From Prediction to Multimodal Evaluation: Human-in-the-Loop Workflows for Museum Conservation

Anonymous ACL submission

Abstract

Recent advances in multimodal large language models (MLLMs), retrieval-augmented generation (RAG), and vision-language systems have accelerated multimodal AI research over images, documents, metadata, and scientific measurements in cultural heritage settings. However, existing work remains fragmented across computer vision, NLP, conservation science, and digital humanities, with limited discussion of evidence grounding, provenance tracking, and human-in-the-loop verification in expert-facing workflows. This survey organizes the literature around multimodal evidence grounding for conservation, spanning direct conservation applications, enabling technologies such as multimodal retrieval and knowledge graphs, and broader cultural AI tasks including artwork captioning and culturally grounded evaluation. Recurring limitations include weak grounding between generated outputs and observable evidence, fragmented evaluation protocols, insufficient provenance modeling, and limited expert-centered benchmarks. These gaps motivate future work on provenance-aware reasoning, evidence-grounded generation, and human-in-the-loop multimodal evaluation.

1 Introduction

Recent advances in multimodal large language models (MLLMs), retrieval-augmented generation (RAG), vision-language systems, and scientific imaging have accelerated research on multimodal reasoning over images, documents, metadata, and scientific measurements. A growing body of work applies these systems to cultural heritage, museums, archives, collection management, and related institutional settings. Earlier surveys and policy reports have mapped machine learning applications for cultural heritage, digitized collections, cultural access, museum planning, and heritage protection (Fiorucci et al., 2020; Murphy and Villaespesa, 2020; Pasikowska-Schnass and Lim, 2023; Colace et al., 2025; Thiel and Bernhardt, 2024). In parallel, museum-oriented work has examined the benefits and risks of computer vision for collection tagging, discovery, and access (Villaespesa and Murphy, 2021; Villaespesa and Crider, 2021), while broader cultural management research has discussed AI in cultural policy, arts-led development, digital artwork, creative labor, wellbeing, and sectoral transformation (Murphy and Franklin, 2025; Andrews and Hawcroft, 2024; Duester, 2024; Frost and Vargas, 2024; Rizzi and Bertola, 2025; Connock, 2024).

A workflow-oriented perspective organizes this fragmented research space around evidence grounding, verification, provenance tracking, and expert validation rather than standalone prediction tasks. The central challenge is not only integrating multimodal observations, scientific measurements, metadata, provenance records, and expert annotations into traceable workflows, but also evaluating whether generated outputs remain faithful to observable evidence and domain interpretation. In this context, evidence may include visual observations, scientific imaging signals, archival provenance records, conservation reports, metadata fields, and human annotations that must remain attributable and verifiable throughout the workflow. Current multimodal systems are effective at visual similarity matching and surface-level feature alignment, yet still struggle to capture deeper contextual and cultural meaning beyond statistical associations. Reliable evaluation therefore often requires human-in-the-loop validation, particularly in settings involving uncertainty and domain-specific interpretation. This perspective differs from broader cultural heritage AI surveys that mainly focus on digitization, retrieval, or artwork understanding (Fiorucci et al., 2020; Colace et al., 2025), as well as cultural-bias and cultural-alignment research that primarily studies fairness and representation outside evidence-intensive human workflows (Gal-

legos et al., 2024; Pawar et al., 2024; Adilazuarda et al., 2024).

Existing research can be organized into three layers: (1) direct human-facing applications such as condition assessment, restoration support, and technical documentation; (2) enabling technologies including multimodal retrieval, knowledge graphs, scientific imaging, provenance-aware systems, and RAG pipelines; and (3) broader multimodal cultural AI tasks such as artwork captioning, VQA, and culturally grounded multimodal evaluation. Across these layers, recurring technical challenges include multimodal evidence grounding, provenance-aware retrieval, observation–inference separation, hallucination control, uncertainty calibration, and human-centered evaluation in multimodal workflows. The main contributions are summarized as follows:

- We introduce a workflow-oriented technical taxonomy for multimodal systems spanning scientific imaging, multimodal retrieval, knowledge integration, evidence grounding, provenance tracking, and human-in-the-loop verification.
- We synthesize datasets, models, and infrastructures across scientific imaging, vision-language modeling, RAG, knowledge graphs, MLLMs, and agent systems, and map their roles in conservation-oriented multimodal workflows.
- We identify recurring challenges in multimodal alignment and evaluation, particularly in evidence grounding, observation–inference separation, provenance-aware reasoning, and human-centered assessment in multimodal workflows.

2 Survey Methodology and Scope

2.1 Review Type

This paper adopts a curated narrative review methodology with structured thematic synthesis rather than a fully systematic meta-analysis. This choice reflects the interdisciplinary and unevenly distributed state of the field: conservation-specific AI studies remain relatively sparse, while relevant methods and infrastructures are scattered across computer vision, multimodal learning, museum informatics, heritage science, archives, cultural-bias evaluation, and digital humanities.

Relevant papers were identified through iterative literature search, citation snowballing, and manual expert curation across scientific publications, technical reports, and recent preprints. The reviewed literature spans early semantic-web and heritage-infrastructure work through recent 2026 studies, with greater emphasis on post-2018 work because of the rapid growth of deep learning, MLLMs, and RAG-based systems.

The final cited set was selected to support the survey's core analytical categories rather than to provide exhaustive bibliometric coverage. Because the review process relied on iterative snowballing and expert-guided literature expansion rather than a preregistered systematic-review protocol, the survey should be interpreted as a structured technical synthesis rather than a reproducible PRISMA-style meta-analysis. The review therefore prioritizes work that informs how multimodal systems represent, retrieve, align, verify, or evaluate evidence in human-facing workflows.

2.2 Direct, Adjacent, and General Heritage Relevance

Because conservation-specific AI work remains sparse, this survey distinguishes between direct conservation studies and conservation-relevant enabling research. We categorize prior work according to its degree of conservation relevance:

- **Direct conservation relevance**: work that directly addresses conservation, restoration, condition assessment, deterioration monitoring, conservation documentation, scientific imaging for material analysis, or intervention support.
- **Adjacent conservation relevance**: work that does not directly solve a conservation task but provides enabling infrastructure, datasets, multimodal reasoning methods, knowledge graphs, RAG systems, or documentation technologies relevant to conservation workflows.
- **General Heritage heritage relevance**: broader museum, cultural heritage, artwork understanding, or cultural-bias research that informs conservation indirectly through access, interpretation, governance, or representational validity.

This distinction is used throughout the survey to avoid scope inflation. Many enabling technologies discussed in this survey originate outside conservation-specific AI; their relevance emerges from the convergence of multimodal reasoning, scientific imaging, provenance infrastructures, institutional documentation, and human-centered workflows.

3 Museum Conservation as an Evidence-Grounded Multimodal Workflow

3.1 Why Conservation Differs from General-purpose museum AI systems

General-purpose museum AI systems often focuses on access, discovery, educational interpretation, or visitor interaction. These are important, but they do not exhaust conservation needs. Conservation workflows require reasoning over material change, uncertainty, treatment histories, physical evidence, and documentation and provenance constraints. Computer vision for museum collections may correctly detect objects or generate tags (Villaespesa and Murphy, 2021; Villaespesa and Crider, 2021); artwork captioning datasets may improve semantic description (Stefanini et al., 2019; Garcia and Vogiatzis, 2018; Lu et al., 2024; Becattini et al., 2023); and MLLMs may support natural-language questions about artworks (Bin et al., 2024; Ferrato et al., 2025). Yet conservation decisions require a different evidential standard. A visually plausible description can still be evidentially unsupported if it fabricates material properties, ignores prior treatment, or confuses visible observation with inferred cultural meaning.

This difference is visible in direct conservation work, such as cultural relic protection datasets (Yu et al., 2022), edge and crack-related detection for conservation (Targa et al., 2025), transformer-based restoration of cultural relic images (Wu et al., 2026), and digital restoration workflows that translate computational reconstruction into physical intervention (Kachkine, 2025). Archaeological site conservation reviews similarly highlight monitoring, deterioration, and preservation planning rather than only semantic interpretation (Casillo et al., 2025). These studies point to a more demanding question: can AI support conservation reasoning that is materially grounded, traceable, and verifiable?

3.2 Conservation Documentation as Evidence Infrastructure

Conservation documentation is not a passive archive. From an AI systems perspective, it functions as a multimodal memory, retrieval, and evidence-tracking layer. Institutions use documentation to record object condition, intervention history, treatment rationale, uncertainty, provenance, material analysis, and expert disagreement. This includes condition reports, treatment records, conservation vocabularies, environmental monitoring logs, photographic records, scientific imaging metadata, and decision logs. Standards and communities such as ICOM-CC provide conservation-oriented documentation standards (International Council of Museums Committee for Conservation, 2024), while CIDOC CRM supports semantic interoperability for cultural heritage documentation (Doerr, 2003). FAIR principles provide a broader data stewardship framework (Wilkinson, 2016), and recent heritage-science work emphasizes the need to operationalize FAIR and linked data practices in heritage settings (Larsson et al., 2025).

The key tension is that most AI systems consume documentation as input, while conservation workflows also require AI outputs to become future documentation. If a model generates a condition summary, provenance note, or treatment suggestion, that output may enter institutional records. Therefore, conservation AI must support audit trails, uncertainty labels, provenance attribution, post-hoc verification and correction mechanisms, and retrieval traceability. Knowledge graphs such as CHAD-KG and museum collection KGs provide relevant representational infrastructure (Barzaghi et al., 2025; Li et al., 2025); CIDOC-based construction with LLMs suggests how structured heritage knowledge may be partially automated (Wang and Zhang, 2025); collaborative ontology construction further highlights how metadata structure and ontology design influence representation and retrieval (Chen et al., 2026). Related work on provenance-aware agent workflows and retrieval-grounded scientific QA further illustrates the growing importance of traceable retrieval and evidence tracking in AI systems (Souza et al., 2025; Lála et al., 2023). However, the unresolved question is not merely how to build knowledge graphs or retrieval systems, but how to ensure that AI-generated claims entering these workflows remain evidentially grounded, traceable, and verifiable. A workflow-level example illustrating multimodal evidence integration and expert validation is provided in Appendix A.

4 An Iterative Taxonomy of Conservation AI

Figure 1 presents an iterative taxonomy for museum conservation AI. Unlike a flat layered taxonomy, this framework emphasizes feedback loops among retrieval, reasoning, validation, documenta-

tion, and governance. Expert review updates institutional records; conservation decisions generate new treatment documentation; governance audits revise access policies; and uncertainty logs inform future retrieval and model use. The framework can therefore be interpreted as an iterative retrieval, reasoning, validation, and documentation loop for multimodal AI systems.

The taxonomy clarifies why isolated performance metrics are insufficient. A model may perform well on a visual task but fail to produce auditable evidence traces. Conversely, a knowledge graph may improve retrieval while reproducing historically incomplete metadata. Conservation AI therefore requires both technical capability and documentary accountability.

5 Data, Datasets, and Benchmarking

5.1 From Artwork Understanding to Conservation Evidence

Many datasets used in cultural heritage AI were built for artwork understanding rather than conservation. SemArt, Artpedia, ArtCap, VISCOUNTH, and related datasets support retrieval, captioning, VQA, or visual-semantic alignment (Garcia and Vogiatzis, 2018; Stefanini et al., 2019; Lu et al., 2024; Becattini et al., 2023). These resources are valuable for model development, but most do not include conservation records, scientific imaging, treatment history, uncertainty annotations, or material analyses. Captioning work on iconography, ceramics, ancient artworks, and Thangka imagery further improves cultural description (Cetinic, 2021; Liu et al., 2023; Zheng et al., 2023; Zhong et al., 2026), and diffusion-based augmentation and dense annotation pipelines can improve data collection (Cioni et al., 2023; Lin et al., 2025). Yet from a conservation perspective, these datasets remain mostly **Adjacent**: they support semantic access but do not directly evaluate conservation reasoning.

Direct conservation datasets and projects are rarer. The Dunhuang cultural heritage protection project combines AI with conservation needs in a major heritage context (Yu et al., 2022); edge detection for cultural heritage conservation explicitly targets visual features relevant to conservation workflows (Targa et al., 2025); and digital/-physical restoration work highlights intervention risks (Kachkine, 2025). Architectural heritage datasets and archaeological site conservation reviews expand the scope toward built and spatial heritage (Abu Talib et al., 2025; Casillo et al., 2025), but museum-object conservation still lacks widely adopted multimodal benchmarks combining RGB, scientific imaging, 3D scans, treatment records, provenance files, and expert condition annotations. A structured comparison of representative datasets and benchmarks is provided in Appendix C.

5.2 Benchmark Gap

The benchmark gap is not only a lack of data volume. It is a lack of conservation-specific supervision. Current benchmarks usually reward classification, retrieval, captioning, or question answering, while conservation requires observation–inference separation, provenance grounding, material-evidence alignment, uncertainty calibration, and expert review. Emerging benchmarks such as Appear2Meaning, VULCA-BENCH, RAVENEA, AlignMMBench, and VaseVQA move toward structured cultural metadata inference, visual culture understanding, multimodal retrieval augmentation, Chinese multimodal alignment, and domain-specific ancient pottery reasoning (Jiang et al., 2026; Yu et al., 2026; Li et al., 2026; Wu et al., 2025; Ge et al., 2026). These works are important because they expose failures in cultural and multimodal reasoning. However, they remain mostly **Adjacent** rather than direct conservation benchmarks unless they are connected to material condition, treatment documentation, or intervention evaluation. Current benchmarks also rarely evaluate whether multimodal systems hallucinate unsupported material properties, provenance claims, or restoration rationales.

6 Multimodal Perception, Imaging, and Spatial Reasoning

6.1 Computer Vision: Pattern Recognition Without Evidence Interpretation

Computer vision has long been used for classification, pattern extraction, visual retrieval, and representation learning in artworks (Castellano and Vessio, 2021; Zhao et al., 2021; Scaringi et al., 2025). More recent systems use MLLMs or graph-augmented models for artwork captioning, classification, and retrieval (Bin et al., 2024; Jiang et al., 2024; Fanelli et al., 2025; Bai et al., 2025). These systems improve semantic recognition and interactive analysis, but conservation-oriented workflows require a different question: what does the visual pattern imply for condition, material change, or

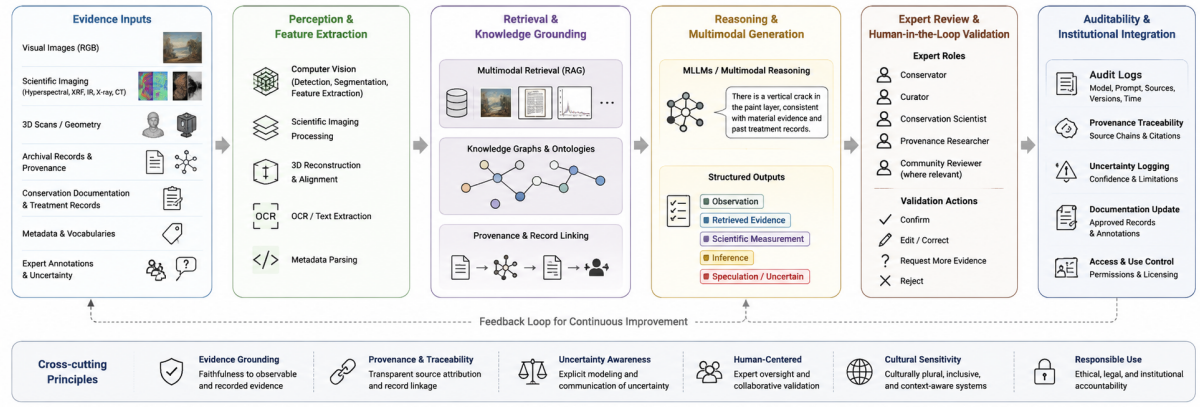

Figure 1: Human-in-the-loop verification workflow for evidence-grounded multimodal conservation AI. Expert review iteratively updates multimodal retrieval, reasoning, documentation, and auditability.

treatment?

A crack detector may identify line-like structures, but it cannot by itself determine whether the crack is active, historically stable, structurally dangerous, visually intrusive, or relevant to treatment. A style classifier may recognize iconographic features, but it does not assess pigment instability or prior overpainting. This gap explains why generic visual performance cannot be equated with conservation reliability. Conservation-oriented computer vision must connect visual features to material evidence, documentation history, and expert judgement.

6.2 Scientific Imaging: Material Evidence and Its Limits

Scientific imaging is central to conservation because it can reveal evidence not visible in standard RGB images. Hyperspectral and multispectral imaging can support pigment identification, spectral mapping, and condition analysis (Cucci et al., 2016). XRF and related depth-profiling or subsurface imaging techniques provide elemental and material information relevant to historical paintings (Alfeld and Broekaert, 2013). Infrared reflectography and visible/infrared spectroscopy can reveal underdrawings, compositional changes, or hidden features (Delaney, 2010). X-ray, CT, and other modalities can provide structural and subsurface evidence.

The key tension is that scientific imaging provides material evidence, but it does not automatically produce grounded interpretation. Pigment identification must be interpreted in relation to object history, dating, prior restoration, environmental exposure, and treatment options. Degradation detection must distinguish meaningful change from acquisition artifacts, lighting variation, or sensor noise. Cross-instrument generalization is a major unresolved problem: models trained on one museum's imaging setup may fail under another institution's sensor calibration, spectral resolution, illumination protocol, or documentation practice.

Multimodal registration and fusion remain especially difficult. Conservation workflows may require aligning RGB photographs, hyperspectral cubes, XRF maps, infrared reflectograms, 3D surfaces, and textual records. Each modality carries different spatial resolution, uncertainty, and acquisition constraints. Current AI systems often treat modalities separately, while conservation reasoning requires synchronized evidence across material, spatial, temporal, and documentary sources. This creates a multimodal alignment problem in which visual, spectral, spatial, and documentary evidence must remain mutually consistent across retrieval and reasoning workflows. Representative imaging modalities, evidence types, and associated AI challenges are summarized in Appendix D.

6.3 3D and Spatial Heritage

Spatial documentation contributes to conservation by preserving geometry, surface structure, deformation, and context. Photogrammetry and 3D scanning are established methods for heritage recording (Remondino and Campana, 2011). Neural rendering techniques such as NeRF (Mildenhall, 2020) and emerging generative or dense 3D annotation methods (Lin et al., 2025) offer new opportunities for documentation and visualization. However, conservation-grade spatial AI requires more than perceptually convincing reconstruction. It must

support geometric fidelity, repeatable acquisition, scale accuracy, long-term archival formats, uncertainty annotation, and compatibility with treatment records.

This creates a tension between computer graphics benchmarks and conservation needs. A reconstruction may appear visually convincing but still fail to preserve measurement accuracy or surface details needed for deterioration monitoring. Gaussian splatting and neural rendering methods may support visualization and interaction, but their conservation value depends on whether they can be audited, remeasured, and aligned with institutional documentation.

7 Knowledge Graphs, RAG, MLLMs, and Agents

7.1 Knowledge Infrastructures

Knowledge graphs, linked data, CIDOC CRM, and FAIR practices provide the representational backbone for evidence-grounded conservation AI. Archival and museum AI scholarship emphasizes access and born-digital records (Jaillant, 2022), while museum AI collections highlight practical institutional applications (Thiel and Bernhardt, 2024). CHAD-KG models digitization paradata (Barzaghi et al., 2025), CIDOC-based LLM construction explores automated KG generation (Wang and Zhang, 2025), and MuseKG proposes collection-scale museum graph infrastructure (Li et al., 2025). Work on scholarly debate KGs further shows how LLMs and ontological engineering can structure cultural heritage texts (Schimmenti et al., 2025).

The analytical point is that knowledge graphs improve structure but do not guarantee truth. Museum records can be incomplete, outdated, colonial, contested, or unevenly digitized. A KG can formalize uncertainty poorly if the ontology assumes stable entities and relations. Collaborative ontology work explicitly frames interoperability as an equity issue (Chen et al., 2026), reminding us that evidence infrastructures are also cultural and political infrastructures.

7.2 RAG and Provenance Search

RAG systems are especially relevant because they can connect generative models to institutional records. RAG for the Getty Provenance Index demonstrates natural-language provenance search (Henrickson et al., 2026); museum QA work compares retrieval and context strategies (Ramos-Varela et al., 2025); visual RAG for intangible cultural heritage supports recognition in culturally specific contexts (Fan et al., 2025); and RAVENEA extends multimodal retrieval-augmented visual culture understanding (Li et al., 2026). These systems move beyond closed-book generation and make evidence retrieval part of the reasoning process.

Yet RAG systems inherit the limits of their corpora. If provenance records are incomplete, contested, or biased, retrieval may amplify gaps. If retrieved passages are not cited or attributed, generative fluency may conceal evidential weakness. Conservation RAG therefore requires provenance-faithfulness metrics, source-level attribution, contradiction detection, and confidence labels that distinguish archival evidence from speculative inference.

7.3 MLLMs and Agents

MLLMs enable interactive analysis of paintings, artifacts, images, and collection records (Bin et al., 2024; Ferrato et al., 2025; Tiet et al., 2025). Captioning and multimodal generation work has expanded from general artwork captions to localized descriptions, iconographic captions, crowdsourced cultural heritage captioning, and multi-agent captioning (Lian et al., 2024; Cetinic, 2021; Reshetnikov and Marinescu, 2025; Bai et al., 2025). Agentic and debate-based systems may improve reasoning diversity and cultural alignment (Ki et al., 2025), and hybrid cognitive alignment research suggests a broader human–AI collaboration perspective (Lu and Yan, 2026).

However, MLLMs are dangerous in conservation if they blur the boundary between what is visible, what is retrieved from records, what is inferred from context, and what is speculative. This problem is especially salient in structured metadata inference (Jiang et al., 2026), artwork QA (Ferrato et al., 2025), and multimodal alignment surveys (Shu et al., 2025). For conservation, the central requirement is not simply better answers, but answer decomposition into evidence, inference, uncertainty, and recommended expert review. Representative modeling paradigms and their conservation-related tensions are summarized in Appendix E.

8 Evaluation and Expert Validation

8.1 Why Generic Metrics Fail

Generic metrics such as accuracy, BLEU, CIDEr, IoU, F1, or Recall@K remain useful for narrow technical evaluation, but they do not capture conservation reliability. A generated caption can be fluent but unsupported by evidence. A retrieval system can return visually similar artworks while missing provenance-sensitive distinctions. A segmentation model can detect cracks but provide no information about active deterioration or treatment risk. An MLLM can produce plausible historical context while fabricating provenance. Evaluation in conservation AI must therefore assess both technical performance and evidence grounding. Representative evaluation dimensions and their operational limitations are summarized in Appendix G.

8.2 Expert Validation Protocols

Human-in-the-loop evaluation is too vague a standard unless the evaluation protocol is explicitly defined. Conservation AI evaluation should specify who evaluates, what is evaluated, and how disagreement is recorded. Expert validation may involve conservators assessing material plausibility, curators checking interpretation, conservation scientists evaluating imaging claims, provenance researchers verifying archival evidence, and domain or community reviewers evaluating culturally sensitive interpretation. Inter-expert agreement should be measured where appropriate, but disagreement should not always be treated as noise. In contested heritage contexts, disagreement may reflect legitimate interpretive plurality.

A robust protocol should include blind review where possible, randomized model outputs, model-name masking, independent annotation before discussion, adjudication rules, examples of pass/fail criteria, uncertainty labels, and failure-case review. Expert review should also calibrate AI outputs against conservation records, including condition reports, treatment histories, imaging evidence, and provenance files.

8.3 Cultural and Multimodal Misalignment

Cultural and multimodal misalignment in conservation AI is not merely a representational fairness problem. It can affect provenance interpretation, treatment prioritization, multilingual metadata generation, sensitive provenance documentation, source-community knowledge representation, and contested historical interpretation. Prior work shows that LLMs and MLLMs often encode dominant cultural assumptions, display uneven multilingual cultural understanding, and rely heavily on surface-level visual or linguistic alignment in cross-cultural reasoning tasks (Wang et al., 2024; Myung et al., 2024; Chiu et al., 2025; Yu et al., 2026; Gallegos et al., 2024). Within conservation workflows, these failures may distort attribution records, suppress minority interpretations, or introduce culturally inconsistent rationales into institutional documentation pipelines.

Related work further investigates cultural value alignment, multilingual cultural reasoning, implicit bias, identity-aware reasoning, culturally aware adaptation, and social bias in language and multimodal systems (Tao et al., 2024; Pawar et al., 2024; Adilazuarda et al., 2024; Li et al., 2024; Feng et al., 2025; Yu et al., 2025; Zhao et al., 2025; Wu et al., 2025). Work on museum collection bias and biological artefacts further demonstrates that historical collection practices can systematically shape downstream AI representations and evaluations (Wehi et al., 2012). Representative examples of cultural misalignment and evaluation challenges are summarized in Appendix B. Therefore, cultural validity must be evaluated as part of multimodal evidence grounding rather than treated as an isolated fairness objective.

9 Synthesis: Current Limitations of Conservation AI

Current computer vision systems support recognition, retrieval, and captioning (Castellano and Vessio, 2021; Zhao et al., 2021; Scaringi et al., 2025; Bin et al., 2024; Jiang et al., 2024; Fanelli et al., 2025; Bai et al., 2025). However, visual accuracy alone does not provide conservation reasoning because detected features may correspond to damage, artistic structure, restoration traces, or imaging artifacts. Recognition performance therefore remains insufficient for evidence-grounded conservation workflows.

Scientific imaging extends observation beyond RGB data through hyperspectral imaging, XRF, IR, X-ray, CT, and related modalities (Cucci et al., 2016; Alfeld and Broekaert, 2013; Delaney, 2010). Yet benchmark standardization, cross-modal alignment, and cross-institution transferability remain limited because datasets and acquisition protocols vary substantially across collections.

Knowledge graphs and RAG systems improve grounding by connecting model outputs to institutional records (Barzaghi et al., 2025; Wang and Zhang, 2025; Li et al., 2025; Henrickson et al., 2026; Ramos-Varela et al., 2025; Fan et al., 2025; Li et al., 2026). However, retrieval quality remains bounded by incomplete, historically imbalanced, or contested documentation. Retrieval alone therefore does not guarantee evidence grounding unless uncertainty and provenance conflicts are explicitly modeled.

MLLMs improve multilingual access, descriptive synthesis, and interaction with visual and archival materials (Bin et al., 2024; Ferrato et al., 2025; Tiet et al., 2025; Lian et al., 2024; Cetinic, 2021; Reshetnikov and Marinescu, 2025; Bai et al., 2025). However, they often blur the boundary among visible observations, retrieved evidence, expert inference, and speculation (Jiang et al., 2026; Shu et al., 2025), creating risks for documentation reliability and multimodal alignment.

Governance frameworks identify risks involving copyright, fairness, transparency, and cultural sensitivity (Pansoni et al., 2023; Murphy et al., 2025; Westenberger and Farmaki, 2025; UNESCO, 2021). However, these principles rarely become operational evaluation procedures such as source attribution, audit trails, uncertainty logging, and expert validation.

Taken together, current systems remain fragmented across perception, material evidence, retrieval, multimodal grounding, and governance. They can often recognize, retrieve, or generate cultural information, but they still struggle to support evidence-grounded, uncertainty-aware, and verifiable conservation reasoning.

10 Open Challenges and Research Agenda

The next step is to build multimodal systems that support evidence-grounded and verifiable conservation workflows. Future benchmarks should combine RGB images, scientific imaging, 3D scans, treatment records, provenance metadata, conservation vocabularies, and uncertainty annotations. Tasks may include condition assessment, material-evidence verification, treatment-record retrieval, provenance-grounded summarization, observation–inference classification, and expert-reviewed conservation QA. Evaluation protocols should further distinguish **Direct** conservation tasks from **Adjacent** cultural understanding tasks and explicitly label whether claims are directly observed, scientifically measured, retrieved from records, inferred by the model, or speculative. Such schemas would support finer-grained hallucination analysis and expert review of specific reasoning layers.

RAG and MLLM systems additionally require provenance-faithfulness metrics evaluating archival consistency, retrieval attribution, citation coverage, contradiction with provenance files, and unsupported provenance claims. Provenance faithfulness may become analogous to factuality evaluation in NLP, but adapted to archival uncertainty and incomplete or contested records. AI-generated conservation records also require standardized auditability through model-version tracking, retrieval traces, uncertainty logging, reviewer corrections, and downstream usage records. More broadly, multimodal systems should support multilingual metadata and, where appropriate, culturally grounded review workflows.

Human validation protocols also remain underdeveloped. Evaluation workflows should specify evaluator roles, disagreement handling, uncertainty labels, and failure-case review procedures. Expert validation may involve conservators, curators, conservation scientists, provenance researchers, and domain reviewers depending on the task. Longitudinal workflow studies are additionally needed to determine whether AI systems reduce effort, improve documentation quality, or instead introduce additional review burden.

11 Conclusion

This survey argues that museum conservation AI should be evaluated as evidence-grounded multimodal workflows rather than standalone prediction systems. Current systems remain limited by weak material grounding, incomplete records, hallucination risk, insufficient observation–inference separation, limited expert validation, and under-operationalized governance. While computer vision, scientific imaging, knowledge graphs, RAG, MLLMs, and agents each contribute important capabilities, none independently solves conservation reasoning. Progress therefore depends not only on stronger models, but also on multimodal workflows that support evidence traceability, uncertainty tracking, provenance-aware retrieval, culturally grounded interpretation, expert review, and auditable institutional documentation over time.

Ethical Considerations

Governance as Evaluation

Cultural heritage AI governance has been discussed in ethical frameworks, policy reports, responsible AI initiatives, and copyright analyses (Pansoni et al., 2023; Murphy et al., 2025; Westenberger and Farmaki, 2025; UNESCO, 2021). These works identify risks involving transparency, accountability, authorship, data access, rights, and institutional responsibility. In conservation AI, governance must be operationalized as evaluation. Transparency requires source attribution, audit logs, uncertainty statements, and model-version tracking, while fairness requires multilingual metadata coverage, culturally diverse representation, and sensitivity to historically imbalanced collections.

Copyright and licensing are particularly important because museums often hold complex rights over images, digitized records, and derivative outputs. AI systems trained on or generating from museum data may create uncertainty around reuse, attribution, and downstream publication. Dataset provenance and historical collection imbalance may also propagate institutional, geographic, and cultural biases into downstream multimodal systems and evaluation benchmarks. Governance therefore shapes not only legal compliance but also whether AI-generated conservation records can be stored, shared, verified, or cited. Representative governance frameworks and their operational gaps are summarized in Appendix F.

Hallucination and Evidential Misattribution

MLLMs and retrieval-based systems may generate plausible but unsupported provenance claims, material descriptions, historical interpretations, or treatment rationales. In conservation workflows, such outputs may enter institutional records, collection databases, or future conservation documentation, creating risks of evidential contamination and false attribution. These risks are amplified when fluent multimodal outputs obscure uncertainty, source limitations, or the boundary between direct observation and model inference. Failure to communicate uncertainty may further cause speculative outputs to be interpreted as verified conservation evidence. Human expert validation therefore remains necessary before AI-generated outputs are incorporated into institutional workflows or archival records.

Labor and Institutional Adoption

AI adoption may also change professional workflows and review practices. Work on cultural policy, arts-led AI, digital art labor, wellbeing, and creative sectors highlights that AI systems can reshape expertise, institutional responsibility, and professional oversight (Andrews and Hawcroft, 2024; Duester, 2024; Frost and Vargas, 2024; Rizzi and Bertola, 2025; Connock, 2024; Murphy and Franklin, 2025). In conservation settings, practical concerns include annotation burden, expert review time, accountability for AI-assisted decisions, and the increased workload required to verify generated outputs. Over-reliance on AI-generated summaries or recommendations may also reduce critical expert review and increase the risk of accepting unsupported interpretations. Evaluation should therefore consider workflow usefulness, verification cost, and expert burden in addition to model performance.

Limitations

This survey focuses primarily on museum-centered conservation workflows and does not attempt exhaustive coverage of archaeology, built heritage management, urban preservation, digital humanities, creative AI for art generation, or non-museum heritage institutions. Archaeological site conservation is discussed where it informs conservation AI more broadly (Casillo et al., 2025), but the primary unit of analysis remains multimodal conservation evidence workflows for museum objects.

The surveyed literature also relies heavily on rapidly emerging 2025–2026 work, including preprints and recent proceedings. As MLLMs, multimodal agents, and retrieval systems continue to evolve rapidly, some technical capabilities, evaluation practices, and benchmark assumptions discussed in this survey may change over time.

Many of the evaluation dimensions highlighted in this survey, including evidence grounding, provenance attribution, uncertainty calibration, and expert-in-the-loop validation, still lack standardized benchmarks, annotation protocols, and large-scale expert-reviewed datasets. Expert-centered evaluation may also be difficult to scale because conservation expertise is domain-specific, institution-dependent, and often unavailable for large benchmark construction.

In addition, many conservation datasets containing scientific imaging, treatment histories, prove-

nance records, or expert annotations remain inaccessible due to institutional, legal, preservation, or copyright constraints. Finally, although this survey discusses multilinguality and cultural bias, the current literature remains skewed toward English-language publications and institutions with substantial digitization infrastructure.

References

Manar Abu Talib, Iman Ibrahim, and Manar Anwer Abusirdaneh. 2025. Reusability and benchmarking potential of architectural cultural heritage datasets for generative ai: An analytical study. *Expert Systems with Applications*.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in llms: A survey. In *Proceedings of EMNLP 2024*, pages 15763–15784.

Matthias Alfeld and José A.C. Broekaert. 2013. Mobile depth profiling and sub-surface imaging techniques for historical paintings—a review. *Spectrochimica Acta Part B: Atomic Spectroscopy*, 88:211–230.

Hannah Andrews and Aurora Hawcroft. 2024. Articulating arts-led ai: artists and technological development in cultural policy. *European Journal of Cultural Management and Policy*, 14:12820.

Longju Bai, Angana Borah, Oana Ignat, and Rada Mihalcea. 2025. The power of many: Multi-agent multimodal models for cultural image captioning. In *Proceedings of NAACL-HLT 2025*, pages 2970–2993.

Sebastian Barzaghi, Arianna Moretti, Ivan Heibi, and Silvio Peroni. 2025. Chad-kg: A knowledge graph for representing cultural heritage objects and digitisation paradata. *Preprint*, arXiv:2505.13276.

Federico Becattini, Pietro Bongini, Luana Bulla, Alberto Del Bimbo, and 1 others. 2023. Viscounth: A large-scale multilingual visual question answering dataset for cultural heritage. *ACM Journal on Computing and Cultural Heritage*, 19(6).

Yi Bin, WENHAO SHI, Yujuan Ding, Zhiqiang Hu, Zheng WANG, Yang Yang, See-Kiong Ng, and Heng Tao Shen. 2024. GalleryGPT: Analyzing paintings with large multimodal models. In *ACM Multimedia 2024*.

Mario Casillo, Francesco Colace, Rosario Gaeta, Angelo Lorusso, and Michele Pellegrino. 2025. Artificial intelligence in archaeological site conservation: Trends, challenges, and future directions. *Journal of Computer Applications in Archaeology*, 8(1):224–241.

Giovanna Castellano and Gennaro Vessio. 2021. Deep learning approaches to pattern extraction and recognition in paintings and drawings: an overview. *Neural Computing and Applications*.

Eva Cetinic. 2021. Towards generating and evaluating iconographic image captions of artworks. *Journal of Imaging*, 7(8):123.

Anne Hunnell Chen and 1 others. 2026. Interoperability as equity: Collaborative cultural heritage knowledge graphs as a tool to shape inclusive ontologies. *Journal of Open Humanities Data*, 12.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. Culturalbench: A robust, diverse, and challenging cultural benchmark by human-ai cultural teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701.

Dario Cioni, Lorenzo Berlincioni, Federico Becattini, and Alberto Del Bimbo. 2023. Diffusion based augmentation for captioning and retrieval in cultural heritage. In *ICCV Workshops*.

Francesco Colace, Rosario Gaeta, Angelo Lorusso, Michele Pellegrino, and Domenico Santaniello. 2025. New ai challenges for cultural heritage protection: A general overview. *Journal of Cultural Heritage*, 75:168–193.

Alex Connock. 2024. British tv and ai: explore and exploit. *European Journal of Cultural Management and Policy*, 14:13225.

Costanza Cucci, John K. Delaney, and Marcello Picollo. 2016. Hyperspectral imaging for art investigation. *Physics Reports*, 583:1–89.

John K. et al. Delaney. 2010. Visible and infrared imaging spectroscopy of picasso's harlequin musician. *Applied Spectroscopy*, 64(6):584–594.

Martin Doerr. 2003. The cidoc conceptual reference module: An ontological approach to semantic interoperability of metadata. *AI Magazine*, 24(3):75–92.

Emma Duester. 2024. Digital art work and ai: a new paradigm for work in the contemporary art sector in china. *European Journal of Cultural Management and Policy*, 14:12470.

Tao Fan, Hao Wang, Yuehua Zhao, Shu Zhou, and Bin Shi. 2025. When mllms meet ich: A visual retrieval-augmented generation-based method for intangible cultural heritage image recognition—take shadow puppetry as a case. *Information Processing & Management*.

Nicola Fanelli, Gennaro Vessio, and Giovanna Castellano. 2025. Artseek: Deep artwork understanding via multimodal in-context reasoning and late interaction retrieval. *Preprint*, arXiv:2507.21917.

Ruixiang Feng, Shen Gao, Xiuying Chen, Lisi Chen, and Shuo Shang. 2025. Culfit: A fine-grained cultural-aware llm training paradigm via multilingual critique data synthesis. *arXiv preprint arXiv:2505.19484*.

Alessio Ferrato, Carla Limongelli, Fabio Gasparetti, Giuseppe Sansonetti, and Alessandro Micarelli. 2025. Exploring the potential of multimodal large language models for question answering on artworks. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*.

Marco Fiorucci, Marina Khoroshiltseva, Massimiliano Pontil, Arianna Traviglia, Alessio Del Bue, and Stuart James. 2020. Machine learning for cultural heritage: A survey. *Pattern Recognition Letters*, 133:102–108.

Sophie Frost and Lauren Vargas. 2024. Cultural work, wellbeing, and ai. *European Journal of Cultural Management and Policy*, 14:12825.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1173.

Noa Garcia and George Vogiatzis. 2018. Semantic art understanding with multi-modal retrieval. In *ECCV Workshops*.

Jinchao Ge, Tengfei Cheng, Biao Wu, Zeyu Zhang, Shiya Huang, Judith Bishop, Gillian Shepherd, Meng Fang, Ling Chen, and Yang Zhao. 2026. VaseVQA: Multimodal agent and benchmark for Ancient Greek pottery. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 1154–1167, Rabat, Morocco. Association for Computational Linguistics.

Mathew Henrickson, Eric Atwell, John Stell, Mark Westgarth, Dibyadyuti Roy, and Noorhan Abbas. 2026. Retrieval-augmented generation for natural language art provenance searches in the getty provenance index. *Academia AI and Applications*, 2.

International Council of Museums Committee for Conservation. 2024. Icom-cc.

Lise Jaillant, editor. 2022. *Archives, Access and Artificial Intelligence: Working with Born-Digital and Digitized Archival Collections*. Bielefeld University Press, Bielefeld.

Yanbei Jiang, Krista A. Ehinger, and Jey Han Lau. 2024. Kale: an artwork image captioning system augmented with heterogeneous graph.

Yuechen Jiang, Enze Zhang, Md Mohsinul Kabir, Qianqian Xie, Stavroula Golfomitsou, Konstantinos Arvanitis, and Sophia Ananiadou. 2026. Appear2meaning: A cross-cultural benchmark for structured cultural metadata inference from images. *Preprint*, arXiv:2604.07338.

Alex Kachkine. 2025. Physical restoration of a painting with a digitally constructed mask. *Nature*, 642(8067):343–350.

Dayeon Ki, Rachel Rudinger, Tianyi Zhou, and Marine Carpuat. 2025. Multiple llm agents debate for equitable cultural alignment. *arXiv preprint arXiv:2505.24671*.

Åsa M. Larsson, Barbro Bornsäter, and Marei Hacke. 2025. Developing practices for fair and linked data in heritage science. *npj Heritage Science*.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models. In *Advances in Neural Information Processing Systems*, volume 37.

Jiaang Li, Yifei Yuan, Wenyan Li, Mohammad Aliannejadi, Daniel Hershcovich, Anders Søgaard, Ivan Vulić, Wenxuan Zhang, Paul Pu Liang, Yang Deng, and Serge Belongie. 2026. Ravenea: A benchmark for multimodal retrieval-augmented visual culture understanding. *Preprint*, arXiv:2505.14462.

Jinhao Li, Jianzhong Qi, Soyeon Caren Han, and Eun-Jung Holden. 2025. Musekg: A knowledge graph over museum collections. *Preprint*, arXiv:2511.16014.

Long Lian, Yifan Ding, Yunhao Ge, and 1 others. 2024. Describe anything: Detailed localized image and video captioning. In *CVPR*.

Xiaoyu Lin, Aniket Ghorpade, Hansheng Zhu, and 1 others. 2025. Denseannotate: Enabling scalable dense caption collection for images and 3d scenes via spoken descriptions. *arXiv preprint arXiv:2511.12452*.

Fang Liu, Mohan Zhang, Baoying Zheng, and 1 others. 2023. Feature fusion via multi-target learning for ancient artwork captioning. *Information Fusion*, 97:101811.

Li Lu and Bei Yan. 2026. Syncing minds and machines: Hybrid cognitive alignment as an emergent coordination mechanism in human–ai collaboration. *Academy of Management Review*.

Yue Lu, Chao Guo, Xingyuan Dai, and Fei-Yue Wang. 2024. Artcap: A dataset for image captioning of fine art paintings. *IEEE Transactions on Computational Social Systems*, 11(1):576–587.

Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G. Rodriques, and Andrew D. White. 2023. Paperqa: Retrieval-augmented generative agent for scientific research. *Preprint*, arXiv:2312.07559.

Ben et al. Mildenhall. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. *ECCV*.

Oonagh Murphy and Matthew Franklin. 2025. Editorial: Artificial intelligence: cultural policy, management, education, and research. *European Journal of Cultural Management and Policy*, 15:15030.

Oonagh Murphy, Arts Council England New Technologies, and Innovation Team. 2025. Responsible ai in practice: The journey of arts council england. Technical report, Arts Council England.

Oonagh Murphy and Elena Villaespesa. 2020. Ai: A museum planning toolkit. Discussion paper, Goldsmiths, University of London, London. The Network is funded by the Arts and Humanities Research Council.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, and 3 others. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. In *Advances in Neural Information Processing Systems*.

Silvia Pansoni, Simona Tiribelli, Marina Paolanti, Francesco Di Stefano, Emanuele Frontoni, Elisabetta Malinverni, and Benedetta Giovanola. 2023. Artificial intelligence and cultural heritage: Design and assessment of an ethical framework. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume XLVIII-M-2-2023, pages 1149–1155.

Magdalena Pasikowska-Schnass and Yeong-Sin Lim. 2023. Artificial intelligence in the context of cultural heritage and museums.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2024. Survey of cultural awareness in language models: Text and beyond. *arXiv preprint arXiv:2411.00860*.

Samuel Ramos-Varela, Jaime Bellver-Soler, Marcos Estecha-Garitagoitia, and Luis Fernando D'Haro. 2025. Context or retrieval? evaluating rag methods for art and museum qa system. In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pages 129–136.

Fabio Remondino and Stefano Campana. 2011. Heritage recording and 3d modeling with photogrammetry and 3d scanning. *Remote Sensing*.

Artem Reshetnikov and Maria-Cristina Marinescu. 2025. Caption generation in cultural heritage: Crowdsourced data and tuning multimodal large language models. In *ACL Workshop*.

Greta Rizzi and Paola Bertola. 2025. Exploring the generative ai potential in the fashion design process: an experimental experience on the collaboration between fashion design practitioners and generative ai tools. *European Journal of Cultural Management and Policy*, 15:13875.

Raffaele Scaringi, Giuseppe Fiameni, Gennaro Vessio, and Giovanna Castellano. 2025. Graphclip: Image-graph contrastive learning for multimodal artwork classification. *Knowledge-Based Systems*, 310:112857.

Andrea Schimmenti, Valentina Pasqual, Fabio Vitali, and Marieke van Erp. 2025. Knowledge graphs generation from cultural heritage texts: Combining llms and ontological engineering for scholarly debates. *arXiv preprint arXiv:2511.10354*.

Dong Shu, Haiyan Zhao, Jingyu Hu, Weiru Liu, Ali Payani, Lu Cheng, and Mengnan Du. 2025. Large vision-language model alignment and misalignment: A survey through the lens of explainability. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 1713–1735.

Renan Souza, Amal Gueroudji, Stephen DeWitt, Daniel Rosendo, Tirthankar Ghosal, Robert Ross, Prasanna Balaprakash, and Rafael Ferreira da Silva. 2025. Prov-agent: Unified provenance for tracking ai agent interactions in agentic workflows. *Preprint*, arXiv:2508.02866.

Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Massimiliano Corsini, and Rita Cucchiara. 2019. Artpedia: A new visual-semantic dataset with visual and contextual sentences in the artistic domain. In *International Conference on Image Analysis and Processing*, pages 729–740.

Yan Tao, Olga Viberg, Ryan S. Baker, and René F. Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3:pgae346.

Laya Targa, Carmen Cano, Álvaro Solbes-García, Sergio Casas, Ester Alba, and Cristina Portalés. 2025. Automated edge detection for cultural heritage conservation: Comparative evaluation of classical and deep learning methods. *Applied Sciences*, 15:8260.

Sonja Thiel and Johannes C. Bernhardt, editors. 2024. *AI in Museums: Reflections, Perspectives and Applications*. transcript Verlag, Bielefeld.

Nathalie Tiet, Fabio Rezende de Souza, and Pierre Nugues. 2025. Aspects of image captioning across multimodal large language models. In *ACM Multimedia Workshops*.

UNESCO. 2021. Recommendation on the ethics of artificial intelligence.

Elena Villaespesa and Seth Crider. 2021. A critical comparison analysis between human and machine-generated tags for the metropolitan museum of art's collection. *Journal of Documentation*, 77(4):946–964.

Elena Villaespesa and Oonagh Murphy. 2021. This is not an apple! benefits and challenges of applying computer vision to museum collections. *Museum Management and Curatorship*, 36(4):362–383.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael R. Lyu. 2024. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384.

Yue Wang and Man Zhang. 2025. Cidoc crm-based knowledge graph construction for cultural heritage using large language models. *Applied Sciences*, 15:12063.

Priscilla M. Wehi, Hemi Whaanga, and Steve A. Trewick. 2012. Artefacts, biology and bias in museum collection research. *Molecular Ecology*, 21(13):3107–3112.

Paula Westenberger and Despoina Farmaki. 2025. Artificial intelligence for cultural heritage research: the challenges in uk copyright law and policy. *European Journal of Cultural Management and Policy*, Volume 15 - 2025.

Mark D. et al. Wilkinson. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018.

Xing Wu, Jun Song, Deyu Gao, and 1 others. 2026. Transformer-based image restoration for cultural relics. *Applied Intelligence*.

Yuhang Wu, Wenmeng Yu, Yean Cheng, Yan Wang, Xiaohan Zhang, Jiazheng Xu, Ming Ding, and Yuxiao Dong. 2025. Alignmmbench: Evaluating chinese multimodal alignment in large vision-language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6541–6558.

Haeun Yu, Seogyeong Jeong, Siddhesh Pawar, Jisu Shin, Jiho Jin, Junho Myung, Alice Oh, and Isabelle Augenstein. 2025. Entangled in representations: Mechanistic investigation of cultural biases in large language models. *arXiv preprint arXiv:2508.08879*.

Haorui Yu, Diji Yang, Hang He, Fengrui Zhang, and Qiufeng Yi. 2026. Vulca-bench: A multicultural vision-language benchmark for evaluating cultural understanding. *arXiv preprint arXiv:2601.07986*.

Tianxiu Yu, Cong Lin, Shijie Zhang, and 1 others. 2022. Artificial intelligence for dunhuang cultural heritage protection: The project and the dataset. *International Journal of Computer Vision*, 130:2646–2673.

Raoyuan Zhao, Beiduo Chen, Barbara Plank, and Michael A. Hedderich. 2025. Makieval: A multilingual automatic wikidata-based framework for cultural awareness evaluation for llms. *arXiv preprint arXiv:2505.21693*.

Wentao Zhao, Dalin Zhou, Xinguo Qiu, and Wei Jiang. 2021. How to represent paintings: A painting classification using artistic comments. *Sensors*, 21(6):1940.

Baoying Zheng, Fang Liu, Mohan Zhang, and 1 others. 2023. Image captioning for cultural artworks: a case study on ceramics. *Multimedia Systems*, 29:3223–3243.

Ping Zhong, Wenjin Hu, Yinqiu Zhao, and Fujun Zhang. 2026. Geo-tcam: a thangka captioning method integrating topic modeling with geometry-guided spatial attention. *npj Heritage Science*.

A Example Workflow: Multimodal Evidence Infrastructure

Running Example: A Cracked Ceramic Vessel as Evidence Infrastructure

A museum holds a ceramic vessel with visible cracking, pigment loss, and an uncertain prior treatment history. The artifact is documented through RGB photography, a 3D surface scan, hyperspectral imaging, prior treatment records, provenance files, environmental monitoring logs, and controlled vocabulary terms.

A computer vision model may segment crack regions and detect surface losses. Scientific imaging may support pigment identification or reveal subsurface changes. A 3D model may preserve geometry for condition comparison and mount design. A knowledge graph or RAG system may connect the vessel to related treatment records, provenance histories, comparable objects, and conservation vocabularies. An MLLM may help summarize the records for a conservator or generate questions for further inspection.

However, the final conservation judgment cannot be delegated to the model. Experts must determine whether cracks indicate active deterioration or stable historical damage, whether pigment change is chemically significant, whether prior intervention constrains treatment options, and whether provenance uncertainty creates cultural or institutional sensitivity. In this workflow, AI functions as evidence infrastructure: it organizes and proposes connections among observations, records, and interpretations, but expert validation, uncertainty logging, and governance review remain central.

B Examples of Cultural Bias and Misalignment

C Representative Datasets and Benchmarks

Challenge	Observed Failure	Potential Impact	Related Technical Issue
Cross-cultural captioning	Descriptions rely on visually similar but culturally incorrect analogies	Misleading metadata and culturally inappropriate interpretation	Surface-level multimodal alignment
Multilingual cultural reasoning	Uneven performance across languages and regions	Loss of culturally specific meaning in multilingual archives	Language imbalance and representation bias
Provenance interpretation	Models over-associate objects with dominant historical narratives	Distorted attribution and provenance documentation	Statistical prior over-reliance
Collection bias	Training data reflects historically imbalanced collections	Suppression of minority or underrepresented perspectives	Dataset bias and historical imbalance
Restoration recommendation	Generated rationales ignore contextual or community-sensitive factors	Culturally inappropriate conservation decisions	Weak contextual grounding

Table 1: Representative examples of cultural bias and multimodal misalignment in conservation-related AI workflows.

Resource / Work	Modality	Task	Relevance	Access / Scale	Conservation Limitation
SemArt, Artpedia	image + text	retrieval, visual-semantic learning	**Adjacent**	public research datasets	limited material evidence, no treatment history
ArtCap, iconographic captioning	image + captions	artwork captioning	**Adjacent**	dataset / task-specific	descriptive fluency rather than conservation reliability
VISCOUNTH	image + multilingual VQA	cultural VQA	**Adjacent**	large-scale VQA	weak linkage to conservation documentation
Dunhuang protection dataset	images, heritage records	heritage protection	**Direct**	project-specific	domain-specific, difficult generalization
Edge detection for conservation	images	edge/damage cue detection	**Direct**	experimental	narrow visual feature focus
DenseAnnotate / 3D captioning	images, 3D scenes, speech	dense annotation	**Adjacent**	emerging	annotation method not conservation-specific
Architectural CH datasets	spatial / architectural	generative AI benchmarking	**Adjacent**	dataset analysis	site/building focus rather than museum objects
Cultural-bias benchmarks	text / multimodal	cultural reasoning	**General Heritage**	benchmark-specific	relevant to interpretation, not material conservation

Table 2: Datasets and benchmarks by conservation relevance. Many resources are useful enabling technologies but do not directly evaluate conservation decision support.

D Scientific Imaging Modalities and AI Challenges

E Modeling Paradigms and Conservation Tensions

Imaging / Data Type	Evidence Provided	Typical Conservation Use	Relevance	AI Challenge
RGB photography	surface appearance, visible damage	documentation, visual comparison	**Direct / Adjacent**	lighting bias, semantic overinterpretation
Hyperspectral imaging	spectral/material signatures	pigment mapping, degradation cues	**Direct**	cross-instrument calibration, high-dimensional fusion
XRF / elemental mapping	elemental composition	material identification, hidden layers	**Direct**	spatial registration, expert interpretation
Infrared reflectography	underdrawings, subsurface features	technical art history, treatment planning	**Direct**	linking hidden evidence to decisions
X-ray / CT	structure, density, internal features	structural assessment, hidden damage	**Direct**	limited public benchmarks, acquisition variability
3D scanning / photogrammetry	geometry, volume, surface deformation	condition comparison, digital twins	**Direct / Adjacent**	conservation-grade fidelity, long-term reproducibility
Text records	treatment history, provenance, notes	audit and decision support	**Direct**	incomplete records, uncertainty and bias

Table 3: Scientific and spatial evidence in conservation AI. The technical challenge is not only modality-specific prediction, but registration, uncertainty tracking, and expert-interpretable fusion.

Paradigm	Inputs	Outputs	Relevance	Main Conservation Tension
Computer vision	RGB, scientific images	labels, masks, features	**Direct / Adjacent**	detects patterns but may not justify conservation significance
Scientific imaging ML	HSI, XRF, IR, X-ray, CT	material/ degradation maps	**Direct**	material evidence requires expert interpretation
3D / spatial AI	scans, photos, point clouds	reconstructions, digital twins	**Direct / Adjacent**	visual fidelity may not equal measurement reliability
Knowledge graphs	metadata, ontologies, records	structured relations	**Adjacent / Direct**	formalization can reproduce incomplete or biased records
RAG	archives, provenance files, images	grounded answers	**Adjacent**	retrieval depends on incomplete institutional corpora
MLLMs	image + text + prompts	explanations, QA, summaries	**Adjacent**	fluency can hide hallucination and inference collapse
Agents	tools, memories, models	multi-step plans	**Adjacent**	accountability and audit become harder

Table 4: Modeling paradigms and conservation relevance. The survey distinguishes direct conservation tasks from enabling technologies that become relevant when embedded in conservation workflows.

F Governance Frameworks and Operational Gaps

G Evaluation Dimensions for Conservation AI

Framework / Area	Scope	Legal / Ethical Focus	Conservation Relevance	Operational Gap
UNESCO AI Ethics	global AI governance	ethics, human rights, accountability	**General Heritage**	broad principles need museum-specific protocols
Responsible AI in arts	cultural institutions	responsible deployment	**General Heritage / Adjacent**	limited conservation-specific metrics
UK copyright analysis	CH research and AI	copyright, licensing, policy	**Adjacent**	rights constraints must map to data workflows
Ethical CH AI frameworks	cultural heritage systems	fairness, transparency, accountability	**Adjacent**	rarely specify audit thresholds
CIDOC CRM / FAIR	semantic/data infrastructure	interoperability, reusability	**Direct / Adjacent**	does not by itself validate AI-generated claims
Institutional audit logs	conservation records	traceability, accountability	**Direct**	need standardization for model-generated records

Table 5: Governance frameworks and their operational gaps for conservation AI. Ethical principles must become measurable evaluation and audit protocols.

Dimension	What It Measures	Example Metric / Protocol	Why Generic ML Metrics Are Insufficient
Technical accuracy	Classification, segmentation, retrieval, caption quality	IoU, F1, Recall@K, CIDEr, exact/partial match	high task accuracy does not ensure conservation reliability
Material-evidence alignment	Whether claims are grounded in visual or scientific evidence	expert imaging review, evidence-label matching	standard metrics rarely evaluate material grounding
Provenance grounding	Consistency with archival and provenance records	source attribution, provenance consistency checks	fluent answers may fabricate provenance
Observation–inference separation	Whether visible evidence, retrieved evidence, inference, and speculation are separated	structured annotation schema; expert adjudication	MLLMs often merge observation with interpretation
Hallucination rate	Unsupported names, dates, materials, provenance, or treatment claims	unsupported-claim annotation; contradiction analysis	generic factuality benchmarks ignore conservation evidence
Uncertainty calibration	Whether confidence reflects evidential strength	calibration curves; confidence vs expert judgement	conservation decisions require explicit uncertainty tracking
Expert agreement	Alignment with conservators, curators, scientists, or community experts	inter-rater agreement; Cohen's/Fleiss' kappa; disagreement logs	single-reference labels ignore interpretive plurality
Workflow usefulness	Practical value in real conservation tasks	task completion, time saving, expert usefulness study	offline metrics do not measure institutional integration
Cultural validity	Respect for multilingual, community, contested, or Indigenous knowledge contexts	culturally grounded review; community consultation	standard benchmarks underrepresent contested heritage
Legal/governance compliance	Copyright, data access, provenance policy, institutional accountability	audit checklist; licensing review; governance protocol	ML metrics do not encode legal or policy constraints
Reproducibility/ auditability	Whether outputs can be traced, reproduced, corrected, and archived	decision logs; model/version records; retrieval traces	black-box outputs reduce institutional trust

Table 6: Evaluation dimensions for conservation AI. The framework operationalizes the claim that conservation AI should be evaluated as evidence infrastructure rather than prediction technology.