure, providing a reusable method for digitizing and organizing literati paintings.

2 Methods

2.1 Data Sources, Motif Taxonomy and Annotation Protocol

In this paper, we construct a WTBR-LP dataset for region annotation, semantic categorization, and knowledge graph generation of the parent themes of literati paintings of dead wood, bamboo, and stone. The data sources include open collection images, collection metadata, inscriptions, description texts, transcription of inscriptions and expert review records. Image interoperability is performed using the IIIF image interface caliber [9], and open collections are mainly from The Metropolitan Museum of Art Open Access, Cleveland Museum of Art Open Access API, Harvard Art Museums API, and Smithsonian Open Access [10, 13]. These sources provide images, work numbers, authors, dates, materials, dimensions, curatorial information, and some descriptive text, and are suitable for building cross-collection sample pools. All of the above open-collection sources can be verified through the official pages.

Sample searching is performed using a combination of English and Chinese keywords. English keywords include "withered tree", "old tree", "bamboo", ‘rock', "scholar rock", "literati rock", and "rock". " "scholar rock" "literati painting" "Chinese painting "; Chinese keywords include "withered tree" "old tree" "bamboo" ‘rock' "scholar rock" "literati painting" "Chinese painting". Chinese keywords include "dead wood", "ancient wood", "bamboo and stone", "strange stone", "sloping stone", "literati painting", "Chinese painting", and "ink". A total of 4,318 records were screened, and 1,184 works, 1,768 images, and 3,024 paragraphs of text were retained after deleting duplicate images, low-resolution thumbnails, non-Chinese pictorial objects, and extraneous decorative motifs. In order to avoid local images of the same work from entering different data sets, the training, validation and test sets are divided by work number with the ratio of 8:1:1. The data organization and evidence chain composition are shown in Fig. 1.

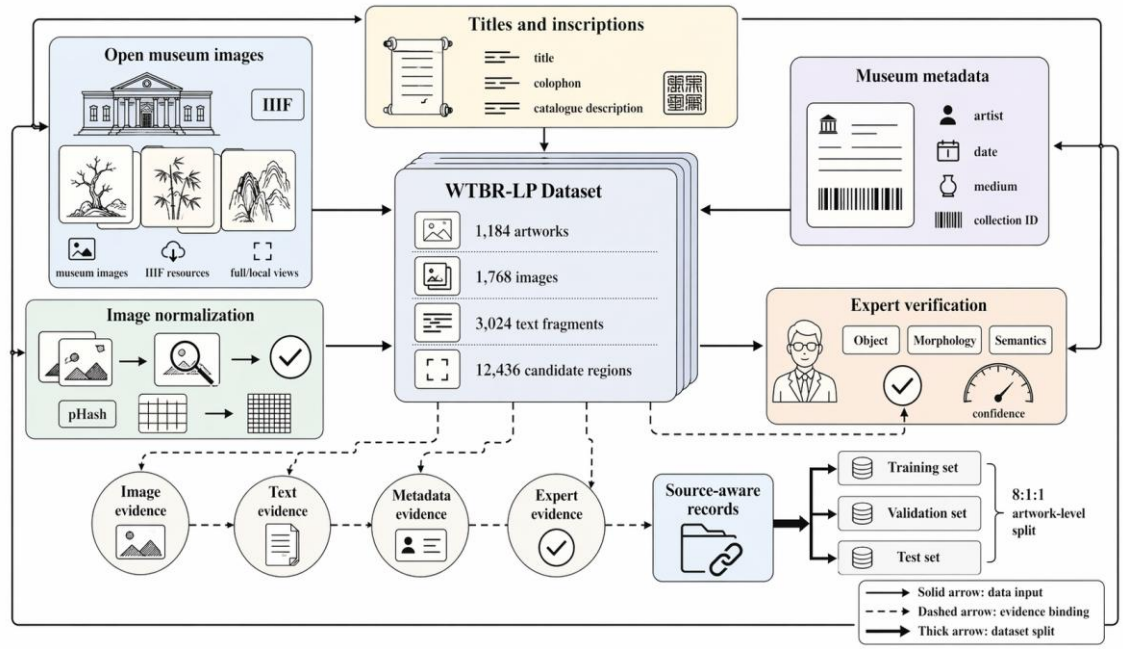

Figure 1: Literati Painting Deadwood and Bamboo Data Organization and Evidence Chain Composition.

In Figure 1, open collection images provide regional evidence, collection metadata provide work-level facts, titles and inscriptions provide semantic trigger words, and expert review records provide high-confidence labels. The four types of evidence are merged under the work number and subsequently enter the regional annotation and map generation process.WTBR-LP dataset composition and labeling system. As shown in Table 1.

Table 1: WTBR-LP dataset composition and labeling system

Data Item	Quantity/Category	Description
Initial Screening Records	4318 entries	Search results from Chinese and English keywords
Deduplicated Works	1184 items	Merged by collection number, title, and image hash
Image Resources	1768 images	Full images, partial images, different resolution versions
Text Segments	3024 segments	Titles, descriptive texts, transcriptions of inscriptions, exhibition descriptions
Candidate Regions	12436 regions	Generated by open vocabulary detection and segmentation model
Object Layer Labels	10 types	Categories such as dead wood, bamboo, stone, slope, inscriptions, seals, etc.
Morphological Layer Labels	18 types	Categories such as bent branches, dry brush strokes, bamboo leaf density, stone surface textures, etc.
Cultural Semantic Labels	14 types	Categories such as ancient meanings, seclusion, integrity, clarity, solitude, ethereal quality, etc.
Training/ Validation/ Testing	946/119/119 items	Divided by works to prevent leakage of partial images
Annotation Consistency	κ=0.87/0.80/0.73	Statistics for object layer, morphological layer, and semantic layer respectively

The labeling system is divided into three layers. The object layer labels the areas of dead wood, bamboo, stone, sloping banks, inscriptions and seals visible in the picture. The morphology layer describes the trend of branches and trunks, the intensity of ink color, the combination of bamboo and leaves, the contour of the stone, the way of chapping, and the sparseness of the lines. The cultural and semantic layer is labeled with "ancient meaning", "seclusion", "temperance", "openness", ‘solitude', and "solitude". Concepts such as "loneliness" and "elegance". Semantic layer labels can not be separated from the evidence alone, each semantic judgment must be bound to at least one image area, text fragment or collection description of the source. The labeling hierarchy and annotation rules are shown in Figure 2.

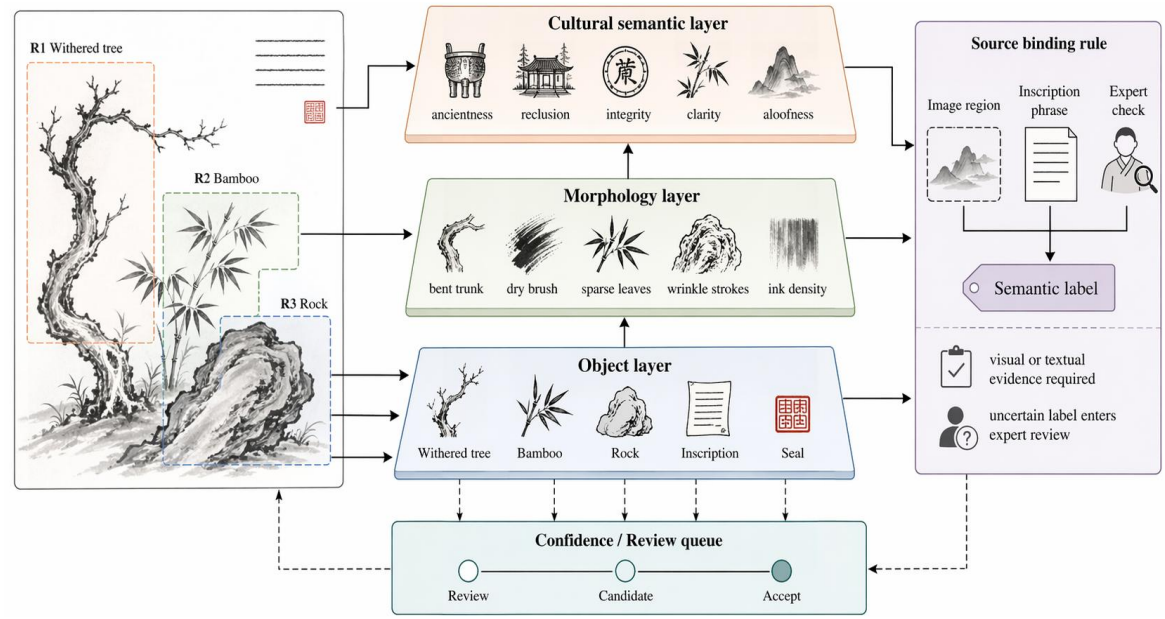

Figure 2: The labeling system and annotation hierarchy of the Kukui Bamboo and Stone Matrix.

In Figure 2, the object layer is responsible for locating visible motifs, the morphological layer is responsible for recording ink and structural features, and the cultural semantic layer is responsible for expressing interpretive concepts. Bottom-up evidence constraints are formed between the three layers, and semantic labels enter the training set only when image or text support exists.

Manual labeling uses a two-round mechanism. In the first round, two annotators independently labeled the object and morphological layers; in the second round, experts in the direction of the history of painting and calligraphy reviewed the cultural semantic layer. The object layer has the highest consistency, with Cohen's κ of 0.87. The morphological layer has a κ of 0.80, with the errors mainly coming from the boundary differences between the sloping stones and the chapped lines at the foot of the mountain, and the dead branches and the lines of the grasses and trees. The κ of the cultural semantic layer is 0.73, indicating that the labels of "ancient meaning", "clear and open", "elegant", etc. are highly dependent on the inscriptions and the context of the works, and cannot be judged entirely by the texture of the images.

2.2 Multimodal Semantic Annotation Model

The goal of the semantic annotation model is to generate multi-labeled results for each candidate region at the object, morphological, and cultural semantic levels. Model inputs include the whole image, candidate regions, title, caption or description text, curatorial metadata, and atlas neighborhoods. The visual side uses CLIP to obtain image and text alignment

representations [14], and Grounding DINO to obtain image and text alignment representations based on "withered tree" ‘bamboo' "rock ", ‘inscription', "seal" and other cues to generate candidate frames [15], and then SAM to generate region masks [16]. The textual side uses visual language models to extract candidate entities and semantic phrases in the title, title trek, and description texts, and BLIP-2 and LLaVA are used to assist in generating textualized descriptions of the image region descriptions and fuzzy regions [17, 18].The modeling sources of CLIP, SAM, BLIP-2, Grounding DINO, and LLaVA can be verified through the open paper pages. The information coupling mechanism of the multimodal semantic annotation models is shown in Figure 3.

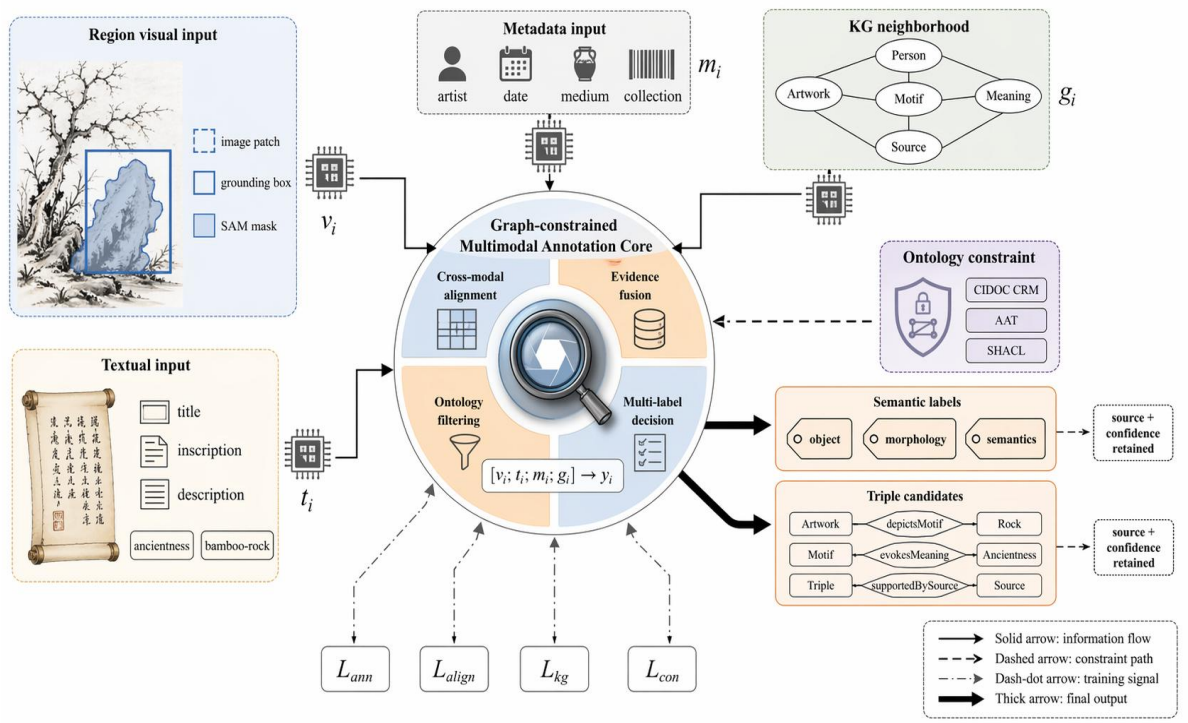

Figure 3: Information coupling mechanism of multimodal semantic annotation models.

In Figure 3, region image vectors provide visual evidence, text vectors provide title and description semantics, metadata vectors provide authorship, chronology, and curatorial context, and atlas neighborhood vectors provide structural constraints between parent topics and concepts. The four types of information are used in the fusion layer to compute regional-level multilabeling results. The multi-label prediction vector, shown in equation (1).

$$\hat{y}_i = \sigma\left(W_f[v_i; t_i; m_i; g_i] + b_f\right) \tag{1}$$

where $\hat{y}_i$ denotes the multi-label prediction vector of the ith candidate region, v_i denotes the regional visual vector, t_i denotes the text vector, m_i denotes the collection metadata vector, g_i denotes the atlas neighborhood vector, W_f and b_f are the parameters of the fusion classification layer, and σ is the multi-label activation function. This structure makes the object labels such as dead wood, bamboo and stone mainly triggered by regional images, and makes the semantic labels such as "ancient meaning", ‘seclusion' and "modesty" subject to both text and atlas neighborhood constraints. The cross-modal alignment loss function is shown in equation (2).

$$\mathcal{L}_{align} = -\frac{1}{N}\sum_{i=1}^{N} log \frac{exp(cos(v_i,t_i)/\tau)}{\sum_{j=1}^{N} exp(cos(v_i,t_j)/\tau)} \quad (2)$$

where $\mathcal{L}_{align}$ denotes the cross-modal alignment loss, N denotes the number of samples in the training batch, $cos(\cdot)$ denotes the cosine similarity, and τ denotes the temperature coefficient. This loss is used to bring the visual evidence of the region closer to the corresponding text phrase and reduce the misleading effect of the title on irrelevant regions. For example, when "bamboo and stone" appears in the title, the model only improves the confidence of the relevant labels locally in the visual region where there is evidence of both bamboo and stone, instead of spreading the work-level title to all regions.

2.3 Knowledge Graph Construction and Experimental Design

WTBR-KG uses works as the central node to connect authors, dates, curatorial sources, image areas, visual motifs, pen and ink forms, inscription texts, and cultural semantic concepts. The ontology design adopts CIDOC CRM as the upper-layer framework for cultural heritage objects, people, events, and information resources [19], uses Getty AAT for terminological alignment of materials, techniques, and art terminology [20], and refers to Linked Art's artwork data modeling approach for organizing artworks, images, people, and curatorial events [21]. The underlying representation uses the RDF triad model [22], the terminology layer uses SKOS to express conceptual hierarchies and synonyms [23], the constraints layer uses SHACL to check relational domains, value domains, and mandatory attributes [24], and the query layer uses SPARQL to support capability problem validation [25]. cidoc CRM, Getty AAT, and Linked Art are cultural heritage verifiable sources in semantic modeling. The design of the core classes and relationships of the Kukki Bamboo and Stone Knowledge Graph is shown in Table 2.

Table 2: Core Classes and Relationships Design of Kukki Bamboo and Stone Knowledge Graph

Type	Name	Description	Example Relationship
Core Class	Artwork	Records of artworks or sections of artworks	createdBy, heldBy
Core Class	VisualMotif	Themes such as dead wood, bamboo, stone, etc.	depictsMotif, coOccursWith
Core Class	MotifRegion	Candidate regions within an image	locatedIn, hasMask
Core Class	BrushworkFeature	Features such as dry brush, wet brush, textures, etc.	hasBrushwork
Core Class	InscriptionText	Inscriptions, titles, descriptive texts	hasInscription, mentions
Core Class	CulturalMeaning	Concepts such as ancient meanings, seclusion, integrity, solitude, etc.	evokesMeaning
Core Class	Person	Authors, inscribers, collectors	createdBy, inscribedBy
Core Class	CollectionSource	Collecting institutions and open data platforms	providedBy
Relationship	depictsMotif	Artwork or region depicting a motif	Artwork/MotifRegion → VisualMotif
Relationship	hasMorphology	Region possessing morphological features	MotifRegion → BrushworkFeature
Relationship	evokesMeaning	Motif or artwork triggering cultural meaning	VisualMotif/Artwork → CulturalMeaning
Relationship	supportedBySource	Evidence source for a triple record	TripleRecord → SourceRecord
Relationship	alignedToTerm	Local term aligned with an external term	LocalTerm → AAT/SKOS Term

Graph Completion uses a combination of embedding scoring and ontology constraints. TransE provides the basis for translational relational modeling [26], RotatE handles antisymmetric, inverse, and combinatorial relations [27], R-GCN is suitable for learning node representations on multirelational graphs [28], and GAT is used for neighborhood weight assignment [29]. In this paper, we use low-complexity embedding scoring in the candidate ternary generation stage, and use SHACL constraints and manual review results to filter relations in unsemantic domains in the final inbound stage. The candidate triad scoring, as shown in equation (3).

$$s(e_s, r, e_o) = \|\mathbf{z}_s + \mathbf{z}_r - \mathbf{z}_o\|_2 \tag{3}$$

where $s(e_s, r, e_o$ denotes the candidate triad score, e_s denotes the subject entity, r denotes the relation type, e_o denotes the object entity, and $\mathbf{z}_s$, $\mathbf{z}_r$, $\mathbf{z}_o$ denote the embedding vectors of the subject, relation, and object, respectively. The lower the score, the higher the consistency of the candidate triad in the embedding space. The loss function is shown in equation (4).

$$\mathcal{L} = \lambda_1 \mathcal{L}_{ann} + \lambda_2 \mathcal{L}_{align} + \lambda_3 \mathcal{L}_{kg} + \lambda_4 \mathcal{L}_{con} \tag{4}$$

where $\mathcal{L}$ denotes the total loss, $\mathcal{L}_{ann}$ denotes the multi-label annotation loss, $\mathcal{L}_{align}$ denotes the cross-modal alignment loss, $\mathcal{L}_{kg}$ denotes the map complementation loss, $\mathcal{L}_{con}$ denotes the ontology constraint loss, and $\mathcal{L}_{con}$ to λ_4 denote the corresponding weights. The four weights in the pre-experiment are set to 1.0, 0.5, 0.3, and 0.2, respectively, and the experimental task, comparison method, and evaluation protocol are shown in Fig. 4.

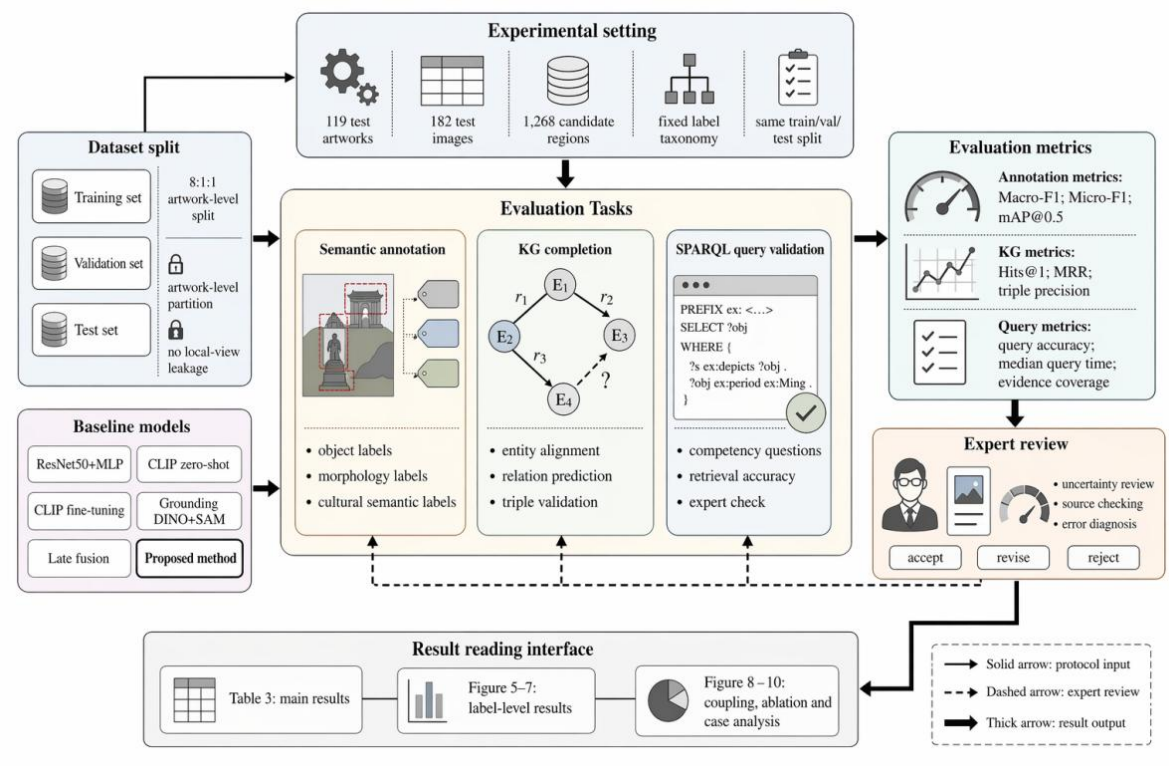

Figure 4: Experimental Tasks, Comparison Methods and Evaluation Protocols.

In Figure 4, the experiments are divided into three types of tasks: semantic annotation, graph completion and query verification. The semantic annotation task evaluates Macro-F1, Micro-

F1, mAP@0.5 and mAP@0.5:0.95; the graph completion task evaluates Hits@1, Hits@10, MRR and ternary precision; the query validation task evaluates the hit rate of competence questions and manual review consistency.

3 Results and Discussion

3.1 Semantic Annotation Performance and Label-Level Analysis

This section tests whether the model is able to stably recognize the withered wood, bamboo and stone matrices and further distinguish morphological features and cultural semantics. The test set contains 119 works, 182 images and 1268 candidate regions. The region-level prediction, artwork-level aggregation results and atlas entry results are retained for evaluation to avoid masking local errors with only whole-image labels. The distribution of misclassification between different labels is shown in Figure 5.

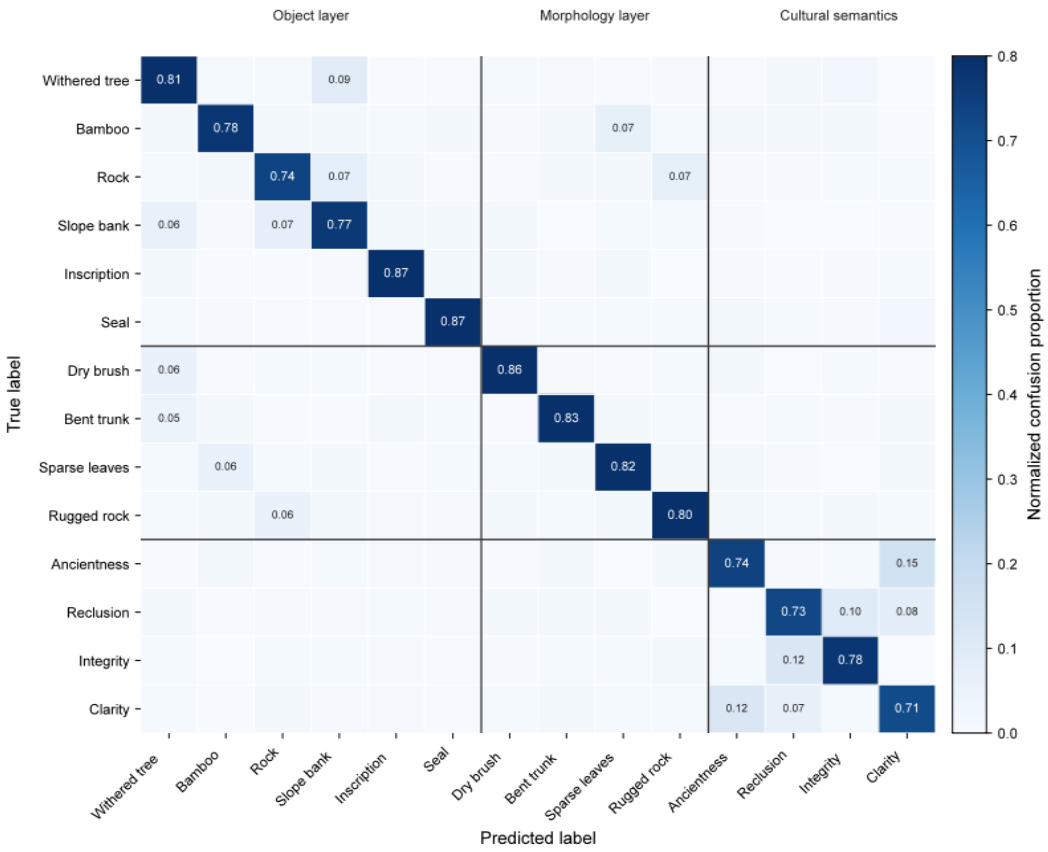

Figure 5: Heatmap of dead wood, bamboo and stone multi-label recognition confusion.

In Figure 5, the confusion between dead wood and slope bank is mainly concentrated in the position of dry brush branches close to the surface, the confusion between stone and slope bank is concentrated in the foot of the mountain area where the chapped lines are dense, and the confusion between bamboo leaves and grass lines mostly occurs in the low-resolution local map. In the semantic layer, the interferences between "ancient meaning" and "openness" are the highest, with a confusion rate of 0.18; the interferences between "integrity" and "solitude" are 0.18; the interferences between "integrity" and "solitude" are 0.15; and the interferences between ‘integrity' and "solitude" are 0.16. "The average recall rate is 0.902 for the object layer, 0.841 for the morphological layer, and 0.786 for the semantic layer, indicating that the semantic

layer is still the main difficulty. The main results of semantic annotation and graph complementation are shown in Table 3.

Table 3: Semantic Annotation and Atlas Completion Main Results

Method	Input Modality	Macro-F1	Micro-F1	mAP@0.5	Hits@1	MRR	Triple Accuracy
ResNet50+MLP	Image	0.667	0.710	0.604	—	—	—
CLIP Zero-shot	Image + Label Text	0.704	0.742	0.636	—	—	—
CLIP Fine-tuning	Image + Label Text	0.761	0.804	0.696	0.781	0.824	0.852
Grounding DINO + SAM	Image Region + Prompt	0.793	0.831	0.742	—	—	—
Late Fusion	Image + Text + Metadata	0.817	0.850	0.758	0.846	0.881	0.887
Proposed Method	Image + Text + Metadata + Graph	0.864	0.891	0.803	0.912	0.938	0.923

In Table 3, the Macro-F1 of this paper is 0.864, which is 0.047 higher than Late fusion and 0.103 higher than CLIP fine-tuning. mAP@0.5 reaches 0.803, which is 0.061 higher than Grounding DINO+SAM, which indicates that region detection and segmentation can only solve the "where" problem. This indicates that region detection and segmentation can only solve the problem of "where", while graphical semantics and graphical constraints still contribute to ‘what' and "what it means". In terms of graph completion, Hits@1 reaches 0.912 and ternary precision reaches 0.923, which indicates that supportedBySource, alignedToTerm and evokesMeaning can reduce the error of "the vocabulary of the title is directly equivalent to the semantics of the work" after constraints are applied. " error. The performance difference between the three types of tasks is shown in Figure 6.

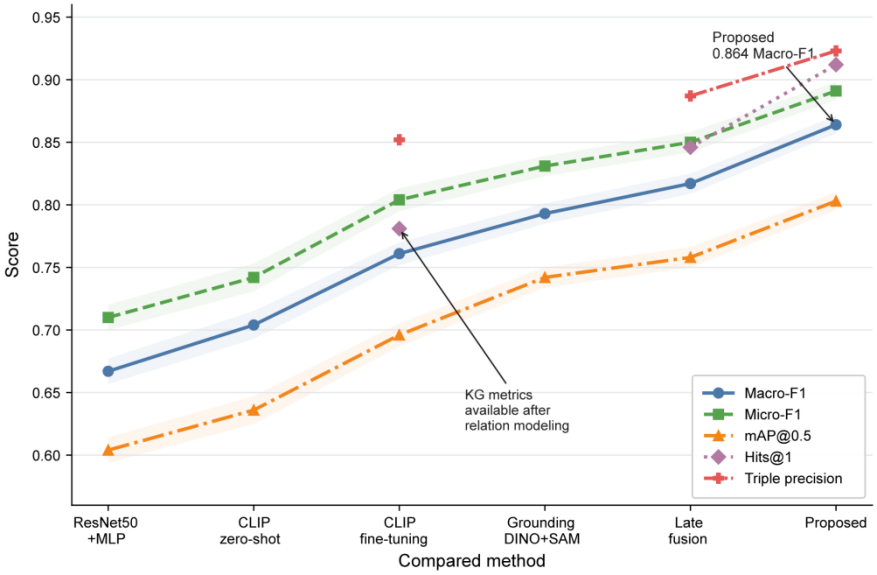

Figure 6: Performance curves of different methods on three types of tasks.

In Figure 6, the Macro-F1 curve has the largest slope from 0.817 to 0.864 after the introduction of text and graph, the mAP@0.5 curve has a smaller increase from 0.758 to 0.803, and the Hits@1 curve has an increase from 0.846 to 0.912. This trend indicates that graph neighborhood has a stronger influence on relationship prediction and semantic attribution, and a relatively indirect influence on region localization. Indirect. The relationship between label granularity, recall and confidence calibration is shown in Figure 7.

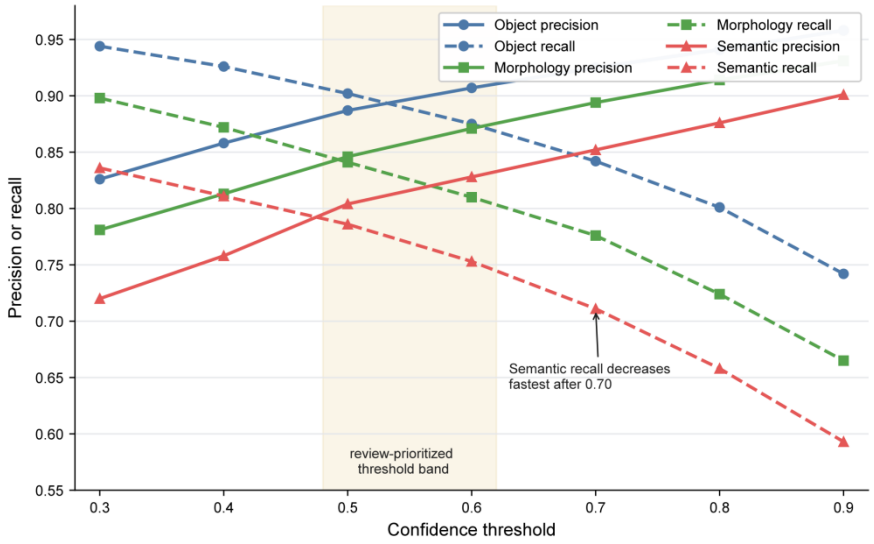

Figure 7: Label granularity, recall and confidence calibration.

In Figure 7, the recall of the object layer reaches 0.902, the morphological layer 0.841, and the semantic layer 0.786 at the confidence threshold of 0.50. After the threshold is increased from 0.50 to 0.70, the precision of the object layer rises from 0.887 to 0.925, while the recall of the semantic layer decreases from 0.786 to 0.711. This result shows that it is not appropriate to adopt a too high uniform threshold for the cultural-semantic layer, and it is suitable to use the uncertain sample prioritization review mechanism. It is suitable to use the mechanism of prioritizing the review of uncertain samples. For labels such as "ancient meaning", ‘elegant' and "clear and open", the system should keep the medium-confidence candidates and show the chain of evidence, and let the experts make the final judgment.

3.2 Ablation, Efficiency and Three-Dimensional Coupling Analysis

The main results show that multimodal fusion is effective, but there is still a need to determine the contribution and computational cost of each module. In this section, the text input, region mask, ontology constraints, graph neighborhood and uncertain sample review interfaces are ablated and the single graph inference time versus manual review time is recorded. The three-dimensional relationship between text completeness, atlas neighborhood density and annotation performance is shown in Figure 8.

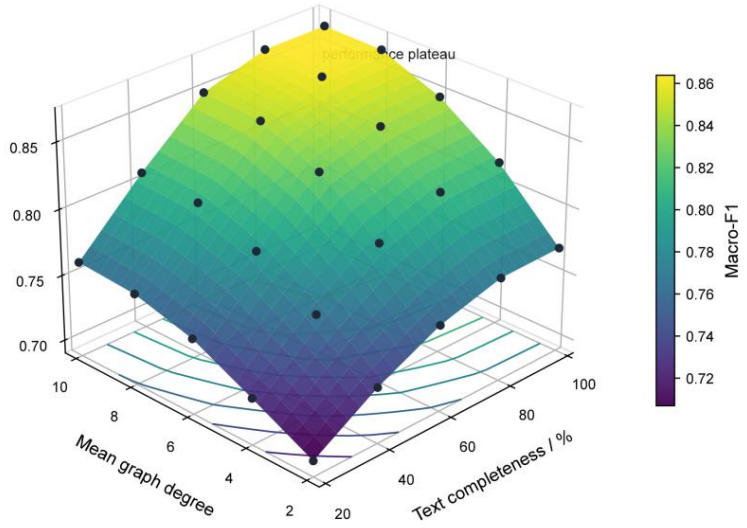

Figure 8: Three-dimensional response surface of text completeness, graph neighborhood density and annotation performance.

In Figure 8, the X-axis is the text completeness in %, the Y-axis is the average neighborhood density of the atlas, and the Z-axis is the Macro-F1. When the text completeness is increased from 20% to 80%, the Macro-F1 increases from 0.702 to 0.861; and the average neighborhood density of the atlas is increased from 2 to 8, the Macro-F1 increases from 0.702 to 0.863. After exceeding 80% of text completeness or average neighborhood density of 8, the performance of the annotations is better than the 80% text completeness or average neighborhood density. After 80% text completeness or 8 average neighborhood degree, the performance enters the plateau area. This result suggests that missing caption and description text can limit cultural semantic judgments, and that too sparse a neighborhood of the map can weaken parent topic co-occurrence and terminological constraints, but too dense a neighborhood does not provide sustained benefits. The coupled relationship between module ablation and inference cost is shown in Figure 9.

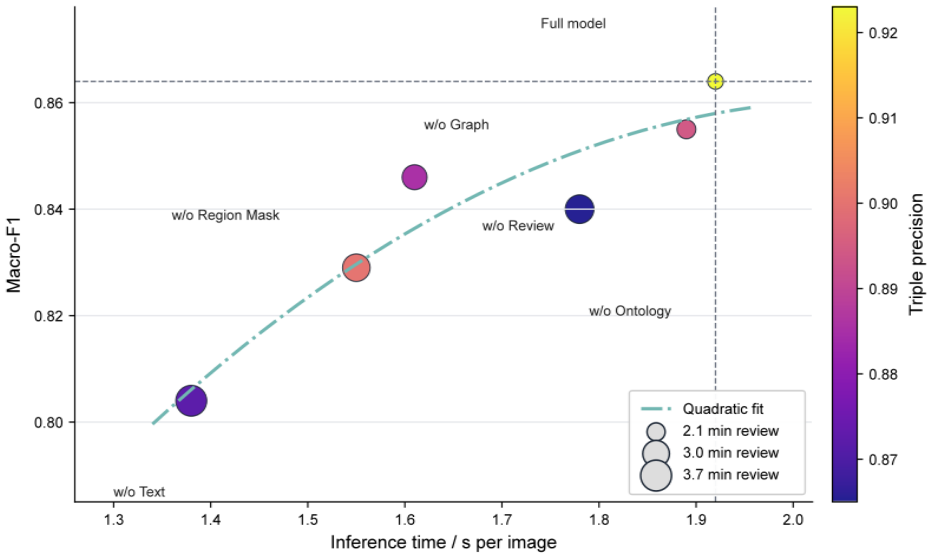

Figure 9: Module ablation coupled with inference cost.

In Figure 9, after removing the text module, the inference time of a single graph decreases from 1.92 s to 1.38 s, but the Macro-F1 decreases by 0.060, and the manual review time increases by 1.67 min. After removing the ontology constraints, the inference time only decreases by 0.14 s, and the triple accuracy decreases from 0.923 to 0.865, which indicates that the ontology constraints have a lower computational cost, but they can filter the relationships that do not fit into the semantic domains effectively. Relationships. After removing the region mask, mAP@0.5 decreases from 0.803 to 0.714, indicating that the local boundaries of the ink image still need to be processed at the region level. Module ablation, efficiency and cost of manual review are shown in Table 4.

Table 4: Module Ablation, Efficiency & Labor Review Costs

Variant	Macro-F1	mAP@0.5	Triple Accuracy	Single Image Inference Time (s)	Manual Review Time (min)
Full Model	0.864	0.803	0.923	1.92	2.05
w/o Text	0.804	0.762	0.872	1.38	3.72
w/o Region Mask	0.829	0.714	0.900	1.55	3.26
w/o Ontology Constraint	0.840	0.797	0.865	1.78	3.41
w/o Graph Context	0.846	0.783	0.885	1.61	2.94
w/o Uncertainty Review	0.855	0.796	0.894	1.89	2.31
Manual Only	—	—	0.941	—	8.70

In Table 4, the text module has the greatest impact on Macro-F1, the region mask has the greatest impact on mAP@0.5, and the ontology constraints have the greatest impact on ternary accuracy. Regarding the manual review time, the average is 2.05 min for Full model and 8.70 min for Manual only, which is a 76.4% decrease in review time. This result indicates that the AI-assisted process mainly reduces the initial screening and evidence aggregation time, and the final judgment of the cultural semantic layer still needs expert verification.

3.3 Knowledge Graph Quality, Case Interpretation and Error Diagnosis

The semantic annotation results have the value of cross-collection search and parent genealogy analysis only when they enter the queryable knowledge structure. This section evaluates the entity scale, triad accuracy, querying capability and case interpretation ability of WTBR-KG and contrasts its application focus with existing cultural heritage multimodal mapping studies.CICHMKG is oriented towards non-heritage multimodal knowledge organization [30], traditional opera multimodal mapping emphasizes ontology modeling and emotion and repertoire identification [31], GAT-based cultural heritage knowledge mapping emphasizes relationship extraction [32], and image-driven batik product knowledge mapping emphasizes the combination of visual patterns and knowledge structures [33]. Compared with these studies, WTBR-KG has a narrower scope of objects, but a finer binding between regional evidence, inscription evidence and cultural semantics.The results of WTBR-KG quality assessment and query validation are shown in Table 5.

Table 5: WTBR-KG Quality Assessment and Query Validation Results

Metric	Value	Description
Entity Nodes	4982	Works, people, areas, motifs, concepts, sources, etc.
Triples	24615	Includes three types of sources: automatically generated, terminology mapping, and manual review
Relationship Predicates	34	Examples include depictsMotif, hasMorphology, evokesMeaning, etc.
Metadata Triple Accuracy	0.976	Accuracy of fields such as author, date, collection source, etc.
Object Triple Accuracy	0.921	Accuracy of relationships regarding objects like dead wood, bamboo, stone, etc.
Morphological Triple Accuracy	0.889	Accuracy of terms like dry brush, texture strokes, density of bamboo leaves, etc.
Semantic Triple Accuracy	0.842	Accuracy of concepts like ancient meaning, reclusiveness, clarity, etc.
Evidence Source Coverage Rate	0.954	Coverage rate of the relationship supportedBySource
Capability Query Accuracy	0.893	Accuracy of 30 SPARQL capability questions verified manually
Median Query Time	0.38 s	In a local database environment

In Table 5, the metadata triad has the highest precision of 0.976; the semantic triad has the lowest precision of 0.842. The difference is consistent with the nature of the task. The author, date and collection sources are mostly from structured fields, and the errors mainly come from missing collection records or translation differences; the cultural semantics need to be combined with images, inscriptions and context of the work, and the sources of errors are more complex. The accuracy of the query for the competency question is 0.893, which indicates that the mapping can support more stable cross-collection search. Typical queries include "searching for works that contain dead trees and strange stones and have ancient semantics in the inscriptions", "searching for works that contain bamboo and stones and have dry brushwork in the morphological layer", "searching for works that have multiple occurrences of the semantics of solitude in the same author system", "searching for works that have multiple occurrences of the semantics of solitude in the same author system". ". The semantic annotation and local knowledge map of typical works are shown in Figure 10.

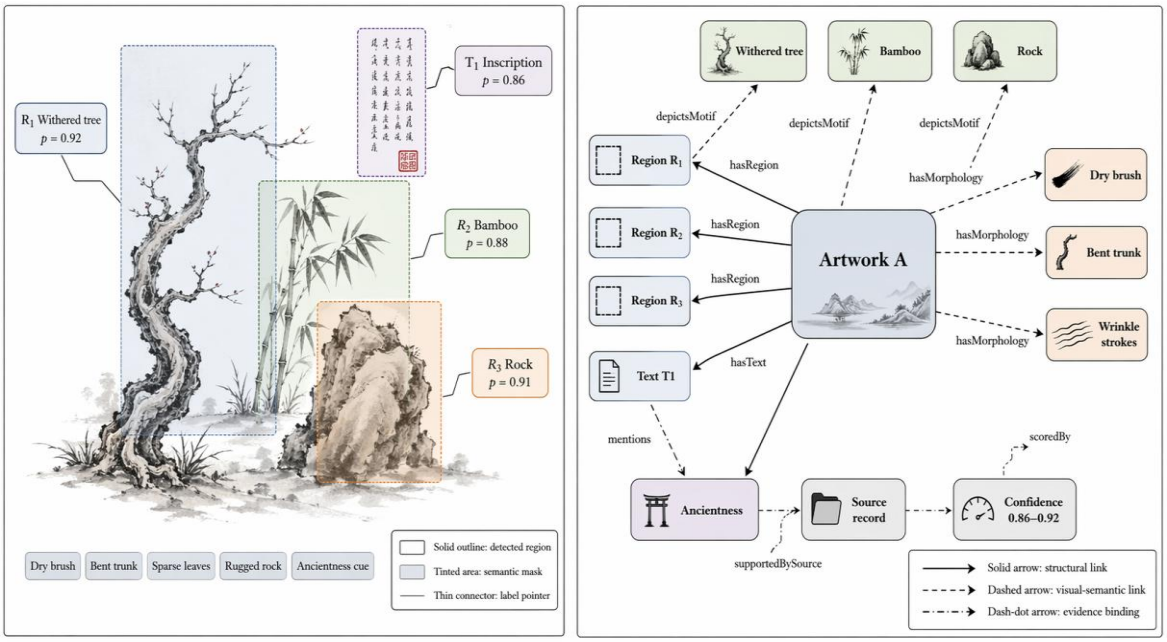

Figure 10: Example of semantic annotation and local knowledge graph of typical works.

In Figure 10, the work nodes are connected to the three themes of dead wood, bamboo and stone through depictsMotif. The dead-wood region is linked with "branches that bend" and "dry brush that cracks", the bamboo region is linked with "thin leaves" and "leaning outwards", and the stone region is linked with "oddness". The bamboo region connects with "thin leaves" and "slanting outward", and the stone region connects with "odd stone contour" and "thick cracked lines". The inscription node connects "ancient meaning" through mentions, and the semantic judgment node connects both the image area and the text fragment through supportedBySource. This structure lets "the work possesses ancient meaning" become one retrospective judgment, and the question raiser can look at the corresponding region, text proof, and confidence degree.

The error diagnosis shows that there are three main types of problems with the model. First, incomplete inscription text reduces cultural semantic recall. Some open collections only provide English summaries, lacking original inscriptions or accurate transcriptions, and the model can only rely on inscriptions and images to make judgments, leading to the omission of labels such as "yiqi" and "hanyi". Secondly, the region segmentation is unstable at the weak boundary of the ink. Withered branches, sloping banks, and chapped lines on stone surfaces share the texture of dry brushwork, and candidate regions are prone to merging or breaking. Third, high-frequency cultural semantics generate a priori bias. When the bamboo and stone are co-occurring, the model is prone to give "modesty" or "openness", but if the inscription emphasizes the contexts of social intercourse, elegant gathering, or chronicle of travel, such high-frequency semantics will obscure the meaning of more specific works.

From the deployment point of view, WTBR-KG is more suitable for three kinds of tasks. First, it can be used to batch generate candidate labels and evidence chains in collection organization, reducing the time for manual reading from zero. Secondly, it can be used in academic research to search for the main topic and discover the relationship, for example, comparing the semantics of the inscription "Dead Wood and Strange Stones" in different collections. Thirdly, in digital exhibition, we can connect individual works to similar themes, related characters and semantic concepts, so that the audience can enter the knowledge reading with evidence path from image browsing.

4 Conclusion

This paper constructs a multimodal semantic annotation and knowledge graph generation method around the parent theme of literati paintings of dead wood, bamboo and stone. Based on open collection images, titles, inscriptions, collection metadata and expert labels, the study refines dead wood, bamboo and stone from generic visual objects into object, morphological and cultural semantic layers, and transforms the model outputs into knowledge graph relations with sources and confidence levels.

(1) This paper establishes the caliber of WTBR-LP dataset and a three-layer labeling system. The dataset contains 1,184 works, 1,768 images, 3,024 paragraphs of text and 12,436 candidate regions, covering dead wood, bamboo, stone, sloping banks, inscriptions, seals, and their morphological and semantic labels, which provide reviewable objects for the calculation of localized matrices of literati paintings.

(2) This paper proposes an annotation model in which image regions, text, metadata and map neighborhoods are jointly involved. In the pre-experiment, the Macro-F1 of this paper reaches 0.864, the mAP@0.5 reaches 0.803, the Hits@1 reaches 0.912, and the ternary precision reaches 0.923. Compared with the purely manual annotation, the AI-assisted process reduces the average time of review from 8.70 min to 2.05 min.

(3) In this paper, we construct a WTBR-KG knowledge graph to form a queryable relationship among works, characters, motifs, ink patterns, inscription texts, cultural semantics, and curatorial sources. The current limitations mainly come from the missing inscription text, the weak boundary segmentation error of ink and wash, and the long-tailed distribution of cultural semantics. Follow-up studies can expand the high-quality inscription transcription data, introduce finer pen and ink labels, and carry out cross-domain robustness validation across different collections, eras, and painting schools.

About the Author

Cao Jikang was born in Linyi City, Shandong Province, P. R. China in 1994. He received his Bachelor's degree and Master's degree from Xi'an Academy of Fine Arts, and his Doctoral degree (PhD) from Thonburi University, Bangkok, Thailand. He is currently a faculty member of the School of Art and Design, Jiangxi Institute of Technology, with his main research direction focusing on fine arts education.

References

[1] Sturman, P. C. (2022). Inscriptional practices of the Song literati: Revisiting Su Shi's Old Tree, Rock, and Bamboo. Archives of Asian Art, 72(1), 75-95.

[2] Ahn, K. (2023). Becoming bamboo: Reassessing Su Shi's painting theory from Deleuze's angle. The Philosophical Forum, 54(3), 161-184.

[3] Zhang, W., Zhang, J. W., Wong, K. K., et al. (2024). Computational approaches for traditional Chinese painting: From the "Six Principles of Painting" perspective. Journal of Computer Science and Technology, 39(2), 269-285.

[4] Liong, S. T., Huang, Y. C., Li, S., et al. (2020). Automatic traditional Chinese painting classification: A benchmarking analysis. Computational Intelligence, 36(3), 1183-1199.

[5] Wan, J., Zhang, H., Zou, J., et al. (2024). WuMKG: A Chinese painting and calligraphy multimodal knowledge graph. Heritage Science, 12, 159.

[6] Wan, J., Chen, S., Zeng, Q., et al. (2026). A multi-path fusion with knowledge augmentation framework for multimodal NER in Chinese painting. npj Heritage Science, 14, 265.

[7] Liang, W., De Meo, P., Tang, Y., et al. (2024). A survey of multi-modal knowledge graphs: Technologies and trends. ACM Computing Surveys, 56(11), Article 273.

[8] Rei, L., Mladenić, D., Dorozynski, M., et al. (2023). Multimodal metadata assignment for cultural heritage artifacts. Multimedia Systems, 29, 847-869.

[9] International Image Interoperability Framework Consortium. (2023). IIIF Image API 3.0. IIIF.

[10] The Metropolitan Museum of Art. (2026). Open access at The Met. The Met.

[11] The Cleveland Museum of Art. (2026). Open access API. Cleveland Museum of Art.

[12] Harvard Art Museums. (2026). Harvard Art Museums API. Harvard Art Museums.

[13] Smithsonian Institution. (2026). Smithsonian Open Access. Smithsonian Institution.

[14] Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning (Vol. 139, pp. 8748-8763).

[15] Liu, S., Zeng, Z., Ren, T., et al. (2024). Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In European Conference on Computer Vision (pp. 38-55).

[16] Kirillov, A., Mintun, E., Ravi, N., et al. (2023). Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 4015-4026).

[17] Li, J., Li, D., Savarese, S., et al. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the 40th International Conference on Machine Learning (Vol. 202, pp. 19730-19742).

[18] Liu, H., Li, C., Wu, Q., et al. (2023). Visual instruction tuning. In Advances in Neural Information Processing Systems (Vol. 36, pp. 34892-34916).

[19] Bekiari, C., Bruseker, G., Canning, E., et al. (2024). Definition of the CIDOC Conceptual Reference Model, Version 7.1.3. CIDOC CRM Special Interest Group.

[20] Getty Research Institute. (2026). Art & Architecture Thesaurus. Getty Vocabularies.

[21] Linked Art. (2026). Linked Art data model. Linked Art Community.

[22] Cyganiak, R., Wood, D., & Lanthaler, M. (2014). RDF 1.1 concepts and abstract syntax. W3C Recommendation.

[23] Miles, A., & Bechhofer, S. (2009). SKOS simple knowledge organization system reference. W3C Recommendation.

[24] Knublauch, H., & Kontokostas, D. (2017). Shapes Constraint Language. W3C Recommendation.

[25] Harris, S., & Seaborne, A. (2013). SPARQL 1.1 query language. W3C Recommendation.

[26] Bordes, A., Usunier, N., Garcia-Duran, A., et al. (2013). Translating embeddings for modeling multi-relational data. In Advances in Neural Information Processing Systems (Vol. 26, pp. 2787-2795).

[27] Sun, Z., Deng, Z. H., Nie, J. Y., et al. (2019). RotatE: Knowledge graph embedding by relational rotation in complex space. In International Conference on Learning Representations.

[28] Schlichtkrull, M., Kipf, T. N., Bloem, P., et al. (2018). Modeling relational data with graph convolutional networks. In The Semantic Web: ESWC 2018 (pp. 593-607).

[29] Veličković, P., Cucurull, G., Casanova, A., et al. (2018). Graph attention networks. In International Conference on Learning Representations.

[30] Fan, T., Wang, H., & Hodel, T. (2023). CICHMKG: A large-scale and comprehensive Chinese intangible cultural heritage multimodal knowledge graph. Heritage Science, 11, 115.

[31] Fan, T., Wang, H., & Hodel, T. (2023). Multimodal knowledge graph construction of Chinese traditional operas and sentiment and genre recognition. Journal of Cultural Heritage, 62, 32-44.

[32] Wang, Y., Liu, J., Wang, W., et al. (2024). Construction of cultural heritage knowledge graph based on graph attention neural network. Applied Sciences, 14(18), 8231.

[33] Wu, X., Yuan, Q., Qu, P., et al. (2025). Image-driven batik product knowledge graph construction. npj Heritage Science, 13, 20.