TCFP-Sight: Ontology-Gated DAIE Reasoning for Auditable Appreciation of Traditional Chinese Figure Painting

Anonymous Author(s)

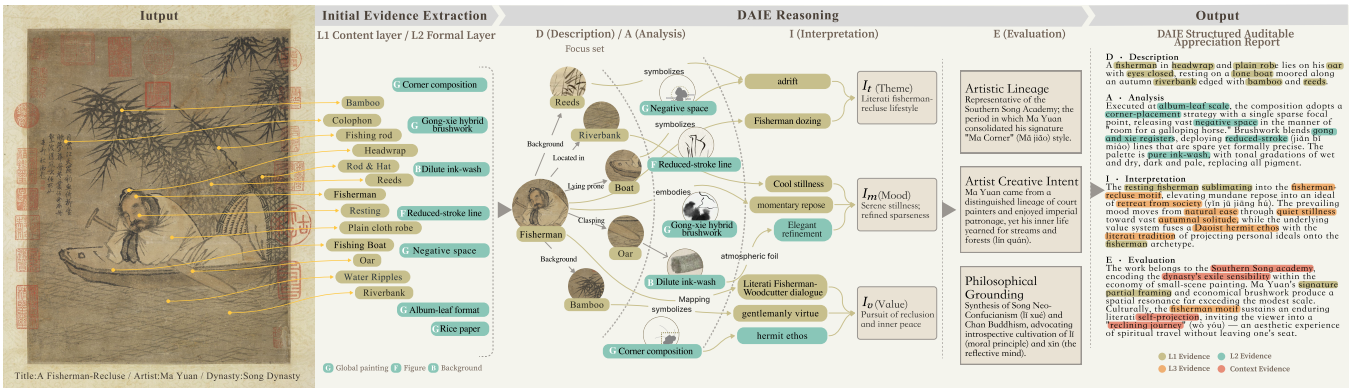

Figure 1: TCFP-Sight traces the DAIE reasoning path over Fisherman Resting on Autumn River (Ma Yuan, Southern Song): from L1/L2 visual evidence → Focus Set (D) → formal-language analysis (A) → theme–mood–value interpretation (I) → cultural-historical evaluation (E), producing a fully auditable appreciation report.

Abstract

Enabling artificial intelligence to transcend surface-level visual perception and achieve deep cultural-semantic understanding of Chinese figure painting is a core open challenge in Cultural Heritage Computing. Although multimodal large language models (MLLMs) can generate fluent art commentary, they systematically fail at cross-level cognitive transitions—producing conclusions that are semantically coherent yet lack grounding in low-level visual evidence—a failure mode we term cultural hallucination. Existing retrieval-augmented generation (RAG) approaches, moreover, lack stage-structured reasoning constraints and domain-specific ontological support. We therefore present **TCFP-Sight** (Traditional Chinese Figure Painting Sight), an ontology-gated multi-stage reasoning framework for auditable appreciation of Chinese figure painting. The system is grounded in a structured prior knowledge space integrating Panofskyan iconology with Chinese painting theory into a computable semantic schema, instantiated across a knowledge graph of 1,323 annotated figure paintings spanning the Tang through Qing dynasties. An Ontology Gating mechanism prevents premature activation of high-level cultural meaning before low-level evidential foundations are secured, mandating that all output claims carry explicit evidence links and confidence scores. Evaluated on a newly constructed benchmark of 100 classical paintings, TCFP-Sight achieves consistent gains over generic MLLMs and RAG baselines across appreciation quality and hallucination suppression dimensions, while improving evidence traceability. Ablation studies further confirm that Ontology Gating is the most critical anti-hallucination component, and that the staged reasoning pipeline is an indispensable complement to domain knowledge-driven architectures.

CCS Concepts

• **Computing methodologies** → **Knowledge representation and reasoning**.

Keywords

Cultural Heritage Computing, Traditional Chinese Figure Painting, Iconology, Ontology Gating, Appreciation Reasoning

1 Introduction

Enabling artificial intelligence to transcend surface-level visual description and achieve deep cultural-semantic understanding of Chinese figure painting stands as one of the most pressing open challenges in Cultural Heritage Computing [18, 27]. As a primary visual medium of Chinese civilization, Chinese figure painting encodes intricate narrative logic, aesthetic paradigms, and philosophical meaning [28]. From the perspective of iconology and art-critical methodology, the complete appreciation of a figure painting is by no means a singular object-recognition task; it constitutes a rigorous, progressive chain of cognitive reasoning—from the objective description of visible content, through the analysis of formal language and brushwork structure, to the interpretation of iconographic motifs and cultural meaning, and finally to a comparative evaluation situated within its historical context [23, 24]. This layered cognitive structure forms the methodological foundation of art-historical inquiry. Any semantic leap across levels, or any inference drawn before lower-level visual evidence has been sufficiently established, undermines the traceability upon which credible appreciation depends.

Conference'17, Washington, DC, USA
2026. ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Recent years have seen MLLMs achieve remarkable performance on general visual understanding tasks, with growing applications to cultural heritage image analysis [4]. Yet existing research consistently reveals a systemic deficiency: while MLLMs demonstrate perceptual competence in recognizing image content, they struggle to correctly reason through cultural meaning, frequently breaking down at cross-level transitions in the cognitive reasoning chain [21]. This limitation is particularly acute in Chinese figure painting. The domain's distinctive expressive logic renders the meaning of visual symbols highly context-dependent: the combination of pine, *qin* zither, and mountain dwelling, for instance, activates the cultural theme of reclusion (*yinyi*) rather than a neutral landscape classification; brushwork and ink carry strong affective and aesthetic orientation; and the interpretive stance of a work must be rigorously anchored to its specific dynastic context [28]. These domain-specific characteristics constitute barriers that general-purpose models cannot readily overcome.

Although RAG partially mitigates these issues by incorporating external knowledge bases [30], existing methods still exhibit two core deficiencies. First, a reasoning paradigm gap: the one-shot generation lacks staged constraints, causing high-level cultural interpretations to be triggered before low-level visual evidence is established. Second, a knowledge base gap: existing repositories built on Western frameworks lack ontological coverage of the core generative logic in Chinese painting theory—evolving from brushwork to *qiyun* (rhythmic vitality) and *yijing* (artistic conception) [20]. Consider a *gaoshi* painting from the Ming–Qing transition: a generic MLLMs describes "a serene, Zen-inflected mood" despite desolate brushwork (L2 misattribution), maps the recluse–pine composition onto *linquan gaozhi* rather than the historically appropriate loyalist (*yimin*) stance, and produces no claim traceable to visual evidence. These two deficiencies converge in the domain of Chinese figure painting to produce a distinctive failure mode: **cultural hallucination**, in which system outputs are semantically coherent yet cannot be meaningfully grounded in actual visual evidence or historical context [21].

Overcoming cultural hallucination and achieving auditable computational art appreciation calls for a structural rethinking: recasting appreciation from black-box one-shot generation into a structured reasoning chain of discrete, sequentially activated cognitive stages, each governed by independent knowledge-access constraints and standardized output formats. We present **TCFP-Sight** (Traditional Chinese Figure Painting Sight)—an ontology-gated reasoning framework that structures appreciation as four progressive stages: **D**escription , **A**nalysis **I**nterpretation, and **E**valuation, collectively termed **DAIE**[3, 25, 26] —for auditable appreciation of Chinese figure painting. The framework is grounded in a prior knowledge space comprising a domain ontology that unifies iconological and Chinese-painting-theory categories, an ontology-aligned knowledge graph supplying instance evidence and historical context, and a symbolic rule base enabling controlled cross-level semantic triggering. An Ontology Gating mechanism enforces hard semantic boundaries across stages, ensuring that no higher-level cultural meaning is activated until the evidential basis at the current stage has been firmly established [20]. The main contributions of this paper are as follows:

- We propose **ontology-gated DAIE reasoning**—a four-stage appreciation pipeline of Description, Analysis, Interpretation, and Evaluation enforced by an Ontology Gating mechanism—that prevents premature activation of high-level cultural meaning before lower-level evidential foundations are established.
- We propose a **Chinese figure painting knowledge space** integrating Panofskian iconology with Chinese painting theory to address the ontological gap in formal-language modeling for East Asian painting absent from existing Western-oriented art knowledge bases.
- We validate the effectiveness of **auditable DAIE appreciation reasoning** via a newly constructed 100-painting benchmark, demonstrating consistent gains over MLLM and RAG baselines across six dimensions spanning appreciation quality, hallucination suppression, and evidence traceability.

2 Related Work

We survey three threads directly bearing on our work—computational iconology, MLLMs for art appreciation, and RAG for cultural heritage—noting that systematic integration across all three remains limited, particularly in cross-level cultural reasoning and appreciation traceability.

2.1 Computational Iconology and Chinese Painting Theory

Panofsky's iconological paradigm establishes a progressive logic of semantic construction from surface to depth: from the experiential recognition of natural objects, through the semantic decoding of iconographic motifs and narrative themes, to the deep cultural foundations and historical spirit of the work [22, 23]. This highly stratified cognitive schema carries substantial computational potential. Following this trajectory, Impett and Moretti were the first to translate core iconological categories into quantifiable discrete operations, confirming the technical feasibility of migrating humanistic interpretive theory into formal computational methods [14]. Subsequent research has achieved sustained and substantial progress across iconographic taxonomy construction [7], AI operability translation [10], semantic annotation of visual cultural heritage [1, 11, 24], and metadata interoperability [31].

The interpretive logic of Chinese painting theory aligns closely with Panofsky's three-layer framework, it harbours an independent and comprehensive formal language system of its own. As Xie He's "structural use of the brush" (*gǔfǎ yòngbǐ*), codified as the second of the Six Principles, establishes [13], brushwork and ink tonality in Chinese painting are not merely technical categories but the primary visual medium through which a painter's personality and deeper cultural meaning are conveyed [12]. Although the ICON ontology [24] represents the most systematic computational operationalisation of Panofsky's tiers to date, its scope is confined to Western iconographic tradition; existing work targeting Chinese painting [5, 40] likewise lacks both a formal-language-layer ontology and explicit symbolic mapping mechanisms that anchor painting-theory vocabulary to low-level visual features.

2.2 Multimodal LLMs for Art Appreciation

Multimodal large language models (MLLMs) have emerged as a primary driver of automated art analysis by jointly processing images and text [33]. General-purpose models such as GPT-4V, LLaVA, and Qwen-VL now reliably handle object recognition, compositional description, and basic affective inference [2, 19, 37]; domain-adapted counterparts further extend this capacity—through adapter injection or large-scale art corpus fine-tuning—to affective expression, formal style analysis, and large-scale pictorial understanding [4, 16, 35]. Decomposed, evidence-chain prompting has additionally been shown to suppress hallucination and improve alignment with human aesthetic judgment [15], yet the applicability of these advances remains largely confined to general Western art corpora.

Benchmark studies on Chinese painting and indigenous cultural imagery reveal a substantial performance gap between state-of-the-art MLLMs and human annotators on metaphorical, phonosemantic, and iconographically deep understanding tasks, while cross-cultural interpretation studies further show that models systematically fall into iconographic level confusion and historical context misattribution [34, 36, 38, 39].

The core bottleneck, in short, is not perceptual but *semantic corroboration failure*—the inability to ascend the evidential chain from visible elements to culturally grounded iconological interpretations, which underlies both cultural hallucination and interpretive non-auditability [4, 34].

2.3 Retrieval-Augmented Generation for Cultural Heritage

Retrieval-augmented generation (RAG) mitigates hallucination by injecting external knowledge at inference time, decoupling parametric knowledge from factual retrieval [17]. KG-RAG advances this by organising retrieved chunks around graph structure, enabling relation-aware association that improves precision on complex, multi-hop queries [41]; GraphRAG further extends cross-document reasoning through community summarisation, synthesising globally coherent answers beyond any single retrieved passage [9].

In the cultural heritage domain, ArtRAG incorporates structured contextual knowledge into visual art understanding [30], while graph-enhanced frameworks have been validated for intangible cultural heritage inheritor information retrieval [29] and domain knowledge question answering [32]. CIDOC-CRM[6] provides the ontological foundation supporting semantic interoperability across these systems [8].

However, existing KG-RAG systems treat retrieval as a stage-blind global operation: premature surfacing of high-level cultural meaning candidates causes cross-level semantic drift before the evidential basis is secured [41]. Moreover, general-purpose ontologies such as CIDOC-CRM lack the domain-specific formal vocabulary of Chinese painting theory—covering neither brushwork classification nor ink-wash conventions—and thus cannot enforce precise semantic constraints across reasoning stages [8].

These limitations collectively point to the need for an integrated approach that provides formal-language ontological coverage, cross-level reasoning traceability, and stage-specific knowledge gating.

3 Prior Knowledge Space

To support constrained and auditable appreciation reasoning, we formalize the system's knowledge infrastructure as a prior knowledge space $\mathcal{K} = (\mathcal{O}, \mathcal{G}, \mathcal{R})$: $\mathcal{O}$ enforces ontology gating, $\mathcal{G}$ supplies instance evidence and historical calibration, and $\mathcal{R}$ triggers cross-layer semantic candidates. as detailed in Sections 3.1–3.3.

3.1 TCFP Ontology ($\mathcal{O}$)

The TCFP ontology $\mathcal{O}$ constitutes the conceptual grammar of the appreciation reasoning process. Its theoretical foundation integrates Panofsky's three-layer iconological pathway— pre-iconographical description, iconographical analysis, and iconological interpretation [23]—with the core categories of Chinese painting theory: brushwork (*gufa yòngbǐ*), rhythmic vitality (*qìyùn*), and artistic conception (*yìjìng*) [12, 13]. $\mathcal{O}$ is compatible with ICON[24] and CIDOC-CRM[6] while extending both with an *L2 Formal Layer* that ontologises the brushwork–*qìyùn*–*yìjìng* logic specific to East Asian painting. This integration translates the progressive interpretive logic inherited from art history and painting theory into an annotatable, compositional, and formally reasoned semantic schema suited to classical Chinese figure painting.

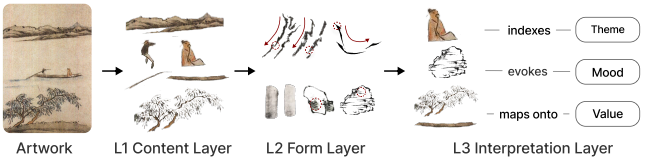

Figure 2: $\mathcal{O}$ represents a painting as three progressive semantic layers: L1 (*what* is depicted), L2 (*how* it is painted), and L3 (*why*— cultural meaning and value).

Structurally, $\mathcal{O}$ organizes domain knowledge into three progressive layers (Figure 2): the **L1 Content Layer** encodes observable visual entities, scene structures, and spatial relations; the **L2 Formal Layer** captures compositional principles, brushwork, ink tonality, and colorization; the **L3 Meaning Layer** encompasses iconographic motifs, narrative themes, aesthetic mood, and value orientations. Each layer is instantiated with a controlled vocabulary of 1,171 terms across 30 third-level categories under 8 second-level classes (L1: 639 terms; L2: 121 terms; L3: 411 terms), defining admissible slots, relation types, and enumerable labels to support annotation consistency and stage-specific retrieval.

Functionally, the operative mechanism of $\mathcal{O}$ is **Ontology Gating**: following the unidirectional layer progression in Figure 2, the Description stage is restricted to L1 fields; Analysis to L2; only at Interpretation and Evaluation does the system enter the L3 semantic space. Stage-dedicated system prompts enforce this as hard constraints, ensuring that no high-level cultural meaning is activated before its lower-level evidential basis has been firmly established.

3.2 TCFP Knowledge Graph ($\mathcal{G}$)

$\mathcal{G}$ constitutes the instance evidence layer of the system. Following the taxonomic structure of $\mathcal{O}$, we semi-automatically annotate 1,323 Chinese figure paintings by 488 artists using multi-source data comprising 36 authoritative painting catalogues, museum archival texts, and encyclopaedic sources, producing structured JSON records that are subsequently instantiated as a graph-structured knowledge

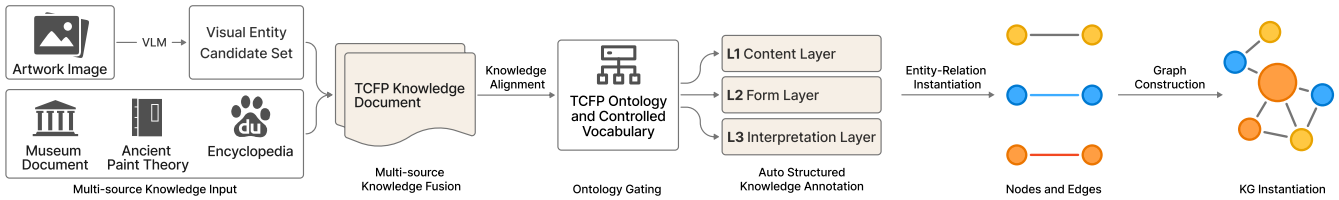

Figure 3: Construction pipeline of TCFP-KG. Dual input streams—painting images (VLM-extracted L1/L2 candidates) and textual sources —are aligned through the TCFP Ontology and Controlled Vocabulary into three annotation layers, then instantiated as a Neo4j graph with typed evidence nodes and support edges.

space (Figure 3). The dynastic coverage (Five Dynasties through Qing) is deliberately designed to span the principal stylistic periods, thematic traditions, and contextual shifts of Chinese painting history, thereby enabling historically calibrated interpretation.

L1–L3 concept nodes are linked by entity, spatial, co-occurrence, support, counter-evidence, and cross-artwork comparative edges; colophons, poems, and painting-theory passages are represented as independent text-evidence nodes, enabling explicit provenance citation in interpretive chains. $\mathcal{G}$ serves three functions within the pipeline: *stage-specific subgraph retrieval* (each DAIE stage accesses its corresponding semantic-layer subgraph), *interpretation calibration* (verifying candidates against historical co-occurrence patterns and counter-evidence relations), and *cross-artwork comparison* for contextualised evaluation.

3.3 Symbolic Mapping Rules ($\mathcal{R}$)

$\mathcal{R}$ constitutes the cross-layer triggering mechanism, distilled from regularities observed in $\mathcal{G}$. Each rule takes the form

$$c_{L1}\ [\wedge\ c_{L2}]\ \Rightarrow\ y_{L3}, \tag{1}$$

where $c_{L1} \subseteq C_1$ denotes an L1 content-cue combination (figures, objects, scenes, actions), $c_{L2} \subseteq C_2$ optional L2 formal cues, and y_{L3} the triggered L3 motif, theme, mood, or value candidate. The operative insight is combinatorial: the co-occurrence of *literatus–qín–pine–mountain dwelling* triggers the reclusion (*yǐnyì*) candidate far more reliably than any single cue in isolation.

Each rule is configured with a **Minimal Evidence Set** (MES)—the minimum cue combination required to activate agiven candidate—and an initial weight encoding its iconographic authority, yielding a ranked candidate set with conditional constraints rather than a forced single conclusion. $\mathcal{R}$ thus functions as a knowledge-constrained forward hypothesis generator: it nominates L3 candidates from structured L1–L2 observations, which $\mathcal{G}$ subsequently validates against historical instances, support chains, and counter-evidence relations.

4 DAIE Reasoning Framework

4.1 Overview

Grounded in $\mathcal{K}$, TCFP-Sight models appreciation as a four-stage DAIE pipeline rather than one-shot free-form generation. Formally:

$$\mathcal{Y}_{\text{DAIE}} = \mathcal{F}(X; \mathcal{K}) \tag{2}$$

where X is the input artwork with metadata, the prior knowledge space $\mathcal{K}$, and $\mathcal{Y}_{\text{DAIE}}$ is a structured DAIE report in which every claim carries evidence links, confidence scores, and uncertainty markers. The overall pipeline architecture is illustrated in Figure 4. Key notation is summarised in Table 1.

Table 1: Key Notation

Symbol	Meaning	Symbol	Meaning
X	artwork and metadata	I_t, I_m, I_v	I (Theme / Mood / Value)
$\mathcal{K}$	Prior knowledge space	F_D	D-stage content focus set
$\mathcal{O}$	TCFP ontology	$\mathcal{Y}_{\text{DAIE}}$	Structured DAIE output
$\mathcal{G}$	TCFP knowledge graph	$\Phi_{\mathcal{G}}(\cdot)$	KG correction operator
$\mathcal{R}$	Symbolic mapping rules	$\Psi_{\mathcal{R}}(\cdot)$	Rule-triggered generator
O_{L1-L3}	Ontology layer	C_{L1}, C_{L2}	Full candidate set
C_D, R_D	Entity, Relation	L_A, M_A	micro/macro-level formal

4.2 Multimodal Input Representation and Initial Evidence Extraction

The system receives the target image X and structured metadata (title, artist, dynasty) as a dual-modal input. We deploy the state-of-the-art **Gemini 3.1 Pro** strictly as a front-end perception engine to extract observable candidates from X while deliberately suppressing high-level cultural interpretations. Concurrently, the TCFP ontology $\mathcal{O}$ and its controlled vocabulary provide two complementary interventions: **constraint** (locking outputs to predefined slots and labels) and **semantic enhancement** (mapping open-vocabulary predictions to standardised terms), ensuring cross-stage annotation consistency and retrieval efficacy.

This process yields two foundational candidate sets (Figure 4, Front-end Observation). The C_{L1} content candidate set enumerates objectively observable visual entities—figures, objects, landscape elements, and their spatial relations; the C_{L2} formal candidate set deconstructs the painting's formal language, covering macro-level compositional layout and micro-level brushwork and colorimetry. All candidates are bound to source region labels and initial confidence scores.

4.3 Description (D)

The D stage distils the full C_{L1} candidate set (Section 4.2) into a focus set $F_D = [C_D, R_D]$—the minimal *figure–object–scene* combination C_D and their interactions R_D that constitute the descriptive backbone of the painting.

Focus extraction proceeds in three steps. **(1) Anchor identification**: an MLLM integrates visual-salience priors from X with metadata cues to identify the primary figure; $\mathcal{O}$ then normalises and classifies the result according to its role hierarchy. **(2) Relation-centred expansion**: entities in explicit interaction or spatial relation with primary figure are admitted from the L1 candidate set, yielding the preliminary focus set $F_D^{(0)}$ **(3) KG salience validation**: the KG correction operator $\Phi_{\mathcal{G}}$ matches $F_D^{(0)}$ and $CL1$ against

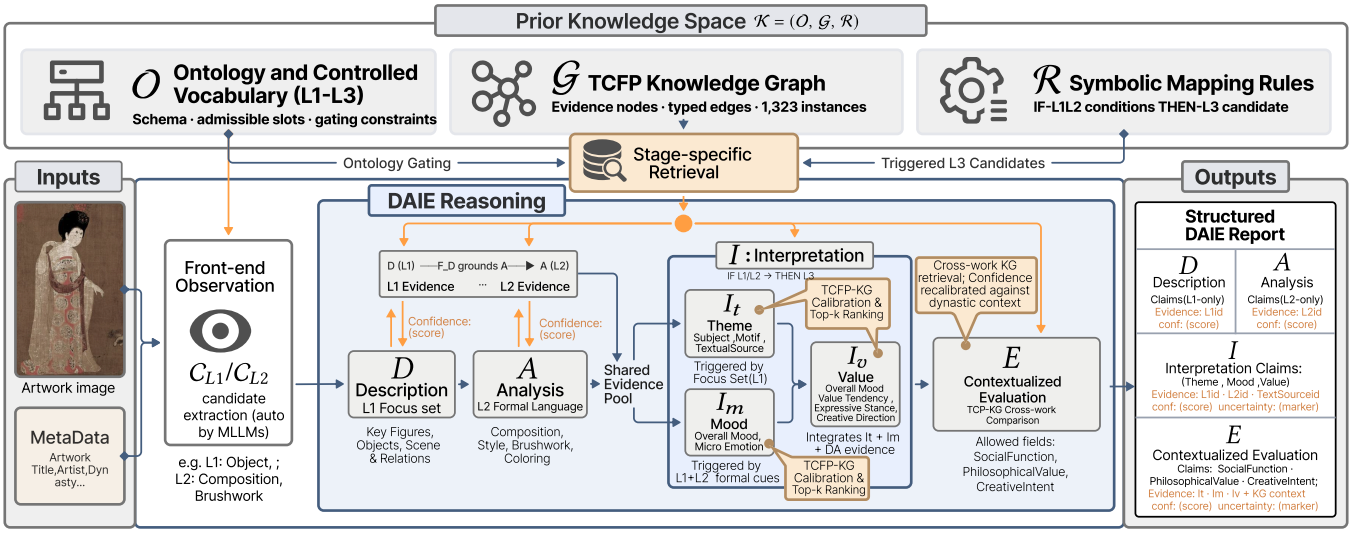

Figure 4: Overall architecture of TCFP-Sight. The system operates under a shared prior knowledge space $\mathcal{K} = (\mathcal{O}, \mathcal{G}, \mathcal{R})$ providing field constraints, evidence retrieval, and symbolic triggering. Given an input artwork X and its metadata, the front-end observation module extracts ontology-constrained C_{L1}/C_{L2} candidates. These feed into four ontology-gated stages—Description (L1 only), Analysis (L2 only), Interpretation ($I_t/I_m/I_v$, L3), and Evaluation—each retrieving stage-specific evidence subgraphs from TCFP-KG. The output is a structured DAIE report in which every claim carries evidence links, confidence scores, and uncertainty markers.

comparable-work patterns in $\mathcal{G}$, suppressing perceptually incidental entities and recovering iconographically expected entities that already reside in $CL1$ but fell below the front-end confidence threshold, scored by co-occurrence frequency with the anchor figure across KNN-retrieved comparable works, yielding:

$$F_D = \left[C_D,\ R_D\right] = \Phi_{\mathcal{G}}\left(F_D^{(0)}, C_{L1}\right)\Big|_{\mathcal{O}_{L1}}, \quad F_D \subseteq C_{L1} \tag{3}$$

F_D is rendered in a $\langle$*Who, Where, Action, Relations*$\rangle$ format with evidence links and confidence scores, strictly gated to L1 fields by $\mathcal{O}$, establishing the auditable factual backbone for subsequent DAIE stages.

4.4 Analysis (A)

The A stage analyses *how* the painting was constructed, restricted to L2 fields and anchored on A:

$$A = \Phi_{\mathcal{G}}\left[L_A(C_D),\ M_A(R_D)\right]\Big|_{\mathcal{O}_{L2}}, \quad A \subseteq C_{L2} \tag{4}$$

where L_A captures micro-level formal units—brushwork, ink tonality, and colorimetry applied to C_D entities—and M_A the macro-level compositional structure—spatial layout, density–void rhythm, and visual flow—organised around R_D relations. $\Phi_{\mathcal{G}}$ further validates L_A and M_A against historical formal conventions in $\mathcal{G}$.

Following a *form-first* principle, L_A and M_A serve as high-priority formal evidence for mood interpretation I_m in the Interpretation stage, ensuring that affective claims are grounded in observable pictorial organisation rather than free-floating sentiment.

4.5 Interpretation (I)

The I stage crosses from visible facts (L1,L2) into the L3 semantic space. To address the complex semiotic system of Chinese figure painting, we decompose I into three coordinated branches—I_t (Theme), I_m (Mood), and I_v (Value)—sharing the pipeline structure

$$I_x = \Phi_{\mathcal{G}}\Big(\Psi_{\mathcal{R}}(\cdot)\Big)\Big|_{\mathcal{O}_{L3}}, \qquad x \in \{t,\ m,\ v\} \tag{5}$$

where $\Psi_{\mathcal{R}}$ is the rule-triggered candidate generator, $\Phi_{\mathcal{G}}$ the KG-based cross-validator, and $(\cdot)$ denotes branch-specific trigger inputs. The three branches differ in their trigger inputs, semiotic dimensions, KG validation logic, and target L3 output slots, as summarised in Table 2.

Crucially, I_v (Value) synthesises all prior outputs (F_D, A, I_t, I_m) for global semantic integration. For instance, when a recluse motif (I_t) co-occurs with an austere brushwork register (A) and a cold, isolated mood (I_m), $\Psi_{\mathcal{R}}$ generates competing candidates including *yimin* loyalist melancholy versus generic mountain-forest aspiration (*linquan gaozhi*). $\Phi_{\mathcal{G}}$ then applies genealogical validation to verify which stance is historically plausible given the specific dynastic context.

To support auditable outputs, the system assigns a claim-level confidence score to each interpretation candidate c before generating the final report. This score synthesizes four signals: (1) aggregated confidence of supporting L1/L2 observations; (2) activation strength of triggered rules in $\Psi_{\mathcal{R}}$; (3) consistency score from the graph validation operator $\Phi_{\mathcal{G}}$; and (4) evidence chain completeness (i.e., explicit links to content/formal evidence and available textual nodes). Formally, it is defined as:

$$\mathrm{Conf}(c) = \lambda_1 s_{\mathrm{obs}}(c) + \lambda_2 s_{\mathrm{rule}}(c) + \lambda_3 s_{\mathrm{graph}}(c) + \lambda_4 s_{\mathrm{evd}}(c),$$

where $\sum_i \lambda_i = 1$, with each component normalized to $[0, 1]$. This score guides both report presentation and output control: high-confidence candidates become primary interpretations, while plausible lower-confidence ones are retained as alternatives with explicit uncertainty markers, avoiding the generation of overconfident single conclusions.

Table 2: Interpretation Stage (I): Three Coordinated Branches

Branch	Evidence Inputs	Semiotic Target & Reasoning	KG Validation Logic	L3 Output Slots
I_t (Theme)	F_D, A	Focus-set-grounded combinatorial reasoning with formal cues as supplementary evidence, triggered by $\mathcal{R}$	Co-occurrence reward; counter-evidence penalty	Subject, Motif
I_m (Mood)	F_D, A	Figure emotion, object imagery, scene perception, and formal features; combinatorial reasoning triggered by $\mathcal{R}$	Co-occurrence consistency check	Overall Atmosphere, Figure Emotion, Thermal Perception
I_v (Value)	F_D, A, I_t, I_m	Integration of I_t and I_m outputs weighted by $\mathcal{R}$-triggered activation over F_D elements; candidate arbitration by aggregate score	Genealogical validation against dynastic context; anachronism suppression	Value Tendency, Expressive Stance, Philosophical Orientation

4.6 Evaluation (E)

The E stage situates the artwork within broader historical and cultural coordinates rather than merely recapping prior findings. Formally:

$$E = f_e\big(\Phi_{\mathcal{G}}(I_t,\ I_m,\ I_v,\ D,\ A);\ \mathcal{O}\big) \tag{6}$$

$\Phi_{\mathcal{G}}$ performs cross-artwork retrieval conditioned on I-stage outputs, recovering comparable works, transmission lineages, and biographical/textual nodes that jointly define the historical context for X.

Meanwhile, confidence scores from the preceding D, A, and I stages are passed to the E stage and dynamically calibrated against the cross-artwork contextual evidence retrieved by $\Phi_{\mathcal{G}}$. The structured output $E = (S_f,\ P_v,\ C_p)$ covers three evaluative dimensions: S_f (*Social Function*) assesses the ritual or diplomatic role the work played in its original social network; P_v (*Philosophical/Value Positioning*) assesses its continuity, variation, or subversion of canonical thought; and C_p (*Creative Positioning*) locates the work within a stylistic lineage or the artist's oeuvre. This design enforces scholarly comparability and prevents the impressionistic praise typical of unconstrained MLLM outputs, which often lacks explicit ties to historical evidence or artistic genealogy.

5 Experiments

We evaluate TCFP-Sight along four research questions. **RQ1 (Overall Quality)** : Does TCFP-Sight surpass general-purpose MLLMs and RAG baselines in cultural-iconographic interpretation (CI) and argument transparency (AT)?**RQ2 (Component Necessity)** : Does each component—ontology gating, DAIE staging, KG verification, symbolic rules—yield measurable gains? **RQ3 (Auditability Anti-hallucination)**: Does the framework quantifiably reduce cultural-interpretation hallucinations and improve evidence traceability? **RQ4 (Narrative Coherence)** : How does the system handle multi-scene handscrolls requiring cross-segment identity consistency?

5.1 Experimental Setup

Benchmark. We compile a benchmark of 100 Chinese figure paintings spanning five dynasties (Tang, Song, Yuan, Ming, Qing) and six subject categories (Elegant Gatherings, Court Lady, Religion and Mythology, Historical Narratives and Genre Scenes), balanced across scrolls, handscrolls, and album leaves. Each painting is accompanied by its image, metadata (title, artist, dynasty, collection), and reference texts drawn from museum catalogues and art-historical scholarship. To test performance both within and beyond knowledge-graph coverage, the benchmark comprises 50 In-KB and 50 Out-of-KB paintings.

Among these, 35 artworks form a fine-grained **Gold subset** via stratified sampling. Each Gold report contains expert-annotated D-stage entity-and-relation lists (L_1), A-stage formal-feature tags (L_2), I-stage theme/mood/value candidates with explicit evidence links to D and A entries (L_3), and E-stage contextual evaluations. Two art-history graduate students perform initial annotation independently; a senior specialist adjudicates disagreements. Inter-annotator agreement before adjudication is $\kappa = 0.79$ for L_1 entity/relation labels, $\kappa = 0.76$ for L_2 formal features, $\alpha = 0.74$ for L_3 theme/value ordinal ratings, and $\kappa = 0.82$ for hallucination four-category classification. Additionally, 8 handscroll artworks (each segmented into 5–8 scene patches, totalling ~48 patches) constitute a separate narrative subset for RQ4.

Baselines. All systems use GPT 5.4 as the backbone LLM, with temperature $T = 0.3$, maximum output length of 4,096 tokens, and retrieval depth $k = 8$ for retrieval-augmented systems. To isolate TCFP-Sight's architectural contribution from backbone capacity and knowledge content, we enforce a same-backbone, vary-mechanism principle across four baselines. **B1 (MLLM-Direct)** uses no external knowledge and generates outputs in one shot, serving as the lower-bound reference. **B2 (MLLM-Structured)** likewise uses no external knowledge, but structured prompts explicitly elicit staged D→A→I→E output with visual-evidence citation, directly testing whether prompt engineering alone can substitute for architecture-level constraints. **B3 (Text-RAG)** accesses the same corpus as TCFP-Sight via flat document retrieval and generates in one shot after retrieval. **B4 (ArtRAG [30], adapted)** retains the ArtRAG multi-granularity subgraph pipeline but replaces its original knowledge graph with the identical Chinese figure painting corpus. TCFP-Sight alone operates under ontology-gated DAIE reasoning supported by a structured prior knowledge space $\mathcal{K} = (\mathcal{O}, \mathcal{G}, \mathcal{R})$.

Evaluation Protocol. Evaluation is organised into three complementary tiers.

Tier 1 — LLM-as-a-Judge (automated, all 100 paintings). We employ GPT-5.4 as the judge model with dimension-specific rubrics, scoring each system output independently on a 1–5 Likert scale across four dimensions: **Factual Grounding (FG)** measures the accuracy of L_1 entity/relation and L_2 formal descriptions; **Formal Analysis Depth (FA)** measures the richness of compositional and brushwork analysis; **Cultural-Iconological Interpretation (CI)** measures the depth and cultural accuracy of L_3 theme, mood, and value readings; **Argument Transparency (AT)** measures whether

Table 3: Main results. FG/FA/CI/AT are LLM-as-a-Judge scores (1–5) on all 100 paintings; ESR, GCR, and Hallucination Total are on the 35-painting Gold subset. Bold indicates best per column; underline indicates second best. TCFP-Sight is further split into in-KG and out-of-KG subsets.

System	Subset	FG↑	FA↑	CI↑	AT↑	ESR↑	GCR↑	Hall. Total↓
MLLM-Direct (B1)	All	2.59	3.07	3.38	2.33	0.737	0.524	350
MLLM-Structured (B2)	All	**3.50**	**3.43**	3.50	3.23	0.881	0.743	428
Text-RAG (B3)	All	2.79	3.24	3.49	2.54	0.703	0.548	370
ArtRAG [adapted] (B4)	All	2.54	3.04	3.52	2.29	0.645	0.567	381
TCFP-Sight	All	3.23	3.16	**3.91**	**3.39**	**0.914**	**0.916**	**122**
	in-KG	3.18	3.28	3.98	3.40	0.912	0.921	51
	out-of-KG	3.28	3.04	3.84	3.39	0.917	0.911	71

Table 4: Hallucination classification counts on the 35-painting Gold subset (instance counts; lower is better).

System	OH↓	FM↓	IH↓	CM↓	Total↓
MLLM-Direct (B1)	273	32	22	23	350
MLLM-Structured (B2)	247	42	70	69	428
Text-RAG (B3)	286	29	27	28	370
ArtRAG [adapted] (B4)	285	35	37	24	381
TCFP-Sight	**84**	**20**	**13**	**5**	**122**

semantic claims explicitly cite low-level evidence. The judge is blind to system identity and output order.

Tier 2 — Auditability and hallucination metrics (semi-automatic, Gold subset). Metrics include: **Evidence Support Rate (ESR)** — proportion of I/E-stage claims with explicit L_1/L_2 evidence links; **Cross-stage Violation Rate (CVR)** — proportion of D/A output segments containing L_3 concepts; **Gating Compliance Rate (GCR)** — proportion of output tokens at each DAIE stage belonging to that stage's ontological field set; and **Hallucination Count** decomposed into four types: **OH** (object hallucination), **FM** (formal misattribution), **IH** (iconographic hallucination), and **CM** (contextual mismatch). Two annotators independently label hallucination instances; disagreements are adjudicated by a senior domain expert.

Tier 3 — Expert human evaluation (30 paintings). Five evaluators with Chinese art-history backgrounds conduct blind assessments. Each painting is presented with outputs from TCFP-Sight and the two strongest automatic-metric baselines (order randomised, identity anonymised). Evaluators provide 5-point Likert ratings on the four dimensions and pairwise preference judgements (win/tie/loss). We report mean ± SD, Krippendorff's α, and Wilcoxon signed-rank p-values.

5.2 Main Results

Overall Performance. TCFP-Sight achieves the best results on the five metrics most central to this work. The CI gain over the strongest baseline (B4 = 3.52) is 0.39 points (+11.1%), a margin approximately 2.8× the entire inter-baseline CI range, indicating a qualitative leap in cultural-semantic reasoning that no baseline approaches. GCR exceeds the strongest baseline (B2 = 0.743) by 23.3 % points, confirming that architectural gating enforces stage-level vocabulary compliance beyond what structured prompting achieves. However, TCFP-Sight does not lead on FG or FA , falling below B2's 3.50 and 3.43 respectively—a pattern explained by the hallucination analysis below.

Structured Prompting and the Hallucination Cost. B2 achieves the highest FG and FA, confirming that structured prompts improve surface-level textual quality. Yet B2 simultaneously produces the highest hallucination total (428)—a 22.3% increase over zero-prompt B1 (350). Table 4 reveals the mechanism: while B2 reduces OH (247 vs. B1's 273), its IH (70) and CM (69) surge to 3.2× and 3.0× B1's counts. Structured prompts compel richer content at each stage, but without knowledge-driven constraints the additional output is largely filled with fabricated cultural associations—prompt engineering alone elevates surface fluency rather than cognitive reliability. That TCFP-Sight underperforms B2 on FG/FA is precisely because its controlled vocabulary trades a degree of narrative richness for higher reliability—a trade-off that yields substantial returns on AT, ESR, and hallucination suppression. Importantly, this reduction is not an artefact of generating less content: on the Gold subset, TCFP-Sight produces 459 I/E-stage interpretive claims versus B2's 403, yet its high-level hallucination density ($\frac{\text{IH+CM}}{\text{I/E claims}}$) is 0.039 compared with B2's 0.345—nearly an order of magnitude lower. This finding empirically validates the core argument of Section 1: architectural constraints on the reasoning paradigm are irreplaceable.

Hallucination Category Analysis. TCFP-Sight achieves the lowest count across all four categories. The most pronounced reduction occurs in CM—from 23 (B1) to 5 (−78.3%)—directly validating $\Phi_{\mathcal{G}}$'s effectiveness in dynastic context checking. IH and CM, the two most characteristic failure modes of cultural interpretation, are jointly suppressed from 45 (lowest baseline combined count) to 18 (−60.0%), constituting the primary empirical support for the framework's anti-hallucination capability in high-level cultural reasoning.

In-KG vs. Out-of-KG Generalisation. The two subsets show negligible divergence on AT (3.40 vs. 3.39), ESR (0.912 vs. 0.917), and GCR (0.921 vs. 0.911), indicating that the evidence-linking and gating mechanisms derive their benefit primarily from architectural constraints rather than KG retrieval hits. The in-KG subset shows a modest advantage on FA (3.28 vs. 3.04) and CI (3.98 vs. 3.84); the out-of-KG subset produces marginally more hallucinations (71 vs. 51) but remains substantially below any baseline. TCFP-Sight's core advantages—high auditability and low hallucination—are preserved beyond KG coverage, demonstrating that its gains do not arise from simple memorisation of in-graph instances.

Table 5: Ablation results on the Gold subset. IH/CM are per-painting average counts; Inst. is average total hallucination count per painting; HL rate = (IH + CM)/Gold-annotated I/E interpretive units per painting.

Variant	FG↑	CI↑	AT↑	ESR↑	CVR↓	GCR↑	IH↓	CM↓	Inst.↓	HL rate↓
Full	**3.06**	**3.91**	**3.40**	**0.912**	0.084	**0.916**	**0.37**	0.14	**3.49**	**0.222**
w/o Ontology Gating	2.63	3.57	2.00	0.783	0.147	0.723	0.67	0.30	4.86	0.422
w/o Knowledge Graph	2.71	3.60	3.31	0.905	0.088	0.908	0.52	0.26	4.62	0.339
w/o Symbolic Rules	2.43	3.51	3.11	0.755	**0.064**	0.891	0.48	**0.10**	4.73	0.252
w/o Staged DAIE	2.74	3.80	3.26	0.314	0.160	0.658	0.88	0.63	4.89	0.658

5.3 Ablation Study

We evaluate four ablation variants on the 35-painting Gold subset, each removing one core mechanism.

Staged DAIE Is the Foundation of System Integrity. Collapsing the pipeline into single-pass generation while retaining $\mathcal{K}$ produces the most severe degradation: ESR drops to 0.314 ($\Delta = -0.598$), CVR rises to its highest value (0.160), GCR drops to 0.658, and both IH and CM peak with a high-level hallucination rate of 0.658. Notably, CI declines by only 0.11, showing the model can still produce plausible-seeming interpretive text even as its evidential anchoring has collapsed—confirming that domain knowledge alone is insufficient without staged reasoning.

Ontology Gating Is the Primary Safeguard of Auditability. Removing gating causes AT to drop from 3.40 to 2.00 ($\Delta = -1.40$)—the largest single-metric decline across all variants; GCR drops from 0.916 to 0.723, confirming that field-level enforcement is the direct mechanism behind token-level stage compliance; CVR rises by 75.0%; and the high-level hallucination rate nearly doubles to 0.422. Gating prevents premature activation of high-level judgements by strictly restricting each stage's semantic access.

The Knowledge Graph Primarily Serves Factual Verification. Removing $\Phi_{\mathcal{G}}$ yields only marginal ESR decline and GCR remains largely unchanged, indicating that evidence-linking structure and vocabulary compliance are maintained by the architecture independently of KG availability. However, total hallucination instances rise by 32.4% and the CM rate increases ×1.9, confirming the KG's core role in suppressing dynastic context misattribution—functioning as a factual verification layer rather than an evidence organisation layer.

Symbolic Rules Activate Controlled Interpretation. Removing $\Psi_{\mathcal{R}}$ yields the lowest FG and reduces ESR to 0.755, yet simultaneously records the lowest CVR and CM. This arises because without rule-triggered activation the system generates fewer high-level claims, reducing both violations and interpretive capacity . The rules' primary function is to provide stable, controlled candidate activation pathways for high-level interpretation.

Complementarity. The full model achieves the optimal balance across interpretation quality, auditability, and hallucination control. Each component fulfils a distinct structural role; no component is redundant—the four are mutually complementary rather than substitutable—validating the necessity of co-designing $\mathcal{K} = (\mathcal{O}, \mathcal{G}, \mathcal{R})$ in concert with the reasoning architecture.

5.4 Expert Evaluation and Case Studies

Expert Likert Scores. To complement the limitations of automatic metrics in capturing aesthetic judgement and cultural normativity, five evaluators with Chinese art-history backgrounds (three doctoral candidates and two senior master's students) conducted blind assessments on 30 paintings. Each painting was presented with outputs from TCFP-Sight and the two strongest baselines identified in Section 5.2 (B2 and B4), in randomised order with system identities anonymised. Evaluators rated each output on four dimensions using a 5-point Likert scale and provided pairwise preference judgements (win/tie/loss).

Table 6: Expert evaluation. 5-point Likert scores (mean ± SD) and pairwise preferences of TCFP-Sight vs. the strongest baseline per dimension. α: Krippendorff's alpha; *: $p < 0.05$, ***: $p < 0.001$ (Wilcoxon signed-rank test).

Dimension	Strongest Baseline	TCFP-Sight	W / T / L	Sig.
Factual Grounding (FG)	3.76 ± 1.11	3.28 ± 0.96	7/8/15	$p = 0.063$
Formal Analysis (FA)	3.26 ± 1.15	3.87 ± 0.87	18/5/7	$p = 0.011^{*}$
Cultural Interp. (CI)	3.24 ± 0.92	4.21 ± 0.87	23/4/3	$p < 0.001^{***}$
Argument Transp. (AT)	3.08 ± 1.05	4.07 ± 0.79	22/4/4	$p < 0.001^{***}$
Inter-rater agreement (α)		0.73		

Expert results closely mirror the automatic evaluation trends. TCFP-Sight is significantly preferred on CI and AT —the two dimensions most central to this work's contribution. FA also shows a significant advantage (+0.61, $p = 0.011$), suggesting that domain experts value the layered, structured analytical approach more than what surface-level automated scoring captures. FG is marginally lower (−0.48, $p = 0.063$, not significant), consistent with the controlled-vocabulary trade-off identified in Section 5.2. The convergence between Tier-1 automated and Tier-3 expert evaluations strengthens the robustness of the overall evaluation conclusions. Inter-rater agreement (Krippendorff's $\alpha = 0.73$) indicates substantial agreement among evaluators.

6 Conclusion

TCFP-Sight reframes Traditional Chinese figure painting appreciation as a structured, evidence-bound cognitive reasoning process by grounding a four-stage DAIE pipeline in the prior knowledge space $\mathcal{K}$ (ontology, knowledge graph, and symbolic rules) and enforcing hard stage-layer boundaries through Ontology Gating. This architecture systematically suppresses cultural hallucination while guaranteeing full claim traceability—ablations confirm that Ontology Gating is the single most critical component, and that staged reasoning cannot be replaced by domain knowledge alone. Future work will extend the framework to broader painting genres and develop human–AI collaborative workflows for curatorial and educational practice.

References

[1] Andrea Alfarano, Lorenzo Venturoli, and Darío Negueruela del Castillo. 2025. VQArt-Bench: A semantically rich VQA Benchmark for Art and Cultural Heritage. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. IEEE, 406–416.

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization. *Text Reading, and Beyond* 2, 1 (2023), 1.

[3] Dina Zoe Belluigi. 2017. A framework to map approaches to interpretation. *Journal of Aesthetic Education* 51, 3 (2017), 91–110.

[4] 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 *Proceedings of the 32nd ACM International Conference on Multimedia*. 7734–7743.

[5] Xiaojiao Chen, Yueying Li, Yonghao Chen, Tan Tang, Ruihan Wang, Yifan Wang, Yingchaojie Feng, Wei Chen, and Xiaosong Wang. 2025. ACPAS: an expert-assistance system for authenticating ancient Chinese paintings via LLM-based agents. *npj Heritage Science* 13, 1 (2025), 512.

[6] CRM CIDOC. 2003. Conceptual Reference Model. *URL: https://cidoc-crm. org* (2003).

[7] Leendert D Couprie. 1983. Iconclass: an iconographic classification system. *Art libraries journal* 8, 2 (1983), 32–49.

[8] Martin Doerr. 2003. The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI magazine* 24, 3 (2003), 75–75.

[9] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130* (2024).

[10] Tao Fan, Hao Wang, Yuehua Zhao, Shu Zhou, and Bin Shi. 2026. 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* 63, 3 (2026), 104481.

[11] Achille Felicetti, Aida Himmiche, and Miriana Somenzi. 2025. Knowledge Graphs and Artificial Intelligence for the Implementation of Cognitive Heritage Digital Twins. *Applied Sciences* 15, 18 (2025), 10061.

[12] Wen Fong. 1986. James Cahill," The Compelling Image: Nature and Style in Seventeenth-Century Chinese Painting"(Book Review). *The Art Bulletin* 68, 3 (1986), 504.

[13] Xiaoyan Hu. 2016. The Notion of 'Qi Yun' (Spirit Consonance) in Chinese Painting. *Proceedings of the European Society for Aesthetics* 8 (2016), 247–268.

[14] Leonardo Impett and Franco Moretti. 2017. Totentanz. operationalizing aby warburg's pathosformeln. (2017).

[15] Ruixiang Jiang and Chang Wen Chen. 2025. Multimodal llms can reason about aesthetics in zero-shot. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 6634–6643.

[16] Afshin Khadangi, Amir Sartipi, Igor Tchappi, and Gilbert Fridgen. 2025. CognArtive: large language models for automating art analysis and decoding aesthetic elements. *arXiv preprint arXiv:2502.04353* (2025).

[17] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.

[18] Meng Li, Yun Wang, and Ying-Qing Xu. 2022. Computing for Chinese cultural heritage. *Visual Informatics* 6, 1 (2022), 1–13.

[19] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems* 36 (2023), 34892–34916.

[20] Viktor Matanski, Anton Iliev, Nikolay Kyurkchiev, and Todorka Terzieva. 2025. Ontology-Driven Multi-Agent System for Cross-Domain Art Translation. *Future Internet* 17, 11 (2025), 517.

[21] Shravan Nayak, Kanishk sartini2023iconl Jain, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stańczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 5769–5790.

[22] Erwin Panofsky. 1955. *Meaning in the Visual Arts: Papers in and on Art History*. University of Chicago Press.

[23] Erwin Panofsky. 2018. *Studies in iconology: Humanistic themes in the art of the renaissance*. Routledge.

[24] Bruno Sartini, Sofia Baroncini, Marieke Van Erp, Francesca Tomasi, and Aldo Gangemi. 2023. Icon: An ontology for comprehensive artistic interpretations. *ACM Journal on Computing and Cultural Heritage* 16, 3 (2023), 1–38.

[25] Ralph A Smith. 1968. Aesthetic criticism: The method of aesthetic education. *Studies in Art Education* 9, 3 (1968), 12–31.

[26] Maithreyi Subramaniam, Jaffri Hanafi, and Abu Talib Putih. 2016. Teaching for Art Criticism: Incorporating Feldman's Critical Analysis Learning Model in Students' Studio Practice. *Malaysian Online Journal of Educational Technology* 4, 1 (2016), 57–67.

[27] Nanne Van Noord. 2022. A survey of computational methods for iconic image analysis. *Digital Scholarship in the Humanities* 37, 4 (2022), 1316–1338.

[28] Jing Wan, Hao Zhang, Jun Zou, Ao Zou, Yubin Chen, Qingyang Zeng, Xinrong Li, and Qiya Wang. 2024. WuMKG: a Chinese painting and calligraphy multimodal knowledge graph. *Heritage Science* 12, 1 (2024), 1–18.

[29] Runzhou Wang, Xinsheng Zhang, Qilei Liu, Jinqi Su, Yuezhou Zhang, and Yulong Ma. 2026. Visual information identification and Q&A of intangible cultural heritage inheritors by using enhanced Graph-Retrieval framework. *npj Heritage Science* 14, 1 (2026), 113.

[30] Shuai Wang, Ivona Najdenkoska, Hongyi Zhu, Stevan Rudinac, Monika Kackovic, Nachoem Wijnberg, and Marcel Worring. 2025. ArtRAG: Retrieval-Augmented Generation with Structured Context for Visual Art Understanding. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 6700–6709.

[31] Yue Wang and Man Zhang. 2025. CIDOC CRM-Based Knowledge Graph Construction for Cultural Heritage Using Large Language Models. *Applied Sciences* 15, 22 (2025), 12063.

[32] Liang Xu, Lu Lu, Minglu Liu, Chengxuan Song, and Lizhen Wu. 2024. Nanjing Yunjin intelligent question-answering system based on knowledge graphs and retrieval augmented generation technology. *Heritage Science* 12, 1 (2024), 1–23.

[33] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review* 11, 12 (2024), nwae403.

[34] Haorui Yu, Ramon Ruiz-Dolz, and Qiufeng Yi. 2025. A structured framework for evaluating and enhancing interpretive capabilities of multimodal LLMs in culturally situated tasks. *Findings of the Association for Computational Linguistics: EMNLP* 2025 (2025), 1945–1971.

[35] Zhengqing Yuan, Yunhong He, Kun Wang, Yanfang Ye, and Lichao Sun. 2023. ArtGPT-4: Towards artistic-understanding large vision-language models with enhanced adapter. *arXiv preprint arXiv:2305.07490* (2023).

[36] Chenhao Zhang, Xi Feng, Yuelin Bai, Xeron Du, Jinchang Hou, Kaixin Deng, Guangzeng Han, Qinrui Li, Bingli Wang, Jiaheng Liu, et al. 2025. Can MLLMs Understand the Deep Implication Behind Chinese Images?. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 14369–14402.

[37] Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024. Mm-llms: Recent advances in multimodal large language models. *Findings of the Association for Computational Linguistics: ACL 2024* (2024), 12401–12430.

[38] Tuo Zhang, Tiantian Feng, Yibin Ni, Mengqin Cao, Ruying Liu, Kiana Avestimehr, Katharine Butler, Yanjun Weng, Mi Zhang, Shrikanth Narayanan, et al. 2025. Creating a lens of chinese culture: A multimodal dataset for chinese pun rebus art understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*. 22473–22487.

[39] Wei Zhang, Wong Kam-Kwai, Biying Xu, Yiwen Ren, Yuhuai Li, Yingchaojie Feng, Minfeng Zhu, and Wei Chen. 2025. Cultiverse: Towards cross-cultural understanding for paintings with large language model. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 6710–6719.

[40] Wei Zhang, Jian-Wei Zhang, Kam-Kwai Wong, Yi-Fang Wang, Ying-Chao-Jie Feng, Lu-Wei Wang, and Wei Chen. 2024. Computational approaches for traditional chinese painting: From the "six principles of painting" perspective. *Journal of Computer Science and Technology* 39, 2 (2024), 269–285.

[41] Xiangrong Zhu, Yuexiang Xie, Yi Liu, Yaliang Li, and Wei Hu. 2025. Knowledge graph-guided retrieval augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 8912–8924.