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.

Instance-driven knowledge organization linking temple mural figures and evidence across sites

Liu Zherong
Beijing Institute of Technology
Qiao Yu
qiaoyu@bit.edu.cn

Beijing Institute of Technology
Zhang Zhihua
Shanxi Cultural Relics and Museum Industry Group Co., Ltd

Article

Keywords:

Posted Date: November 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11592497/v1

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

Additional Declarations: No competing interests reported.

Instance-driven knowledge organization linking temple mural figures and evidence across sites

Liu Zherong[1], Qiao Yu[2, *], Zhang Zhihua[3]

[1] Beijing Institute of Technology, School of Design and Arts, Beijing, China

31202666293@bit.edu.cn

[2] Beijing Institute of Technology, School of Design and Arts, Beijing, China

* Corresponding author: Qiao Yu; email: qiaoyu@bit.edu.cn

[3] Shanxi Cultural Relics and Museum Industry Group Co., Ltd., Department of Digital Cultural Heritage Preservation, Shanxi, China

zhihuazhang005@163.com

Abstract

Temple mural images and interpretations remain dispersed across sites, limiting regional comparison. Research centred on individual temples makes it difficult to connect the diverse representations of shared deities and legends across time and space, limiting understanding of regional cultural patterns and their historical development. Here we organize materials from 25 temples in southeastern Shanxi, China, using localized figure instances to link images, spatial records and textual evidence. Images from 15 temples support model training and validation; the trained model then extracts figure instances from the remaining 10 temples. Experts establish image–text correspondence, and canonical semantic objects connect shared figures and stories while retaining local descriptions and sources. With candidate regions held fixed, domain fine-tuning increases segmentation F1 from 72.2% to 79.4%. The resulting knowledge graph contains 3,433 nodes and 4,178 relations. A Guan Yu case study and image- and story-initiated queries illustrate cross-temple retrieval and source tracing, providing an evidence-linked resource for comparing local representations of shared cultural objects.

Introduction

Temple murals are important sources for understanding religion, folk customs and local beliefs. Their figures and stories record how different communities understand shared cultural objects and express local narrative traditions. A geographically bounded collection of temples encompasses diverse cultural elements from different periods, belief systems and cultural backgrounds. The same deity and associated legends acquire distinct cultural representations in different historical and regional contexts [1]. Temples within a region differ in their historical backgrounds and pictorial arrangements [2], while also sharing figures [3], stories [4] and local cultural practices [5]. These similarities and differences concern both the interpretation of individual temples and the distribution, combination and local expression of shared objects across a region. However, figure images, textual accounts and supporting evidence remain dispersed across temple records. Information organization and case studies centred on individual temples make it difficult to examine relationships among these representations systematically, limiting cumulative comparison and an understanding of regional cultural patterns and their historical development.

Systematic research on regional culture involves identifying shared cultural objects and distinguishing the specific content and interpretative levels of their local representations. Sun et al. identified shared themes and differences in content through the analysis of murals across styles [6]. Sartini et al. used the ICON ontology to distinguish visible depictions, iconographical subjects and cultural meanings [7]. These studies provide a basis for theme identification and image interpretation. However, images and scholarly accounts of the diverse representations acquired by the same cultural object in different historical and local contexts remain dispersed across separate records, making systematic comparison at the regional scale difficult. Can an approach to information organization that accommodates both cultural commonalities and local differences bring these dispersed representations and their

interpretative evidence together, thereby supporting systematic research on regional cultural patterns and their historical development?

Regional heritage research uses iconographical comparison, spatial analysis and relational modelling to examine the cultural content of different sites and the connections among them. Gesterkamp combined murals, stele inscriptions, painting criticism and ritual texts to compare the pictorial schemes and ritual arrangements of Heavenly Court paintings at different sites [8]. At the regional scale, Li used GIS to examine the spatiotemporal distribution of grotto temples in Henan in relation to topography, rivers, transport and political and cultural contexts [9]. For heritage classification, Kang et al. integrated spatial adjacency, architectural configuration and institutional affiliation to identify church types [10], whereas He and Liu combined a minimum cumulative resistance model with spatial network analysis to construct and classify heritage corridors sharing cultural themes [11]. To represent relationships among sites, Li et al. used a Mazu cultural knowledge graph to describe temple networks, shared symbols and local practices, examining the coexistence of shared traditions and local differences [12]. Fang and Han incorporated the spatial clustering of intangible cultural heritage resources into a knowledge graph of the Yellow River Basin, using semantic queries to organize cultural entities and their relations [13]. These studies provide perspectives on image interpretation, regional classification and heritage networks, but comparing the depictions of specific figures and stories across temple murals and the evidence supporting their interpretation still calls for observations at the level of individual figure images.

Image processing provides a technical basis for describing cultural images, localizing objects and extracting details. Abgaz et al. combined computer vision and semantic web technologies to enrich cultural heritage images semantically, using image content to describe and retrieve cultural information [14]. For image region extraction, He et al.'s Mask R-CNN jointly

predicts bounding boxes and instance masks, allowing objects within the same image to be localized separately [15]. Liu et al.'s Grounding DINO localizes objects using category names or language descriptions [16], while Kirillov et al.'s SAM generates region masks from prompts such as points and boxes [17]. For mural details, Cao et al. used multiscale features and dual attention to improve the segmentation of blurred boundaries and small objects [18], and Zhang et al. performed fine-grained segmentation of regions such as figure faces and headwear [19]. Beyond object extraction, Fu et al. used cross-modal prompt tuning for few-shot mural classification [20], Zou et al. localized damaged regions using an improved YOLACT model [21], and Baek et al. clustered paintings using multimodal features to compare their colour and texture characteristics [22]. These techniques describe contours, categories or formal features; establishing figure identities and narrative roles still depends on complete scenes and textual evidence.

Knowledge organization uses ontologies and knowledge graphs to describe relationships among heritage objects, image and textual materials, and interpretative evidence. Fan et al. constructed CICHMKG by organizing intangible cultural heritage texts and images around shared cultural entities [23]. Wan et al. integrated dispersed painting and calligraphy resources in WuMKG and implemented cross-modal retrieval [24]. Object locations and structures have also been incorporated into knowledge representation: Yang and Hou used a domain ontology to describe grotto components and spatial relations [25], and Yang et al. integrated the geometry, semantics and attributes of Buddhist niches through a CityGML extension [26]. For image meaning, Carboni and de Luca distinguished represented objects from acts of interpretation using the VIR ontology [27], while Schimmenti et al. separately represented entities, hypotheses and evidence in scholarly debates on cultural heritage [28]. To address record quality, Bobasheva et al. identified errors and omissions in metadata [29], and Homburg et al. documented the acquisition and processing of 3D data [30]. Applying

these representations to murals across temples still calls for correspondence among specific figure images, their original locations and interpretative records, so that shared names do not obscure differences in local representations or the grounds for their interpretation.

Existing studies provide approaches to image extraction, cultural interpretation and information association. Here we examine their integration through localized figure instances for evidence-linked comparison across temple murals. Three problems have yet to be addressed systematically: extracting and locating individual figure instances under damaged conditions; tracing them through their source murals, halls and temples and establishing correspondence with textual evidence; and forming semantic associations across temples while retaining the supporting sources. Using materials from 25 temples in southeastern Shanxi, we develop the figure extraction workflow with images from 15 temples and apply the trained model to the remaining 10 temples. Candidate localization and boundary refinement produce figure instances whose positions are retained in the original images. Experts then establish image–text correspondence using complete mural walls, spatial records and textual materials. Canonical semantic objects are established for object types such as figures, visual representations and stories, organizing temple records into a queryable knowledge graph. Graph records retain original mural contexts, original designations, object types, grounds for associations and the sources of both records. Queries can therefore trace figure images to identity interpretations and cross-temple associations and return the corresponding wall locations and documentary sources for retrieval, comparison and evidence checking across cultural representations.

Methods

2.1 Overview of the method

This study comprises figure instance extraction, within-temple image–text correspondence and cross-temple semantic association.

The methodological framework is defined as follows:

$$\mathrm{F} = (L_E, L_P) \quad (1)$$

Within the information layers, L_E is the multisource evidence layer, which organizes and preserves image, textual and spatial materials; L_P is the figure instance layer, which stores figure regions extracted from murals, instance identifiers and their source locations. These two types of information are linked through correspondence between figures and original images, temple spaces and textual content.

For temple t, its multisource evidence and figure instances are represented as follows:

$$L_E^{(t)} = \mathcal{I}^{(t)} \cup \mathcal{T}^{(t)} \cup \mathcal{S}^{(t)}, \qquad L_P^{(t)} = f_{\mathrm{ext}}\left(\mathcal{I}^{(t)}\right) \quad (2)$$

Here, $\mathcal{I}^{(t)}$, $\mathcal{T}^{(t)}$ and $\mathcal{S}^{(t)}$ denote the image, textual and spatial evidence sets for the temple, respectively, and f_{ext} is the figure instance extraction function.

The connections between $L_E^{(t)}$ and $L_P^{(t)}$ within a temple form a temple-specific knowledge collection:

$$U_t = \left(L_E^{(t)} \cup L_P^{(t)},\ R_{\mathrm{intra}}^{(t)}\right) \quad (3)$$

Here, $R_{\mathrm{intra}}^{(t)}$ is the set of within-temple relations, including spatial containment, image provenance, correspondence between figure instances and textual descriptions, and evidential links supporting relevant interpretations.

After temple-specific image–text correspondence has been confirmed, records from different temples are associated through canonical figures, stories, themes and objects of worship. Each association retains the shared object, its grounds and the sources of both records.

The final knowledge graph is represented as follows:

$$\mathrm{G} = (\{U_t\}_{t=1}^{n}, R_{\mathrm{cross}}) \quad (4)$$

Here, $U_1, U_2, \ldots, U_n$ denote the knowledge collections of the individual temples. Each collection already contains its multisource evidence, figure instances and internal relations. R_{cross} represents evidence-supported cross-temple semantic associations between these temple-specific collections, and n is the total number of temples.

Fig. 1 summarizes the workflow, distinguishing the multisource evidence and figure instance layers from the temple-specific collections and the final graph.

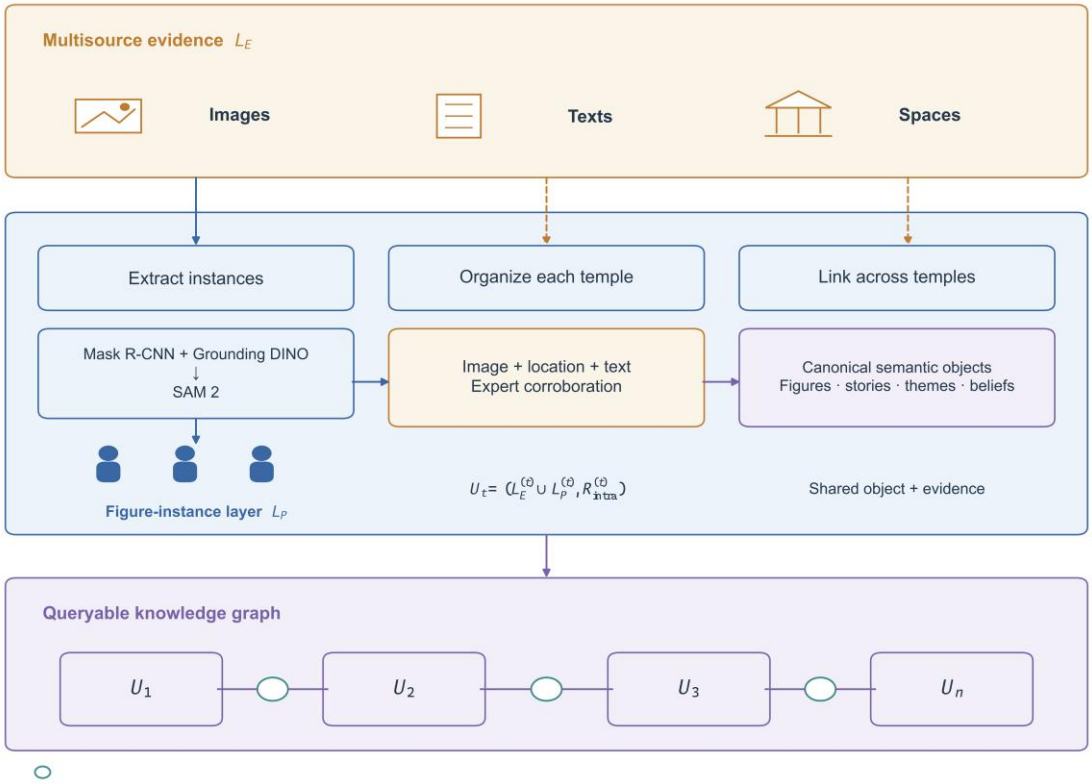


Fig. 1 | Overview of the instance-driven knowledge organization framework. Multisource evidence and figure instances are linked within each temple. Canonical semantic objects connect temple-specific knowledge sets, preserving access to local evidence. Open circles denote shared canonical semantic objects linking temple-specific knowledge collections.

2.2 Data sources

The source materials cover 25 temples in southeastern Shanxi and include mural images, spatial records and textual materials, indexed by temple. The image data have two roles: images from 15 temples are used for model training and validation, whereas images from the remaining 10 temples are processed with the trained model for figure extraction. These 10 temples are not used for model training or parameter selection. The knowledge graph integrates records across the full 25-temple collection.

First, images were acquired using a Sony A7M4 camera equipped with an FE 90 mm f/2.8 Macro G OSS lens, following the principles of non-contact, non-destructive recording and preservation of the original appearance. Photographs were taken sequentially with

approximately 50% overlap to capture both local details and complete wall surfaces. The photographs were stitched in Adobe Photoshop using overlapping regions. After checking for misalignment, duplication and missing areas, the original photographs and complete wall images were stored separately.

Second, spatial records are based on field surveys and verification against documentary materials. They record hall layouts, wall positions, mural distributions and preservation conditions and are also used to determine the photographic coverage. Records are archived by temple, hall and wall and linked to image indexes.

Finally, textual materials are collated by temple and wall and grouped into three categories: heritage and architecture; mural studies; and religion, folklore and stories. They provide spatial facts, image descriptions and cultural interpretations while retaining their sources and scope of applicability.

2.3 Textual data processing

Textual processing uses the temple as the contextual boundary and divides continuous narratives into knowledge units that can be located and supported by evidence. Facts, descriptions and interpretations are distinguished, after which correspondence is established according to objects, spaces and cultural content. Sources and degrees of certainty are retained throughout the organizational process (Fig. 2; Table 1).

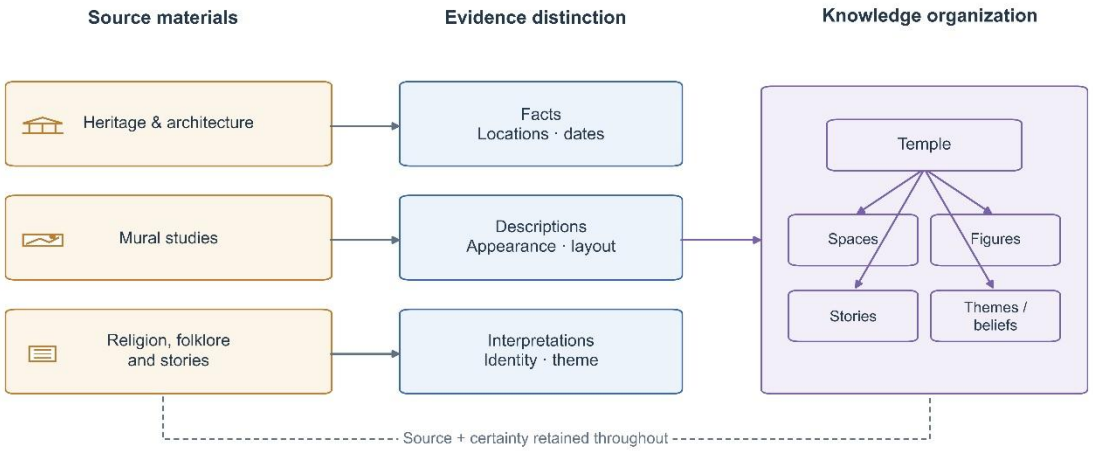


Fig. 2 | Organization of textual evidence. Sources are grouped by research function and differentiated into facts, descriptions and interpretations. Records are then organized around spaces, figures, stories, themes and beliefs within each temple, retaining their provenance and certainty.

Table 1 | Functions of textual materials and approaches to knowledge organization

Material category	**Main sources**	**Content to be confirmed or interpreted**	**Organization within the evidence collection**
Heritage and architectural materials	Heritage surveys, local histories and gazetteers, temple and architectural studies, inscriptions and stelae	Temple locations, dates, historical development, hall composition and wall relationships	Separate temple facts from spatial descriptions; retain temporal scope, location hierarchy and sources
Mural research materials	Studies of mural art, existing digital archives and related image descriptions	Mural layouts, visible figure features, compositional arrangements and thematic judgments	Organize image, figure, story and thematic descriptions separately; retain the objects to which judgments refer and the supporting evidence
Religious, folkloric and story materials	Religious and folklore studies, story sources and related cultural research	Figure identities, story motifs, religious backgrounds and cultural interpretations	Record figures, stories, beliefs and interpretations at distinct levels; retain original degrees of certainty and sources

Spatial hierarchies, visible figure forms and cultural interpretations are stored separately and linked through the same knowledge unit. A knowledge unit is jointly defined by its object, location, content, source and degree of certainty:

$$u_i = (o_i, l_i, c_i, p_i, q_i) \quad (5)$$

Here, o_i, l_i, c_i, p_i and q_i denote the object, spatial location, content, source evidence and degree of certainty, respectively. During organization, repeated narratives are removed, while missing information and the original qualifying expressions are retained.

2.4 Image processing and annotation

Figures vary considerably in scale within complete murals. To preserve figure contours, images are first divided into tiles using overlapping windows, and figures in the 15-temple model-development collection are annotated for training and validation. Images from the remaining 10 temples undergo the same tiling and source indexing before model inference. The tiles retain original image coordinates and spatial provenance.

Image identifiers combine four types of encoded information: the temple code C_t, spatial code C_l, image sequence N_i and tile sequence N_p, with indexes recording the original paths, image dimensions and tile positions.

$$ID = C_t \| C_l \| N_i \| N_p \quad (6)$$

Following tiling and source encoding, annotation is performed at the individual-figure level, recording contours with polygons and retaining the category, bounding box, area and image reference. Categories are standardized against the original images, and empty masks and invalid boxes are removed.

Within the 15-temple model-development collection, annotated images and their annotations are assigned together to training and validation subsets using a fixed random seed. Model

calibration is performed within this collection. The trained model is subsequently applied to images from the remaining 10 temples without further training or parameter selection.

2.5 Model development

To address occlusion, damage and overlapping contours in mural figures, this study uses Mask R-CNN [15] to generate candidates and introduces Grounding DINO [16] into its candidate generation workflow to assist detection. SAM 2 [31] is used to refine figure boundaries (Fig. 3). Manual figure annotations provide supervision for fine-tuning the last three stages of the Mask R-CNN backbone network, the feature pyramid, the region proposal network and the instance prediction heads, while the remaining backbone stages are frozen. The optimization objectives comprise figure classification, bounding box regression and mask prediction, together with objectness classification and bounding box regression in the region proposal network. After training, the model outputs candidate bounding boxes, confidence scores and initial masks.

Grounding DINO is not trained separately; its pretrained parameters remain fixed, and it supplements figure candidates in the same mural image using the "person." prompt. After the two candidate sets are combined, they are filtered by their respective confidence scores. Duplicate boxes are removed through non-maximum suppression (NMS), producing the input regions for SAM 2. Grounding DINO supplements figure localization, and SAM 2 refines figure boundaries. Mask R-CNN parameters are also fixed after fine-tuning, so candidate generation and subsequent mask learning are performed separately.

SAM 2 generates individual masks from bounding boxes and foreground points. Only its prompt encoder and mask decoder are fine-tuned, while the image encoder, memory module and remaining parameters are frozen. Training uses a mixture of bounding box, point and coarse mask prompts and perturbs candidate bounding boxes to accommodate localization errors.

Binary cross-entropy (BCE), Dice loss and intersection over union (IoU) regression constrain pixel classification, region overlap and quality estimation, respectively. Inference uses bounding boxes plus foreground points and outputs masks, instance identifiers and source locations.

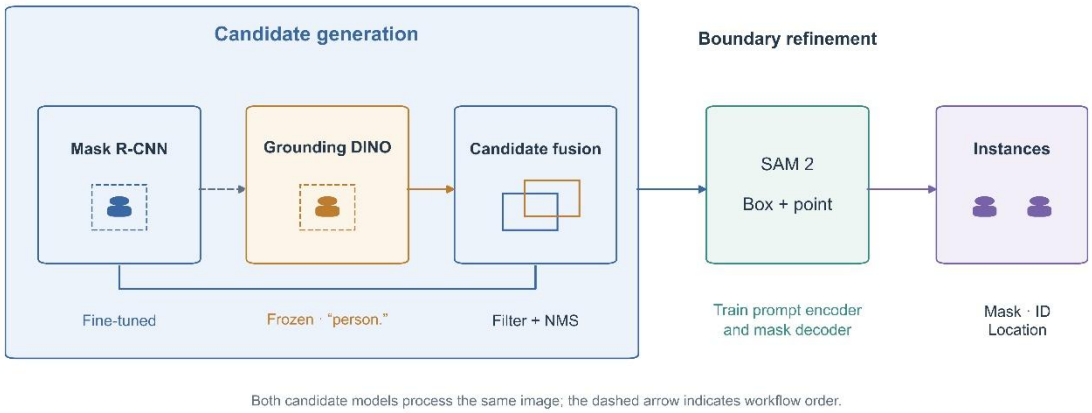


Fig. 3 | Model roles in figure instance extraction. Mask R-CNN supplies supervised candidates, Grounding DINO supplements localization, and NMS merges the candidates. SAM 2 refines figure boundaries using box-and-point prompts. Candidate models remain frozen during SAM 2 fine-tuning.

2.6 Establishing temple-specific knowledge

Temple-specific knowledge is established using the temple as the contextual boundary. The figure instances extracted in Section 2.5 and the textual knowledge units formed in Section 2.3 are organized into a single information collection. Instances are linked to complete murals, walls, halls and temples through image identifiers and indexes, while textual records are linked to the corresponding entries according to their objects and locations. Eight researchers in the relevant discipline independently assessed image–text correspondence and resolved disagreements through discussion at a subsequent meeting and classify the evidence into three confidence levels(Table 2). Confirmed correspondence and its source links constitute within-temple relations; together with multisource evidence and figure instances, these form a temple-specific knowledge collection. Medium-confidence entries retain an inference label and

supporting evidence. Low-confidence entries retain the image, location and description, with the identity marked as undetermined.

Table 2 | Three confidence levels for image–text correspondence and their subsequent processing

Confidence level	Judgment criteria	Typical example	Subsequent processing
High	Location, features and textual records are consistent, allowing explicit correspondence	Leftmost figure in the lower row of the left group on the rear wall of Shishi Village West Temple: the Tiger-Subduing Arhat, leaning on a staff and resting a hand on a tiger	Include after expert verification and retain the source
Medium	The figure grouping or narrative plot supports an identity inference, but explicit correspondence is still lacking	Military figure on the main wall of Xishe Weigong Temple: inferred to be Li Weigong based on the temple context	Retain the inference label and evidence, pending further corroboration
Low	Only localization or recognition of the general figure is possible; identity evidence is insufficient	First arhat on the north wall of Zhangzhuang Guanyin Hall: incomplete figure with an unknown identity	Retain the location and description and mark the identity as undetermined

2.7 Establishing cross-temple relations and constructing the knowledge graph

The graph first connects figure instances, spatial locations and textual content within each temple, then links each temple's knowledge units to a unified system of canonical semantic objects. Entities are distinguished as figures and images, temple spaces, stories and themes, objects of worship, and source materials.

Semantic normalization is performed across records from all temples by checking names, object types and content separately. Alternative names supported by the materials are assigned to the same canonical semantic object, while the original designations are retained. Figures, deities, visual representations, narratives, themes and beliefs receive separate identifiers, and objects of different types are connected through evidence-supported semantic relations (Table 3). Cross-temple semantic associations apply to any pair of distinct temples: knowledge units that refer to the same canonical object can be connected through it, retaining the temple,

location in the source materials and judgment evidence. Source-supported semantic relations between objects of different types are stored separately.

Table 3 | Semantic object normalization and cross-temple association rules

Association category	Normalization and association procedure	Evidence retained
Figures	Check designations and figure judgments, and link entries explicitly referring to the same figure to the same canonical semantic object.	Preserve original designations, figure descriptions and sources of judgments.
Story motifs	Confirm story names from specific plots, and link entries recording the same story to a shared object.	Preserve the sources of story descriptions, identifying inscriptions or research materials.
Themes and beliefs	Classify records separately by thematic system and belief system, and link records of the same type and meaning to a shared object.	Preserve the classification evidence and distinguish direct records from interpretative judgments.
Factual attributes	Use shared date, place-name, material or architectural form objects when the attribute type, comparison granularity and value are consistent.	Preserve original attribute records and their temporal or spatial scope.
Objects of different types	Retain figure, deity, visual representation, story and belief objects separately, connecting them through explicit semantic relations.	Preserve relation types, the evidence supporting their definitions and their sources for cross-level queries.

Within-temple relations, correspondence between knowledge units and canonical semantic objects, and relations between objects are organized into a graph structure:

$$G = (V, E, P),\ V = V_U \cup V_I \cup V_Z \cup V_D$$
$$E = E_U \cup E_I \cup E_Z \cup E_P \quad (7)$$

V_U, V_I, V_Z and V_D are the node sets for temple-specific objects, figure images, canonical semantic objects and source materials, respectively; E_U, E_I, E_Z and E_P are the relation sets for hierarchy, image correspondence, semantics and provenance, respectively. P is the set of properties, including names, locations, paths, evidence and degrees of certainty.

Knowledge graph construction creates nodes from image and textual instances, imports spatial hierarchy, image correspondence, semantic and provenance relations, and checks for duplicates by source node, relation type and target node. Figure images are linked to entries through

indexes, while cross-temple queries retrieve sources by traversing shared objects and semantic paths.

Results

3.1 Image processing and dataset preparation

This project used temple mural images from the Shangdang area of southeastern Shanxi as its source data. The acquired large-scale murals were tiled using 2,048 × 2,048-pixel windows with a 512-pixel overlap between adjacent windows. The final windows were aligned with the boundaries of the original image, allowing figures at tile edges to be checked against adjacent tiles.

Each mural tile was indexed to its original image, with the index recording the temple, wall, original image and window position. For example, BTHGDS001002 denotes the second tile of the first original image of the east gable wall at Beitianhe Guandi Temple (Table 4).

Table 4 | Scale of the mural and image resources from the 25 temples. The table describes the full resource collection; model training and validation use 15 temples, and model inference is subsequently performed on the remaining 10.

No.	Temple code	Temple name	Original images	Cropped images	Tiles containing figures
1	BTHG	Beitianhe Guandi Temple	5	475	130
2	SZZS	Sizhuang Zushi Temple	4	319	23
3	XSWG	Xishe Weigong Temple	10	377	61
4	NSYH	Nanshe Yuhuang Temple	16	1,392	102
5	DBCG	Dongbeichen Guandi Temple	3	312	44
6	SSCX	Shishi Village West Temple	3	1,280	81
7	MZYH	Mazhuang Yuhuang Temple	3	432	27
8	XPYD	Xiangpuyao Village Temple	2	531	54
9	ZZGY	Zhangzhuang Guanyin Hall	3	1,199	63

No.	Temple code	Temple name	Original images	Cropped images	Tiles containing figures
10	DXZG	Dongxianzi Guandi Pavilion	3	477	130
11	HCDX	Huangchuangdi Xiema Hall	3	762	82
12	MJYH	Majia Village Yuhuang Temple	15	752	179
13	GJGG	Guojiagou Guandi Temple	5	912	82
14	KHS0	Kaihua Temple	6	431	231
15	HDGD	Hedong Guandi Temple	5	267	49
Subtotal (15 temples)			86	9,918	1,338
16	YYGH	Yueyang Guanghua Temple	2	399	72
17	WHTW	Wanghuang Tangwang Temple	5	200	30
18	DZHS	Dazhonghan Sanzong Temple	20	1,962	180
19	CJYS	Chenjiayu Sanguan Temple	5	261	13
20	WXDY	Wuxiang Dayun Temple	8	695	51
21	HJY0	Hongji Monastery	3	231	85
22	HHGD	Hehe Guandi Temple	16	1,462	138
23	CYG0	Ciyun Palace	9	603	14
24	HTHG	Huaitan Guild Hall	1	408	10
25	ZCXG	Xigou Guandi Temple, Zhongcun	9	3,339	143
Total			164	19,478	2,074

Figures were annotated individually within each mural tile. Annotators drew polygons along figure contours in Label Studio and stored annotations for distinct figures in the same image separately (Fig. 4). The exported COCO data stored the polygon, bounding box, region area and image reference in a single instance record. During checking, the annotated regions were verified against the images, figure class labels were standardized, duplicate image records were merged and invalid annotations were removed.

Model development used annotated images from 15 temples, whereas images from the remaining 10 temples were reserved for application of the trained model. The manually

annotated dataset used for the model experiments contained 1,338 images and 4,198 figure instances. The dataset was split at the image level, with each image and all its annotations assigned together, yielding 1,200 training images and 138 validation images. Of these, 137 validation images had manual masks suitable for quantitative evaluation, and the remaining image was used for illustration. The image-level experimental subsets belong to the model-development stage and are distinct from the 10-temple inference collection.



Fig. 4 | Example of manual polygon annotation in Label Studio. Outlined figure regions are recorded as separate instances for training and validation.

3.2 Mask R-CNN training and SAM 2 domain fine-tuning

The project used a combined model in which Mask R-CNN and Grounding DINO provided two sets of candidates; after candidate fusion, SAM 2 generated individual masks (Fig. 5).

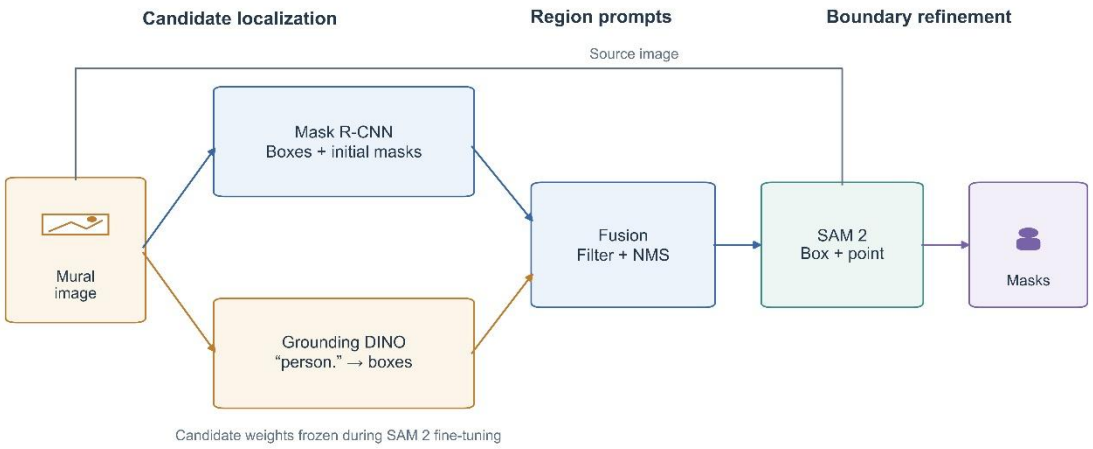


Fig. 5 | Inference workflow for figure instance extraction.

The figure instance extraction workflow uses two models to localize figures in the same image. After filtering and deduplication, the candidates are converted into box and point prompts, which SAM 2 uses with the original image to refine the boundaries. The figures and masks shown are schematic illustrations of the workflow.

Mask R-CNN used a ResNet-50-FPN architecture and was fine-tuned from COCO weights. Images were resized while preserving their original aspect ratios. The model jointly learned classification, bounding box localization and mask prediction, producing confidence scores, bounding boxes and initial contours.

The training and validation batch sizes for Mask R-CNN were both 1, the number of data-loading workers was 0, and the number of trainable backbone layers was 3. The model used stochastic gradient descent (SGD) with a momentum of 0.9 and weight decay of 1 × 10^-4. The 20 training epochs were divided into two stages: the learning rate was 0.001 for the first 10

epochs and was reduced to 0.0003 by StepLR for the remaining 10 epochs. CUDA automatic mixed precision and GradScaler were used, and the inference confidence threshold was 0.5.

During training, the total loss decreased from 0.6198 to 0.1316, and the mask loss decreased from 0.2891 to 0.0942. Classification, candidate bounding box regression and mask losses decreased overall. The final model output figure confidence scores, bounding boxes and initial instance masks (Table 5; Fig. 6).

Table 5 | Total loss and major component losses of Mask R-CNN for each training epoch.

Epoch	Total loss	Classification loss	Box loss	Mask loss	Time (s)
1	0.6198	0.1176	0.1774	0.2891	363.37
2	0.4500	0.0762	0.1457	0.2126	353.72
3	0.3652	0.0564	0.1205	0.1787	301.63
4	0.3138	0.0440	0.1009	0.1621	348.17
5	0.2810	0.0386	0.0884	0.1487	641.59
6	0.2551	0.0329	0.0778	0.1401	340.40
7	0.2395	0.0309	0.0712	0.1334	351.59
8	0.2238	0.0280	0.0650	0.1274	351.74
9	0.2081	0.0253	0.0582	0.1215	351.33
10	0.1954	0.0233	0.0529	0.1166	350.13
11	0.1712	0.0199	0.0403	0.1089	354.84
12	0.1590	0.0182	0.0341	0.1048	344.19
13	0.1525	0.0169	0.0309	0.1029	357.04
14	0.1482	0.0164	0.0289	0.1012	341.51
15	0.1447	0.0159	0.0272	0.0998	352.38
16	0.1412	0.0151	0.0260	0.0985	673.83
17	0.1386	0.0149	0.0248	0.0974	343.02
18	0.1359	0.0142	0.0239	0.0963	2713.98
19	0.1339	0.0142	0.0229	0.0953	351.32
20	0.1316	0.0140	0.0220	0.0942	584.30

(Note: The total loss includes classification, bounding box regression and mask losses, together with two region proposal network losses. The latter two are not listed separately in the table; therefore, the sum of the displayed components does not equal the total loss.)

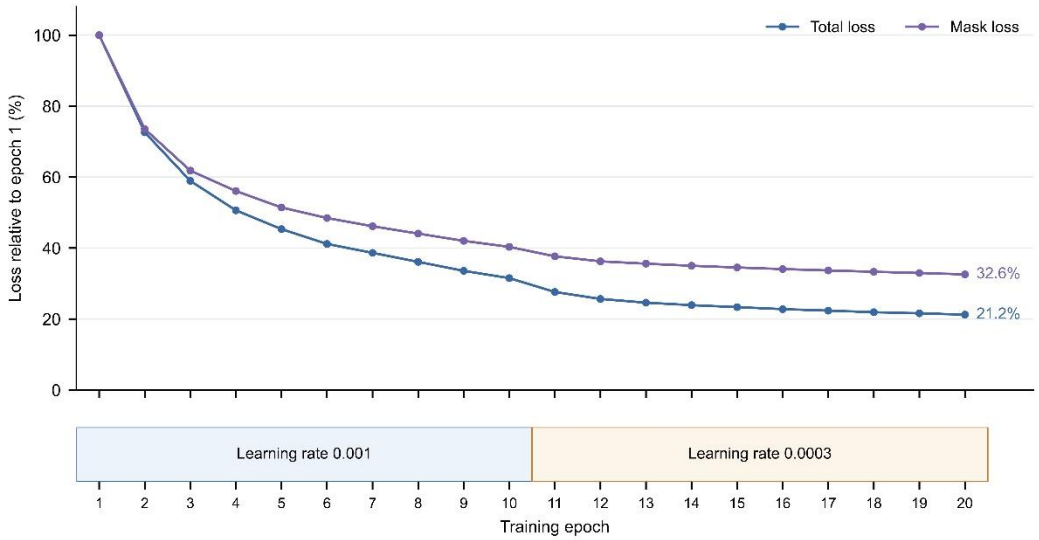


Fig. 6 | Mask R-CNN training loss across 20 epochs. Total loss and mask loss are expressed as percentages of their respective values at epoch 1. The learning rate is 0.001 for epochs 1–10 and 0.0003 for epochs 11–20. The final relative total and mask losses are 21.2% and 32.6%, respectively.

Candidate fusion was performed on the same image. The fine-tuned Mask R-CNN and Grounding DINO with fixed weights each generated candidates; the latter searched for figure regions using the "person." prompt. Candidates were first filtered by confidence, and non-maximum suppression (NMS) was then used to suppress candidates with highly overlapping bounding boxes, preventing repeated segmentation of the same figure. The retained candidate bounding boxes and prompt points were supplied together to SAM 2: prompt points for Mask R-CNN candidates were taken from within the initial masks, whereas Grounding DINO candidates used the centres of their bounding boxes.

During fine-tuning of SAM 2.1 Hiera Base+, the image encoder, memory modules and the rest of the backbone were frozen. Only the prompt encoder and mask decoder were updated, comprising 4,221,329 parameters. 1,080 images from 1,200 training images were used for fine-

tuning and 120 for calibration, both drawn from the original training set, with a random seed of 42. Three prompt types were used: candidate bounding boxes only, candidate bounding boxes with foreground points, and candidate bounding boxes with foreground points and coarse masks. They were mixed in approximate proportions of 25%, 50% and 25%, respectively, with candidate bounding box perturbations of up to 5%. The loss comprised binary cross-entropy (BCE), Dice and IoU quality regression terms. Training used AdamW for 2 epochs, with learning rates of 1×10^{-5} and 5×10^{-6}, respectively.

For parameter selection, 40 of the 120 calibration images were first selected for a broad grid search across 108 configurations. A refined search was then conducted on the calibration set, using end-to-end F1 score as the primary criterion and mean matched IoU and recall as secondary criteria. The final workflow used text prompts. In this workflow, both the Grounding DINO candidate bounding box score threshold and the frozen Mask R-CNN candidate bounding box confidence threshold were set to 0.5. The two sets of candidate bounding boxes were merged using non-maximum suppression with an IoU threshold of 0.5. Grounding DINO's internal text-matching threshold was separately fixed at 0.15. The calibration search covered the Grounding DINO candidate score threshold, SAM 2 prompt type, candidate bounding box expansion ratio and mask threshold. The Mask R-CNN and NMS thresholds were fixed at 0.5 and were not included in the grid search. The final SAM 2 configuration used candidate bounding box and foreground point prompts, 2% outward bounding box expansion and a mask threshold of 0.0. The comparison is shown in Fig. 7.

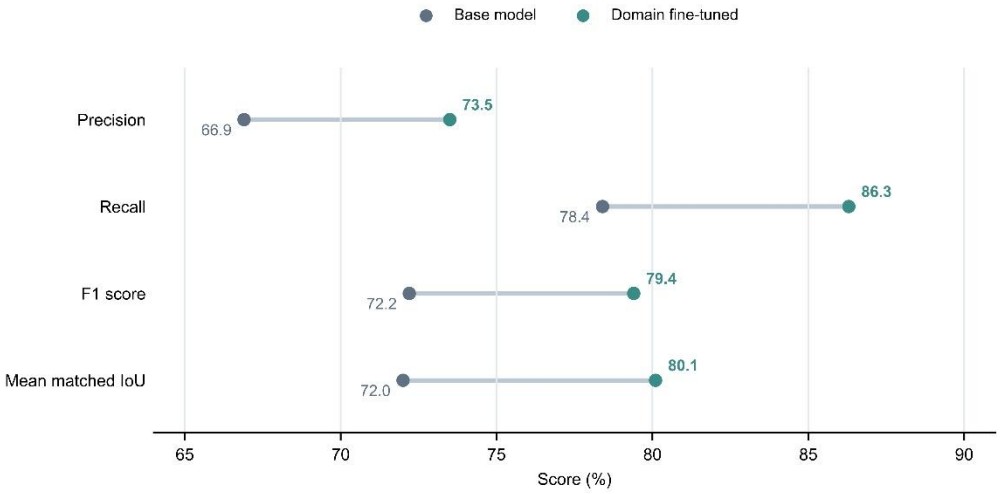


Fig. 7 | Segmentation performance before and after SAM 2 domain fine-tuning with the candidate set held constant. Grey and teal points denote the base and fine-tuned models, respectively. The paired values report precision, recall, F1 and mean matched intersection over union (IoU); connecting lines show the changes after fine-tuning.

After training and calibration on the 15-temple collection, the resulting model and inference settings were applied to mural images from the remaining 10 temples. Candidate fusion and mask segmentation produced localized figure instances for subsequent image–text correspondence and knowledge organization. The examples show extracted regions and their positions in the original images (Fig. 8 and Fig. 9).



Fig. 8 | a) Original deteriorated mural. b) Mask overlay after domain fine-tuning.

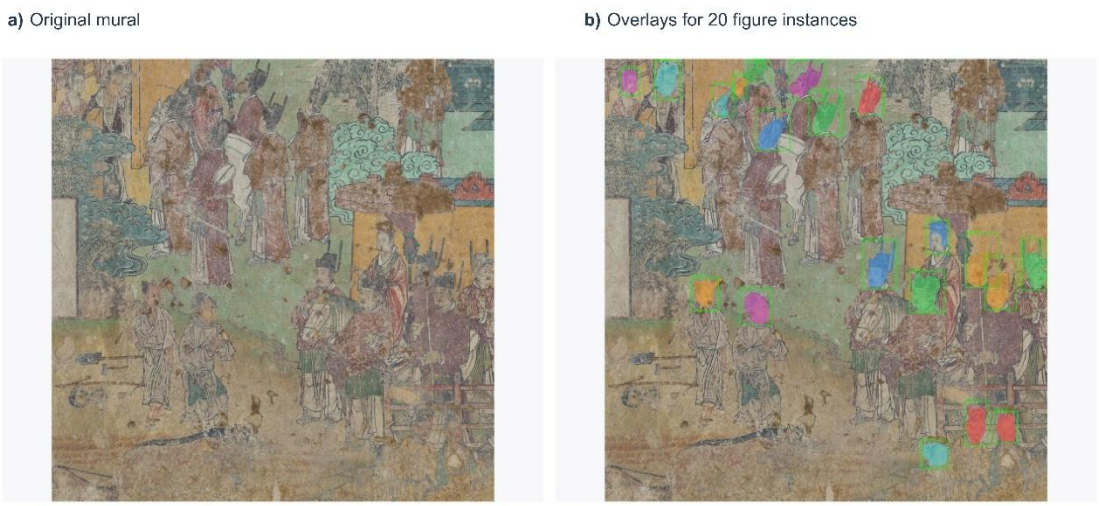


Fig. 9 | Multiple figures. Original mural (left) and predicted overlays for 20 figure instances (right). The overlays show the spatial distribution of extracted regions within the painting.

3.3 Temple-specific information organization and image–text correspondence

The materials for each temple were first divided into basic information and mural information, recording the name, date and location separately from mural themes and layout. They were then subdivided by hall and wall. A five-level path—"mural–mural content–wall and content category–figure or entry–text"—was used.

The materials were divided according to the objects being described. For example, descriptions of figure clothing, the overall composition and the story plot within the same passage were organized under figure, mural information and story entries, respectively. Figure entries retained visible forms, whereas story entries retained plots and related accounts; the entries were then linked through their shared wall location, while sources and expressions of uncertainty were retained.

For the 10 temples outside the model-development collection, figure masks were extracted in batches using the trained combined model, and image IDs and wall codes established their locations within the complete murals. Eight domain experts participated in establishing image–text correspondence. They first made judgments using individual figure regions, complete mural images, figure groupings and textual accounts, then jointly discussed ambiguous or

disputed entries to determine the correspondence and confidence level. Confirmed entries retained the grounds for judgment, such as identifying inscriptions, appearance, clothing and attributes, or position within the depicted episode, and were linked to image references. These model-derived records were incorporated into the same knowledge organization framework as the records from the 15 temples used for model development.

At Dongxianzi Guandi Pavilion, for example, the Guan Yu entry for the west gable wall was linked to an appearance description and figure images, whereas stories were stored as separate entries. After the images were uploaded to Alibaba Cloud, links were used to incorporate them into the temple-specific knowledge collection (Fig. 10).

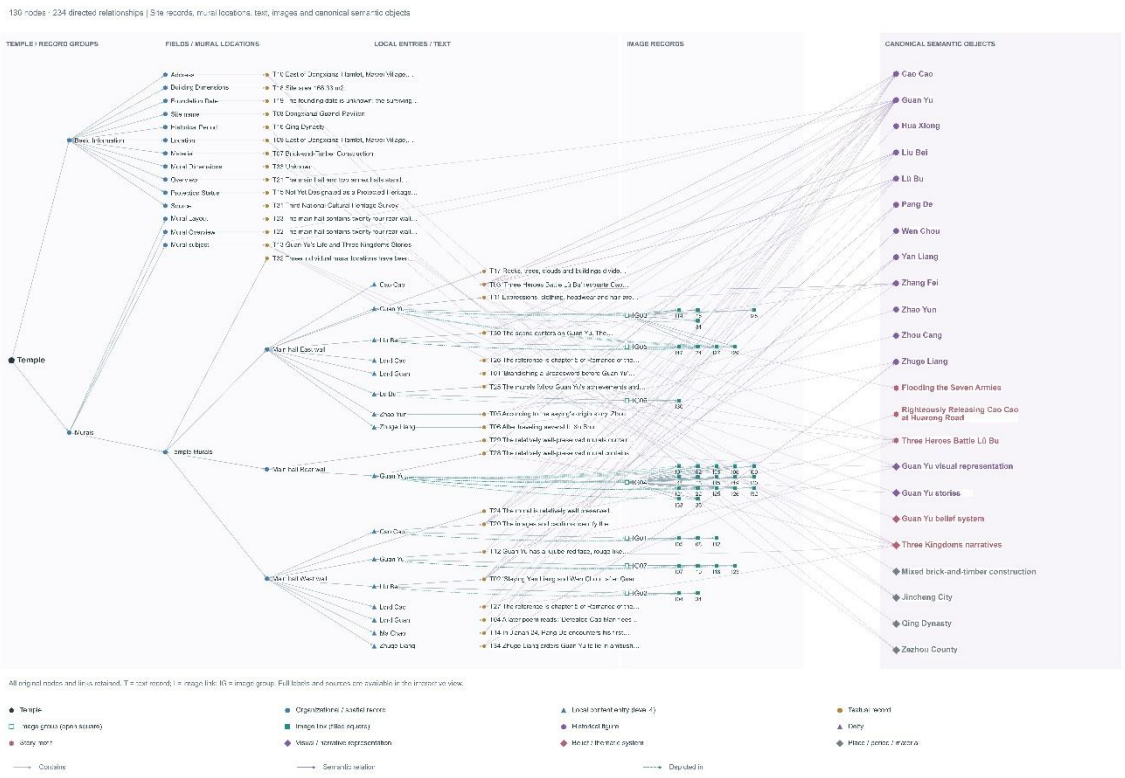


Fig. 10 | Within-temple knowledge organization at Dongxianzi Guandi Pavilion. Mural records, figure descriptions and image references are connected within their temple context. Selected labels identify representative records in the network.

3.4 Cross-temple semantic association

Following the organization of temple-specific information, entries from all temples are consolidated into a master table, retaining temple codes, row numbers, hierarchical paths and full text. Synonymous references to figures, stories and themes are assigned shared identifiers, establishing mappings from entries to canonical semantic objects.

Where supported by the materials, the original designations 'Lord Guan' and 'Guan Yunchang' are normalized to the historical figure 'Guan Yu', while entries concerning divine identity, visual representation, narrative and belief are linked to their respective objects and remain connected through the relations listed in Table 6. Each correspondence retains the original designation, the mural in which it occurs and the source.

Table 6 | Guan Yu-related semantic object categories and permitted relations

Object type	Canonical semantic object	Semantic relations between objects
Historical figure	Guan Yu	Guansheng Dijun—source of divine identity→Guan Yu; Guan Yu stories—about figure→Guan Yu
Deity	Guansheng Dijun	Guan Yu visual representation—depicts divine identity→Guansheng Dijun
Visual representation	Guan Yu visual representation	Linked to Guansheng Dijun through 'depicts divine identity'
Narrative representation	Guan Yu stories	Linked to Guan Yu through 'about figure'
Belief system	Guan Yu belief system	Linked to Guansheng Dijun through 'core deity'

Using the Guan Yu records of Dongxianzi Guandi Pavilion and Guojiagou Guandi Temple as an example, 'Lord Guan' and 'Guan Yunchang' in the materials are normalized to 'Guan Yu', figure descriptions are linked to 'Guan Yu visual representation', and the jointly recorded 'Three Heroes Battle Lü Bu' is linked to the same story object (Table 7 and Fig11). These two temples illustrate the associations formed under a common set of rules; records referring to these objects in other temples are linked using the same procedure.

Table 7 | Shared semantics and records from both Dongxianzi Guandi Pavilion and Guojiagou Guandi Temple

Associated object	Evidence from Dongxianzi Guandi Pavilion	Evidence from Guojiagou Guandi Temple	Rationale for the association and scope of evidence
Figure: Guan Yu	The murals depict events from Guan Yu's life.	The east and west gable walls depict stories from Guan Yu's life.	Both sets of records explicitly mention the same figure.
Visual representation: Guan Yu visual representation	A face the colour of dark red dates, phoenix-shaped eyes, silkworm-shaped eyebrows and a green robe exposing one shoulder.	Guan Yu, wearing a green robe, fights a general dressed in red.	Both sets of records contain descriptions of the depicted Guan Yu.
Story motif: Three Heroes Battle Lü Bu	The story title is recorded, with a citation to Chapter 5 of Romance of the Three Kingdoms.	The mural introduction lists 'Defeating the defenders of the pass: Three Heroes Battle Lü Bu'.	Both sets of records explicitly document the same story.
Narrative representation: Guan Yu stories	Centred on Guan Yu's life, including episodes of combat and his devotion to personal bonds and loyalty.	Stories such as Flooding the Seven Armies are organized around Guan Yu's life.	A shared narrative focus, although specific episodes may differ.
Thematic system: Three Kingdoms narratives	The themes concern episodes from Guan Yu's life and Three Kingdoms stories.	The story record states that all depicted content derives from the Three Kingdoms.	The theme and story records jointly support this classification.
Belief system: Guan Yu belief system	The themes concern episodes from Guan Yu's life and Three Kingdoms stories; the project additionally assigns a belief classification.	The themes concern stories from Guan Yu's life; the project additionally assigns a belief classification.	An inferred project classification; the cited text does not directly document religious practice.

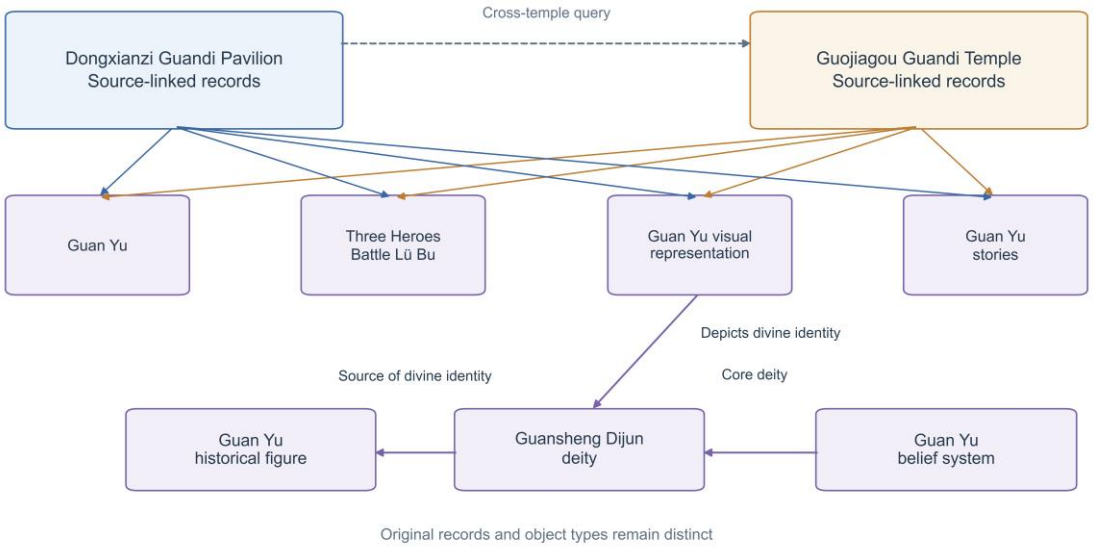


Fig. 11 | Cross-temple semantic associations in the Guan Yu case. Entries from the two temples are associated through shared objects and relations between objects; dashed lines indicate the resulting scope for temple comparison.

In the mapping table, this case contains 16 deduplicated shared objects: seven historical figures, two story motifs, one object each for belief system, narrative representation, thematic system and visual representation, and one object each for period, place name and material. An object referenced by multiple records is counted only once, so this count summarizes the shared scope of the records from the two temples. Table 7 presents evidence from both temples for representative objects and distinguishes direct textual correspondence from project-level semantic classification.

The organized results are saved as a mapping table linking entries from each temple to canonical semantic objects, together with a table of semantic relations between canonical semantic objects. The Guan Yu network across ten temples is shown in Fig. 12.

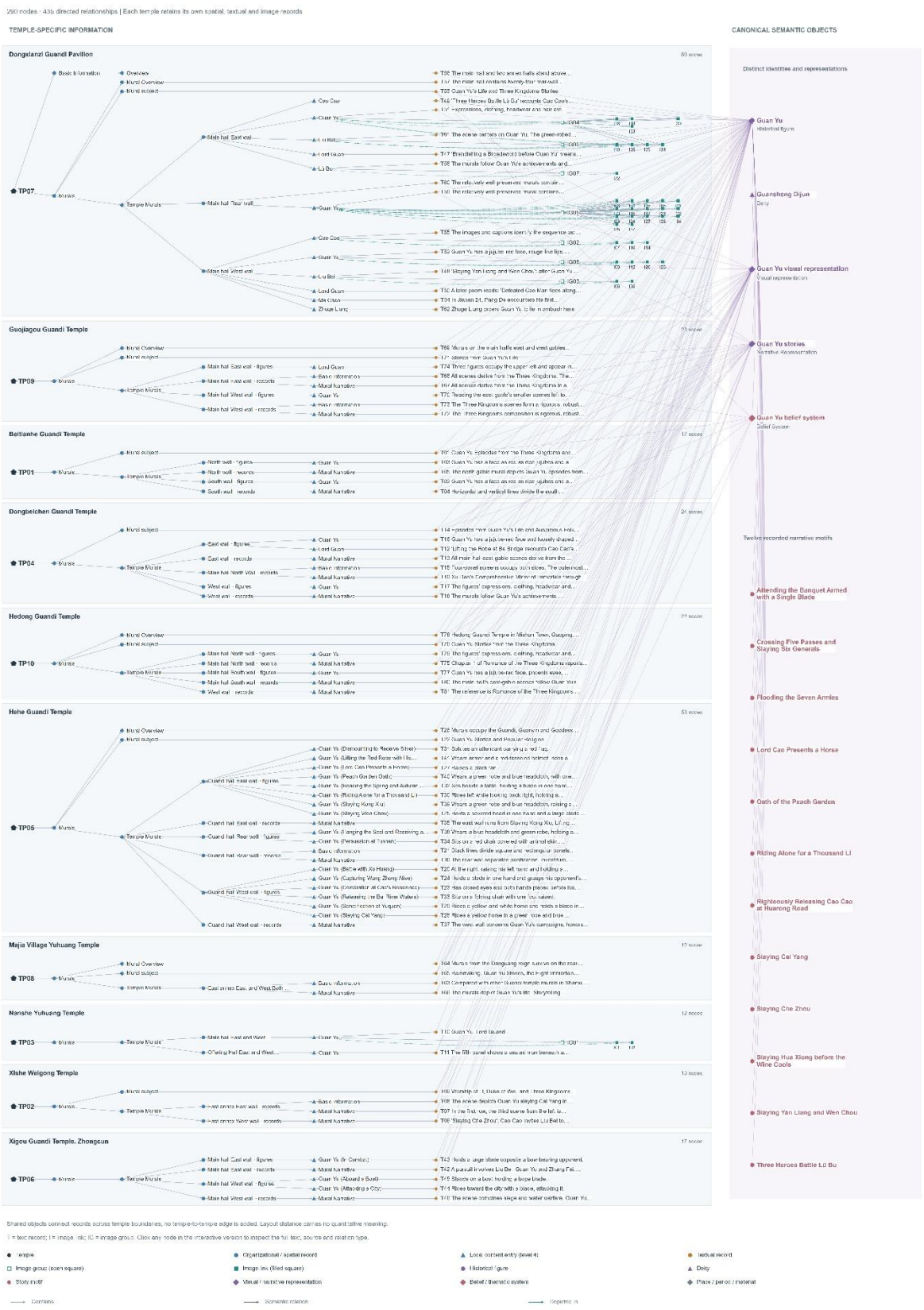


Fig. 12 | Cross-temple record network organized around shared semantic objects. Selected labels indicate representative figure, mural and image-related records. Node colours distinguish temple and spatial records, image records and canonical semantic objects. Connections preserve the temple-specific records associated with Guan Yu across the ten temples represented in this network.

3.5 Knowledge graph construction

During import into Neo4j, nodes are first created using unique identifiers, and source and target nodes are then connected according to the relation tables. Parent and child entries within each temple are linked by ‘contains’ relations, preserving the original organizational order of temples, murals and specific content. Figure image references are attached to the corresponding entries, while entries from different temples are connected through shared canonical semantic objects.

Nodes store names, content and source paths, while relations store correspondence types and the evidence supporting the associations. After import, duplicate identifiers, relation endpoints and image references are checked to verify that the tabular records have been incorporated into the graph.

The project’s 25-temple graph contains 3,433 nodes and 4,178 relations. Figure images from Dongxianzi Guandi Pavilion and Nanshe Yuhuang Temple are grouped and linked according to their corresponding entries, enabling figure records to present textual descriptions alongside image links.

Queries can use either a specific image or a shared semantic object. For example, an image of Guan Yu from Dongxianzi Guandi Pavilion can be used to retrieve the figure entry and mural location. This retrieval can then be extended through ‘Guan Yu’ to relevant records from other temples. Querying ‘Three Heroes Battle Lü Bu’ retrieves the corresponding stories from both temples and displays their respective textual descriptions (Fig. 13 and Fig. 14).

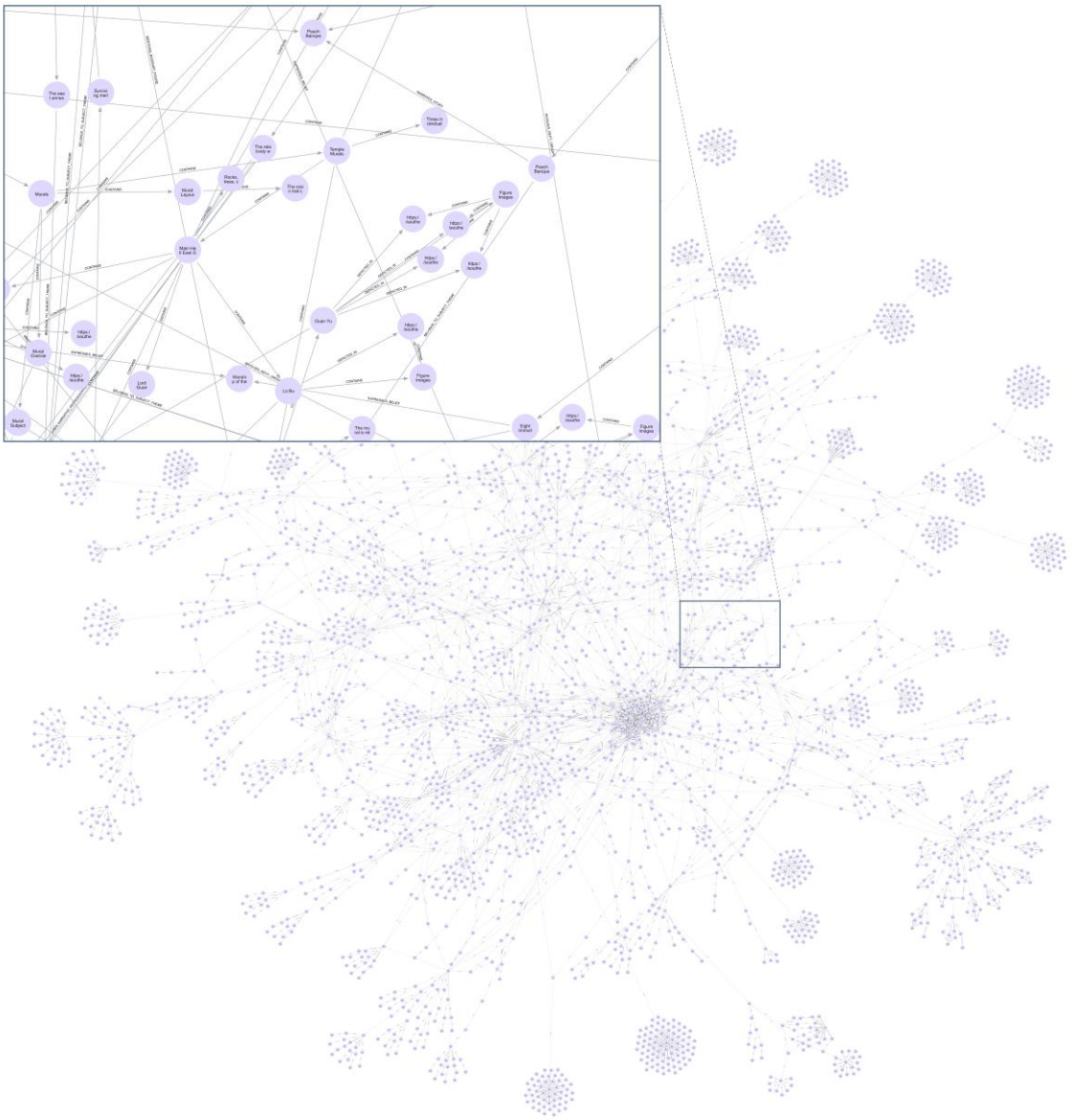


Fig. 13 | Detailed knowledge graph for the 25 temple collections. Nodes and directed connections show the integrated records and their relations. Labels retain the temple-specific records and semantic objects represented in the detailed network; Fig. 14 provides a category-coloured overview of the same collection.

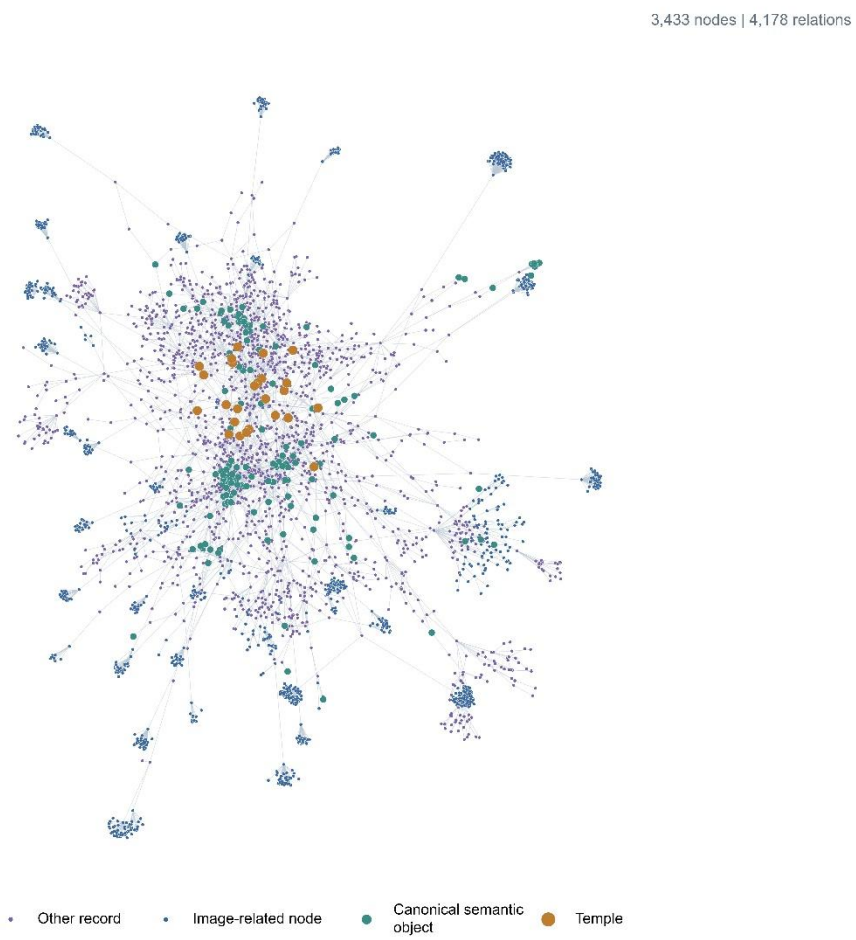


Fig. 14 | Overview of the 25-temple knowledge graph dataset, comprising 3,433 nodes and 4,178 relations. The network is constructed from the node and relation tables. Colours distinguish temple nodes, image-related nodes, canonical semantic objects and other records. Layout positions do not encode geographical distance or cultural similarity.

Discussion

Using materials from 25 temples in southeastern Shanxi, we constructed a cross-temple image–text knowledge graph containing 3,433 nodes and 4,178 relations, bringing figure images, spatial records and scholarly accounts into a unified framework for information organization. Figure instances serve as the unit of image–text correspondence, and shared figures and stories define the scope of thematic comparisons. Individual records retain the original mural contexts and sources of each temple. Model development used images from 15 temples, followed by inference on the remaining 10; the 25-temple graph therefore combines records from model-development and model-application sites.

Figure extraction, expert assessment and semantic normalization turn mural details into objects of comparison with cultural referents and supporting sources. Mask R-CNN generates figure candidates, Grounding DINO supplements localization, and SAM 2 refines boundaries. Original image indexes record temples, walls and tile positions, preserving the spatial relationship between each instance and the complete wall for checking adjacent figures, occlusion and local boundaries. In comparisons with fixed candidates, domain fine-tuning improved segmentation performance. Materials are organized by temple, hall and wall, with figure descriptions and story accounts forming separate entries; accounts of clothing, composition and plot within the same passage are assigned to their respective categories. Eight domain experts assess correspondence using figure regions, complete murals, figure groupings and textual materials. Ambiguous entries retain the grounds for judgment and high, medium or low confidence levels following discussion, and new materials can be added to the existing entries. Names, object types and descriptive content are checked separately before source-supported alternative names are assigned to shared objects. Original designations and sources are retained, while figures, divine identities, visual representations and stories are organized as distinct types. In the regional Guan Yu case across temples in southeastern Shanxi, source-supported references to ‘Lord Guan’ and ‘Guan Yunchang’ are assigned to the same figure, and ‘Three Heroes Battle Lü Bu’ refers to a shared story object, while each temple's figure descriptions and combinations of stories remain separately represented. Applying the same rules to Guan Yu records from ten temples creates a regional thematic collection around shared figures and stories. Figure or story queries return the corresponding images, wall locations, textual accounts and sources from each temple, grounding comparison in specific representations and their interpretative evidence.

Building on existing approaches to knowledge organization, specific figure images and their interpretative records form the units of observation for regional comparison. Previous knowledge organization has addressed heritage spatial semantics [25], image–text retrieval [24], relationships between artworks and related materials [32], and local pattern records [33], while regional studies have examined the spatial clustering of Buddhist sect sites [34] and the spatiotemporal distribution of cave temples and stone carvings [35]. Comparison here focuses on how the same figures and stories are represented in different temples, giving formal observations and interpretations of identity and narrative a common image referent. A geographical boundary brings temples from different periods, belief systems and cultural backgrounds into one collection, within which shared figures and stories define thematic scopes. This organization turns materials previously held in separate temple records into regional resources for systematic comparison, addressing the difficulty of conducting cumulative research on shared figures and their legends across temples.

In further research, these results can be used to establish regional thematic archives of figures and stories and compare figure characteristics, combinations of stories and scene arrangements across temples. Combined with reliable dates, geographical contexts and local historical sources, these archives can support investigations into which figures and stories recur across temples, how their combinations change with period and local context, and how these changes relate to religious practices. For application research, records with explicit image referents and sources can serve as inputs for image–text retrieval [36], question answering [37] and adaptive digital narratives [38], as well as thematic exhibitions. Knowledge graph-based storytelling frameworks provide a related approach to organizing multimedia resources for narrative applications [39]. Future tasks centred on figure

localization, story retrieval and source tracing can evaluate how these applications support information finding, comparative interpretation and the accumulation of regional knowledge.

The study covers 25 temples in southeastern Shanxi. Quantitative performance was assessed on the validation subset of the 15-temple model-development collection, and the trained model was subsequently applied to the remaining 10 temples. Further assessment across regions would establish the geographical scope of model performance. The visual models localize figures and delineate their boundaries; experts establish figure identities, narrative roles and religious meanings using identifying inscriptions, figure groupings and documentary evidence. Shared objects in the graph encode semantic associations between temple records. Investigating historical transmission and the direction of influence between temples requires these associations to be examined alongside dates and other historical sources.

Data availability

Project resources, including a trained Mask R-CNN model, a domain-tuned SAM 2 model, reproducible scripts, a fixed validation set and privacy-reviewed grayscale examples from 25 temples, are available at https://github.com/ZherongLiu/mural-person-segmentation-open

Acknowledgments

Funding for this project comes from the Shanxi Provincial Cultural Relics Protection Special Program, the ‘Shanxi Ancient Architecture Painted Sculpture and Mural Rescue Digital Protection Project.’, and Beijing Institute of Technology University-Level Joint Laboratory Cultivation.

Author contributions

Z.L. conceived and designed the study. Z.L. and Z.Z. collected and curated the mural images, spatial records and documentary materials. Z.L. and Y.Q. developed the computational workflow, constructed the knowledge graph and analysed the results. Z.L. validated the cultural interpretations. Z.L. prepared the figures and drafted the manuscript. Y.Q. supervised the research. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

References

1. Berezkin, R. Temple Painting and Vernacular Narratives: The Case of the Miaoshan Story Mural of Guanyin Monastery, Xinjin, Sichuan (Mid-Fifteenth Century). *Religions* **17**(1), 32 (2026). https://doi.org/10.3390/rel17010032

2. Feng, A. N. Assembling Gods: Nature, Community, and Temple Murals in Southern Shanxi (1300–1700). *Artibus Asiae* **83**(2), 127–170 (2023).

3. Yuan, Y. et al. Study on the dynamics of cultural routes and adaptability of folk beliefs: a case of the Great Tea Road and Guan Yu worship. *npj Heritage Science* **14,** 276 (2026). https://doi.org/10.1038/s40494-026-02567-8

4. Tian, Y. & Zhao, X. Restoring Faith and Form: Challenges and Strategies in the Preservation of Lord Guan Temples in Southern Shanxi. *Religions* **17,** 265 (2026). https://doi.org/10.3390/rel17020265

5. Dean, K. Local Communal Religion in Contemporary South-east China. *The China Quarterly* **174,** 338–358 (2003). https://doi.org/10.1017/S0009443903000214

6. Sun, S., Li, T. & Li, Q. Automated detection and topic mining of ancient murals across different styles. *npj Herit. Sci.* **14,** 112 (2026). https://doi.org/10.1038/s40494-026-02374-1

7. Sartini, B., Baroncini, S., van Erp, M., Tomasi, F. & Gangemi, A. ICON: An Ontology for Comprehensive Artistic Interpretations. *J. Comput. Cult. Herit.* **16**(3), Article 59, 1–38 (2023). https://doi.org/10.1145/3594724

8. Gesterkamp, L. The Heavenly Court: Daoist Temple Painting in China, 1200–1400 (Brill, 2011). https://doi.org/10.1163/ej.9789004184909.i-470

9. Li, Y. Analysis of the relationship between the temporal and spatial evolution of henan grotto temples and their geographical and cultural environment based on GIS. *Herit. Sci.* **11,** 216 (2023). https://doi.org/10.1186/s40494-023-01044-w

10. Kang, F., Li, W., Li, L., Zhou, Z. & Wang, J. Graph-enhanced clustering of late modern churches via multi-dimensional semantic feature integration. *npj Herit. Sci.* **14,** 100 (2026). https://doi.org/10.1038/s40494-026-02360-7

11. He, D. & Liu, Y. Holistic construction and classification of serial heritage corridors using Spatial Design Network Analysis and a minimum cumulative resistance model. *npj Herit. Sci.* **14,** 312 (2026). https://doi.org/10.1038/s40494-026-02542-3

12. Li, H. et al. Tracing the spatiotemporal diffusion of cross-strait Mazu culture: a knowledge graph-driven analysis. Sci. Rep. (2026). https://doi.org/10.1038/s41598-026-70645-9

13. Fang, X. & Han, M. Knowledge graph for intangible cultural heritage in Yellow River Basin integrating spatial characteristics. *npj Herit. Sci.* **14,** 338 (2026). https://doi.org/10.1038/s40494-026-02625-1

14. Abgaz, Y., Rocha Souza, R., Methuku, J., Koch, G. & Dorn, A. A Methodology for Semantic Enrichment of Cultural Heritage Images Using Artificial Intelligence Technologies. *J. Imaging* **7,** 121 (2021). https://doi.org/10.3390/jimaging7080121

15. He, K., Gkioxari, G., Dollar, P. & Girshick, R. Mask R-CNN. In Proc. IEEE International Conference on Computer Vision (ICCV) 2980–2988 (2017). https://doi.org/10.1109/ICCV.2017.322

16. Liu, S. et al. Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection. In Computer Vision – ECCV 2024. Lecture Notes in Computer Science 15105, 38–55 (Springer, 2025). https://doi.org/10.1007/978-3-031-72970-6_3

17. Kirillov, A. et al. Segment Anything. In Proc. IEEE/CVF International Conference on Computer Vision (ICCV) 3992–4003 (2023). https://doi.org/10.1109/ICCV51070.2023.00371

18. Cao, J. et al. Ancient mural segmentation based on multiscale feature fusion and dual attention enhancement. *Herit. Sci.* **12,** 58 (2024). https://doi.org/10.1186/s40494-024-01172-x

19. Zhang, Q. et al. A Mamba based vision transformer for fine grained image segmentation of mural figures. *npj Herit. Sci.* **13,** 204 (2025). https://doi.org/10.1038/s40494-025-01682-2

20. Fu, D., Wang, C., Li, Z. & Hou, D. Parameter-efficient cross-modal prompt tuning for few-shot ancient mural classification. npj Herit. Sci. (2026). https://doi.org/10.1038/s40494-026-02851-7

21. Zou, T., Li, Y., Dong, Y. & An, J. Lightweight high-precision monitoring of damage locations in Dunhuang murals based on improved YOLACT. *Sci. Rep.* **15,** 40731 (2025). https://doi.org/10.1038/s41598-025-24560-0

22. Baek, S. et al. Toward enhanced unsupervised clustering of 20th century Korean paintings via multimodal features. *npj Herit. Sci.* **14,** 76 (2026). https://doi.org/10.1038/s40494-026-02304-1

23. Fan, T., Wang, H. & Hodel, T. CICHMKG: a large-scale and comprehensive Chinese intangible cultural heritage multimodal knowledge graph. *Herit. Sci.* **11,** 115 (2023). https://doi.org/10.1186/s40494-023-00927-2

24. Wan, J. et al. WuMKG: a Chinese painting and calligraphy multimodal knowledge graph. *Herit. Sci.* **12,** 159 (2024). https://doi.org/10.1186/s40494-024-01268-4

25. Yang, S. & Hou, M. Knowledge graph representation method for semantic 3D modeling of Chinese grottoes. *Herit. Sci.* **11,** 266 (2023). https://doi.org/10.1186/s40494-023-01084-2

26. Yang, S., Hou, M. & Fan, H. CityGML Grotto ADE for modelling niches in 3D with semantic information. *Herit. Sci.* **12,** 135 (2024). https://doi.org/10.1186/s40494-024-01260-y

27. Carboni, N. & de Luca, L. An Ontological Approach to the Description of Visual and Iconographical Representations. *Heritage* **2**(2), 1191–1210 (2019). https://doi.org/10.3390/heritage2020078

28. Schimmenti, A., Pasqual, V., Vitali, F. & van Erp, M. Knowledge graphs generation from cultural heritage texts: combining LLMs and ontological engineering for scholarly debates. *J. Doc.* **82**(7), 206–250 (2026). https://doi.org/10.1108/JD-07-2025-0203

29. Bobasheva, A., Gandon, F. & Precioso, F. Learning and Reasoning for Cultural Metadata Quality: Coupling Symbolic AI and Machine Learning over a Semantic Web Knowledge Graph to Support Museum Curators in Improving the Quality of Cultural Metadata and Information Retrieval. *J. Comput. Cult. Herit.* **15**(3), 1–23 (2022). https://doi.org/10.1145/3485844

30. Homburg, T., Cramer, A., Raddatz, L. & Mara, H. Metadata schema and ontology for capturing and processing of 3D cultural heritage objects. *Herit. Sci.* **9,** 91 (2021). https://doi.org/10.1186/s40494-021-00561-w

31. Ravi, N. et al. SAM 2: Segment Anything in Images and Videos. International Conference on Learning Representations (2025). https://openreview.net/forum?id=Ha6RTeWMd0

32. Zhang, Z. Conceptual model construction for the A Comprehensive Collection of Ancient Chinese Paintings based on metadata and ontology. *npj Herit. Sci.* **14,** 63 (2026). https://doi.org/10.1038/s40494-025-02258-w

33. Wu, X., Yuan, Q., Qu, P. & Su, M. Image-driven batik product knowledge graph construction. *npj Herit. Sci.* **13,** 20 (2025). https://doi.org/10.1038/s40494-025-01586-1

34. Yue, X. et al. Grid-scale analysis of the spatial pattern and hotspot detection of Buddhist sects in China. *npj Herit. Sci.* **13,** 449 (2025). https://doi.org/10.1038/s40494-025-02026-w

35. Ma, Q. et al. The spatio-temporal distribution and influencing factors of cave temples and stone carvings in China. *npj Herit. Sci.* **13,** 311 (2025). https://doi.org/10.1038/s40494-025-01879-5

36. Sun, S. et al. A dual-fusion strategy driven multimodal mobile visual search model for Dunhuang murals. *npj Herit. Sci.* **13,** 225 (2025). https://doi.org/10.1038/s40494-025-01788-7

37. Wan, J., Li, X., Zhang, H., Zou, A. & Wang, R. An LLM-based QA system for Chinese Painting and Calligraphy with Knowledge Graphs and external documents. *npj Herit. Sci.* **13,** 643 (2025). https://doi.org/10.1038/s40494-025-02219-3

38. Nappi, M. L., Buono, M., Chivăran, C. & Giusto, R. M. Models and tools for the digital organisation of knowledge: accessible and adaptive narratives for cultural heritage. *Herit. Sci.* **12,** 112 (2024). https://doi.org/10.1186/s40494-024-01219-z

39. Renzi, G., Rinaldi, A. M., Russo, C. & Tommasino, C. A storytelling framework based on multimedia knowledge graph using linked open data and deep neural networks. Multimed. Tools Appl. 82, 31625–31639 (2023).