Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Visually similar prints expose a benchmark validity gap in computational art history

Jiongming Fang

20252011@cuz.edu.cn

Communication University of Zhejiang

Article

Keywords: digital humanities, computational art history, cultural heritage, image retrieval, benchmark validity, print history, provenance

Posted Date: September 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10759054/v1

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

Additional Declarations: No competing interests reported.

Abstract

Digitised cultural collections increasingly use visual retrieval to connect related images, yet visual resemblance does not by itself establish a historical relation between physical objects. This distinction is acute for prints: two sheets may preserve the same composition while deriving from different blocks, states, workshops or copying events. The study tests whether a retrieval benchmark can appear successful because it omits such historically consequential alternatives. A diagnostic repeated-design task is compared with a controlled candidate-pool intervention built from 39 Metropolitan Museum of Art woodcut records. Across 240 transformed queries, Pool A contains an intended same-design target and other-design negatives; Pool C changes nothing except adding one catalogue record explicitly marked as a same-design copy. Raw pixels, histograms of oriented gradients and ORB rank the intended target first for all Pool A queries, but only 187, 177 and 237 times in Pool C. The added copy becomes the unique top result 53, 63 and 3 times. An abstaining ORB/HOG gate also accepts a copy for two queries. These are catalogue-class ranking errors, not findings about matrix identity or authenticity. The results show that candidate composition is part of construct validity: a benchmark of design recurrence cannot support a claim about historical sameness unless relation-confusable candidates are represented. A relation-aware evidence architecture is proposed that separates visual discovery, confusability testing and source-bearing historical judgement. The approach gives digital humanities and cultural-heritage projects a reproducible way to test when useful similarity becomes an invalid historical inference.

Introduction

Digitised museum collections make it possible to search across more images than a scholar could inspect one by one. Visual retrieval can recover recurring compositions, reconnect dispersed objects and support the study of artistic transmission (Arnold and Tilton 2019; Lang and Ommer 2018; van Noord 2022). Its usefulness is especially evident in print culture, where a design may travel through impressions, editions, recut blocks, close copies and reproductions. The same circulation also creates a methodological problem. A system can return an image that looks strikingly right while giving no reliable answer to the historical question that motivated the search.

The problem is not simply that an algorithm sometimes makes a mistake. It is that a benchmark may define the wrong kind of success. If every negative image depicts a different composition, a model can score perfectly by recognising design recurrence. A print historian, however, may need to distinguish two objects that retain the same composition but have different relations to a matrix, state, workshop or copying event. Unless those alternatives enter the candidate set, the evaluation cannot show whether the model preserves the distinction. A strong score may then be valid for visual discovery and invalid for a broader claim about historical sameness.

This gap matters beyond a specialist technical debate. Digital collections are knowledge infrastructures: their representations, metadata and interfaces shape what can be found, compared and reused (Borgman et al. 2019). Museum digitisation has often prioritised access to images while historically

important metadata remains uneven or difficult to integrate (Noehrer et al. 2021). Algorithmic processing adds another representational layer. As Blanke (2024) argues for digital archives more broadly, computational operations do not merely access an archive; they reassemble it around particular questions and associations. A retrieval result is therefore produced jointly by the query image, the model, the candidate set and the catalogue records through which the returned objects are interpreted.

Digital art history treats models, corpora and categories as participants in knowledge production rather than neutral tools applied after interpretation (Impett and Offert 2022; Cha 2026). Computational work in the humanities therefore increasingly emphasises validation, transparency and expert oversight rather than treating benchmark performance as self-explanatory (Hindermann et al. 2026; Karjus 2025). Recent human-centred research on automated image cataloguing likewise shows that plausible machine output still requires specialised review and traceable quality control (Abele et al. 2026). The present study brings that concern to visual retrieval. Its focus is not whether a particular descriptor is generally good or bad, but whether an evaluation contains the alternatives needed to warrant the historical interpretation attached to its score.

Prints provide a demanding case because visual appearance and material history are related but non-equivalent. One impression may come from a particular block; another may be pulled from the same block after wear or alteration; a close copy may be cut by another hand; and a digital photograph may introduce further changes in colour, crop, scale and contrast. Material contact is central to the ontology of printmaking (Abell 2015; Cook 2023; Roberts 2024), while the digital surrogate primarily records visible appearance. Conservation studies demonstrate that sequence and production history may depend on paper, pigments, damage patterns and multivariate material evidence unavailable to whole-image similarity alone (Korenberg et al. 2019; Vermeulen et al. 2020).

The study introduces the term *relation-confusable candidate* for a record that is visually close enough to compete in retrieval but belongs to a different operational relation class for the task under evaluation. The empirical intervention is deliberately narrow. In a Metropolitan Museum of Art (Met) dataset, a record is assigned to an *explicit-copy class* only when its displayed title contains "copy" and its artist prefix is "After". A record without those markers is assigned to the *target class*. These are unary catalogue-derived labels, not pairwise expert judgements about shared matrix, state, originality, copying direction, attribution or authenticity.

The central comparison changes only the candidate pool. Pool A contains an intended same-design target-class record and records of other designs, while excluding the same-design explicit copy. Pool C contains the same query, transformations, scores and Pool A candidates, plus that one copy. Pool B contains only the intended target and copy. Three research questions follow:

RQ1

How does intended-target ranking change when one same-design explicit-copy candidate is added to an otherwise unchanged pool?

RQ2

When the intended target and copy are compared directly, how often does each method rank the intended target first?

RQ3

Does an archived ORB/HOG decision gate that can abstain prevent copy-class records from entering its accepted set?

The article makes three contributions. First, it frames candidate composition as a construct-validity problem for digital humanities evaluation. Second, it supplies a controlled, query-level test of that problem, including observable accepted-copy errors rather than a hypothetical warning. Third, it proposes a relation-aware evidence architecture that records visual discovery, adversarially relevant alternatives and the sources needed for historical judgement as separate but connected objects. This architecture does not automate connoisseurship. It defines what a computational system may responsibly claim and what must remain available for human, documentary and material investigation.

Conceptual framework

From cultural object to digital surrogate

A visual retrieval system does not compare physical prints directly. It compares digital surrogates produced through photography, scanning, colour management, cropping, compression and interface delivery. Let *O* denote a physical cultural object, *S(O)* its digital surrogate and *A(O)* the catalogue assertions currently attached to it. A retrieval function ranks relations among surrogates, such as the similarity between *S(O1)* and *S(O2)*. A historical claim, by contrast, concerns a relation between physical objects, such as whether *O1* and *O2* derive from the same matrix or whether one is a copy after the other. Catalogue assertions may support, qualify or dispute that claim, but they are not created by the similarity score.

This three-part distinction explains why a visually correct result can still be historically underdetermined. The surrogate may preserve composition well enough for discovery while suppressing evidence from paper, watermark, plate or block wear, printing pressure, ink, colourant, verso marks or object history. The catalogue may also be incomplete, institution-specific or revised over time. Drucker's (2011) distinction between humanities data as given and capta as constructed is useful here: the image, metadata fields and benchmark labels are not neutral fragments of the object but selections made for particular purposes.

Figure 1 turns this distinction into an inference boundary. Visual discovery asks which surrogates resemble one another. Relation-confusable evaluation asks whether the ranking survives the presence of an alternative that is visually close but operationally wrong. Historical judgement asks what relation can

be supported by catalogue, provenance, material or specialist evidence. Evidence can move across the stages, but a result from an earlier stage cannot silently inherit the authority of a later one.

Candidate composition and construct validity

Construct validity asks whether an operational measure supports the interpretation assigned to it (Cronbach and Meehl 1955). In information retrieval, a test collection operationalises a task through queries, documents and relevance judgements (Voorhees 2002; Sanderson 2010). Measurement studies further show that a stable or accurate metric can still fail to represent the intended concept (Jacobs and Wallach 2021). Taken together, these traditions imply that benchmark validity depends not only on labels and metrics but also on the alternatives the system is allowed to encounter.

The distinction is important because "same image" can refer to several different constructs: the same file, the same photographed sheet, another impression from one matrix, another state, a copy after the same design or a broadly similar composition. A benchmark containing only other-design negatives operationalises repeated-design discovery. That is a legitimate and often useful task. The validity gap appears only when success on that task is reported as evidence for a relation-sensitive claim that the candidate set never tested.

A relation-confusable candidate has three properties. It is close enough in the model's input space to compete with the intended result; it belongs to the same design family under the working grouping; and it crosses the operational relation boundary relevant to the evaluation. The definition is task-relative. A later state may be correct for a design-family search and wrong for a state-identification task. A photographic reproduction may be useful for iconographic discovery and inappropriate as evidence for object identity. Relation-confusable evaluation therefore begins by stating the relation of interest rather than by choosing a model.

Candidate intrusion provides a direct test. Suppose a method selects the intended target in Pool A. Pool C then adds one relation-confusable record without changing the query or ordinary negatives. If the added record becomes the unique top result, the event is a *strict intrusion*. If it creates a top-score tie, the tie is recorded separately because it exposes unresolved evidence rather than a defensible automatic choice. This intervention isolates a property that a conventional aggregate design-recall score hides: whether success depends on removing a historically consequential alternative before evaluation begins.

Why print history is a transferable stress test

Print history makes the problem visible because its objects are relational by manufacture. Terms such as impression, state, edition, recut and copy connect a sheet to matrices, interventions, sequences and acts of transmission. Major contours and iconographic arrangement may remain stable across those relations, while the decisive evidence lies in carved line, hatching, wear, damage, paper or documentation. Computer vision has demonstrated genuine value in detecting image reuse and

organising art collections (Chung et al. 2015; Kräutli et al. 2021). The present argument does not diminish that achievement. It specifies where recognition of reuse ends and identification of a material-historical relation begins.

The framework also transfers to other cultural domains. Photographs printed from different negatives, casts made from different moulds, manuscript exemplars copied from a shared source, replicas of archaeological objects and later performances of a recorded work can all be visually close while historically distinct. The relevant relation and evidence will change by domain, but the evaluation principle remains: the candidate set should include plausible alternatives that cross the distinction the research claims to preserve.

Materials and methods

Study design and research questions

The research uses a within-query candidate-pool intervention. For pairwise scoring methods, the query, intended target, other-design negatives, image preprocessing and all pairwise scores are held constant between Pool A and Pool C. The sole change is the addition of one same-design explicit-copy record. The design therefore asks how a ranking changes when a historically relevant alternative becomes available, not whether a new model improves a conventional benchmark.

RQ1 is answered by paired Pool A and Pool C outcomes for 240 composite queries. RQ2 is answered by Pool B, which contains only the intended target and explicit copy. RQ3 is answered from the accepted and reviewed decisions of a frozen ORB/HOG gate. A separate Art Institute of Chicago (AIC) corpus supplies an easier repeated-design diagnostic, showing what the methods can achieve when negatives depict other compositions.

Corpora, working groups and operational labels

The primary corpus contains 39 public-domain Met woodcut records organised into 12 checked working same-design groups. Two groups were used during development and ten form the challenge set. Each challenge design contains two target-class records and one explicit-copy record. The 20 target-class records serve as query parents. For each query parent, the identical source object is excluded from retrieval, and the other target-class record in the same working design group is designated as the intended target.

Metadata were obtained from the Met Collection API and frozen in the accompanying manifest (Metropolitan Museum of Art 2026a). A record enters the explicit-copy class when the displayed title contains the string "copy" and the displayed artist prefix is "After". A record enters the target class when the displayed artist contains Dürer, the title lacks "copy" and the prefix is not "After". On 17 August 2026, all 39 object endpoints were requested again and checked for identifier, title, artist fields, medium, object

URL, public-domain status and derived class. Images were also rechecked at the composition level. Met Open Access marks eligible images as CC0 (Metropolitan Museum of Art 2026b).

The labels establish an auditable catalogue-record task. They do not establish that target-class objects are originals, that two target-class objects share a matrix, or that the catalogue's copy description settles copying direction or attribution. The term *working same-design group* similarly denotes compositional grouping rather than matrix identity. Table 1 records these boundaries so that later results cannot be read more strongly than the data permit.

Table 1
Corpus, candidate-pool design and permitted inference

Component	Construction	Permitted inference
Met corpus	39 records in 12 working same-design groups; 2 development and 10 challenge designs; 20 query parents	Catalogue-record ranking within the declared operational classes
Pool A	28 candidates: intended same-design target plus 27 other-design records; same-design explicit copy excluded	Repeated-design retrieval without the difficult copy
Pool C	Pool A unchanged, plus one same-design explicit-copy record	Sensitivity to a relation-confusable candidate, including strict intrusions and ties
Pool B	Two candidates: intended target and same-design explicit copy	Target-over-copy ranking under the operational catalogue boundary
Label boundary	Unary Met classes derived from title and artist-prefix fields	Does not establish matrix identity, original status, copy direction, attribution or authenticity
AIC diagnostic	46 records, 20 working design groups and 2,208 deterministic queries; other-design negatives	Repeated-composition recovery, not relation-sensitive ranking

Source: created for this study from the frozen corpus and candidate manifests.

The AIC diagnostic corpus contains 46 public-domain records in 20 working design groups, obtained through the AIC API and Open Access service (Art Institute of Chicago 2026a, 2026b). It generates 2,208 deterministic transformed queries across four perturbation families; 552 composite queries are used for the repeated-design result reported here. Its negative candidates depict other compositions. It can therefore establish transformation-robust design recovery within that corpus, but it contains no equivalent controlled target-versus-copy intervention.

Query generation and candidate pools

Each of the 20 Met query parents supplies one clean query and 12 composite stress queries: four author-defined severity levels and three fixed-seed replicates at each level. The composite transformation simulates combined changes associated with contact, paper and digitisation after a common canonicalisation step. The analysed intervention uses 240 composite queries. Only query images are perturbed; gallery candidates remain canonical and unperturbed. The severity levels are diagnostic settings and are not estimates of how frequently comparable conditions occur in museum repositories.

Pool A has 28 candidates: the intended same-design target and 27 records from other challenge designs. It excludes the same-design explicit copy. Pool C retains all 28 candidates and adds that one copy. Pool B contains only the intended target and copy. Candidate identities and ordering are stored in a manifest. Pairwise score matrices are pool-invariant for raw pixels, perceptual hash (pHash), histograms of oriented gradients (HOG), ORB-RANSAC, InkTrace and DINOv2. Component-average rank fusion (CAR) is treated separately because it recomputes component ranks inside each pool and is therefore candidate-set dependent.

The primary outcome is intended-target recall at rank 1 (R@1): the designated target must be the unique top result. A copy, other design, top tie or ambiguity counts as zero. A strict copy intrusion requires a Pool A intended-target success followed by the newly added copy as the unique Pool C top result. Ties are detected with an absolute score tolerance of 1e-12 and reported separately. Pool B records target wins, copy wins and ties. Design R@1 is also calculated, but because both Pool B candidates share the working design it cannot measure the target-versus-copy distinction.

Visual methods and decision gate

Six classical outputs are analysed. Raw-pixel similarity uses mean-centred canonical images. pHash supplies a compact perceptual comparison. HOG represents edge orientation (Dalal and Triggs 2005). ORB matches binary local descriptors and applies RANSAC verification of a similarity transform (Fischler and Bolles 1981; Rublee et al. 2011). InkTrace combines background-relative darkness, HOG variants and projection profiles. CAR averages within-pool ranks from InkTrace and ORB.

DINOv2 and SuperPoint-LightGlue are included as exploratory modern baselines rather than confirmatory tests (DeTone et al. 2018; Lindenberger et al. 2023; Oquab et al. 2023). DINOv2 is evaluated on regenerated inputs in Pools A, B and C. SuperPoint-LightGlue is available only for Pool B and uses homography verification, so it cannot supply the controlled A-to-C comparison. Both were run after the challenge designs had been inspected and therefore do not constitute untouched external validation.

The frozen decision gate accepts the ORB top object only when ORB and HOG agree on that object; otherwise the query is routed to review. Its *coverage* is the proportion accepted. Its *selective risk* is the number of accepted copy-class selections divided by accepted decisions. Agreement between the two

descriptors can reduce some errors, but it is not independent documentary evidence: both are derived from the same digital image.

Descriptive analysis, reproducibility and AI use

The candidate intervention produces deterministic paired counts. No null-hypothesis tests or claims of population significance are made. To show dependence on the ten challenge designs, descriptive percentile intervals resample designs, not individual transformed queries, 10,000 times with bootstrap seed 20260801. These intervals are sensitivity summaries for this small constructed corpus; they are not population confidence claims. Query generation uses the archived seed specification rooted at 20260711. The supplementary package contains manifests, operational-label rules, parameter files, checksums, query-level outcomes, frozen and recomputed scores, environment specifications and executable analysis code. Dataset documentation follows the principle that evaluation data should expose their composition, intended use and limits (Gebru et al. 2021).

The surviving archive preserves the clean and composite decision matrices, candidate manifests and all reported outcomes. The original generated query bytes, feature correspondences and masks from the frozen run do not survive. Current execution reproduces the recorded decisions and generates deterministic replacement queries from the declared source, configuration and seed. Figure 3 therefore combines a frozen decision with a deterministic visual reconstruction; it is not presented as a pixel-identical reconstruction of the original query or as evidence of the exact feature correspondences that caused the decision.

OpenAI Codex was used under the author's direction during July–August 2026 solely for limited editorial assistance, including language polishing, stylistic refinement, text formatting, and document preparation. The author independently conceived and conducted the research, performed and verified the analyses and calculations, reviewed the literature and citations, interpreted the findings, and wrote and approved all substantive content. The author reviewed every AI-assisted edit and accepts full responsibility for the manuscript. OpenAI Codex is not an author. No generative AI was used to create or alter any figure imagery.

Results

Repeated-design retrieval is nearly perfect when relation-confusable candidates are absent

On the 552 composite AIC queries, HOG, ORB, InkTrace and CAR each achieve 1.000 design R@1. Raw pixels achieve 0.998 and pHash 0.984. This diagnostic shows that the transformations do not make repeated-design retrieval intrinsically impossible. It remains a design-level result because the negative candidates show other compositions.

The same pattern appears in Met Pool A. Raw pixels, HOG, ORB and CAR place the intended target first for all 240 composite queries; InkTrace succeeds for 237 and pHash for 236. If evaluation ended with Pool A, four methods would appear perfect for the implemented target-record task. The score would correctly describe ranking against other-design negatives, but it would not have encountered the same-design explicit copy.

One added copy changes the apparent result

Table 2 and Fig. 2 report the paired Pool A-to-C intervention. Raw-pixel intended-target counts fall from 240 to 187, HOG from 240 to 177 and ORB from 240 to 237. The added copy becomes the unique top result 53, 63 and 3 times, respectively. InkTrace produces 60 strict copy intrusions. pHash produces 28 strict intrusions and 27 additional transitions to a top tie. CAR records one strict intrusion but 61 ties; because CAR recomputes ranks inside the expanded pool, its large tie count is part of the method's candidate-set dependence rather than a pairwise-score change.

Table 2
Pool A-to-C intervention on 240 composite queries

Method	Pool A target R@1	Pool C target R@1	Strict copy intrusion	A-to-C tie
Raw	240/240 (1.000)	187/240 (0.779)	53/240 (0.221)	0
pHash	236/240 (0.983)	181/240 (0.754)	28/240 (0.117)	27
HOG	240/240 (1.000)	177/240 (0.738)	63/240 (0.263)	0
ORB	240/240 (1.000)	237/240 (0.988)	3/240 (0.013)	0
InkTrace	237/240 (0.988)	177/240 (0.738)	60/240 (0.250)	0
CAR*	240/240 (1.000)	178/240 (0.742)	1/240 (0.004)	61
DINOv2**	239/240 (0.996)	160/240 (0.667)	79/240 (0.329)	0

Source: created for this study. *CAR recomputes ranks inside each pool.* *DINOv2 uses regenerated inputs and is exploratory. Strict intrusion requires a Pool A intended-target success followed by the added copy as the unique Pool C top result.

The exploratory DINOv2 result displays the same structural vulnerability: intended-target R@1 falls from 239/240 in Pool A to 160/240 in Pool C, with 79 strict copy intrusions. This does not establish that learned features are generally inferior to classical features. The representation, preprocessing and evidential status differ, and the challenge designs were already known. It shows that a modern representation can also prefer a visually compelling same-design copy when the evaluation requires the target catalogue class.

The design-cluster distributions in Fig. 2 show that the effect is not evenly spread. Some designs are consistently easy and others account for a larger share of intrusions. This heterogeneity is one reason the transformed queries cannot be treated as 240 independent artworks. It also indicates where expert-

led error analysis is most useful: not at the average score alone, but at design families whose copied and target-class records occupy overlapping visual neighbourhoods.

A perfect design score can conceal relation-sensitive failure

Pool B isolates the target and copy. Across the 240 composite queries, ORB records 237 target wins and 3 copy wins; raw pixels 187 target and 53 copy wins; HOG 177 and 63; and InkTrace 180 and 60. pHash records 181 target wins, 29 copy wins and 30 ties. CAR records 178 target wins, 1 copy win and 61 ties. DINOv2 records 160 target and 80 copy wins. SuperPoint-LightGlue records 240 target wins in its exploratory Pool B run, but it has no controlled A-to-C result and no untouched confirmation set.

Every method nevertheless has design R@1 of 1.000 in Pool B, including methods that select the copy many times, because both candidates belong to the same working design. The metric is not numerically wrong. It answers a different question. This is the empirical form of the validity gap: design recall remains perfect while the operational target-versus-copy decision fails.

An abstaining gate still accepts two copy-class records

The ORB/HOG gate accepts 178 of the 240 composite queries and routes 62 to review, giving coverage of 0.742. Two accepted decisions select the copy rather than the intended target. Selective risk is 2/178, or 0.011. Resampling the ten design clusters produces a descriptive sensitivity interval from 0.000 to 0.038. Both errors occur at severity 4, in two seeded instances of the same query parent and design. Their concentration in one cluster limits generalisation, but the events are sufficient to show that agreement plus abstention does not guarantee a copy-free accepted set.

Figure 3 documents one event. The query parent is Met 388040. The gate selects Met 387744, catalogued as *The Nativity (copy)* with the artist prefix "After" Albrecht Dürer, rather than intended target Met 387391, which lacks a copy marker under the operational rule. Frozen copy/target scores are 0.080000/0.069091 for ORB and 0.885180/0.883284 for HOG. The descriptors agree because the same-design copy provides strong visual evidence. Their agreement does not supply source, material or catalogue evidence for the historical relation.

The reconstructed query shown in Fig. 3 is intentionally blurred, colour-shifted and block-like at severity 4. These artefacts display the declared stress transformation rather than damage to the figure. The event is a real error from the frozen experimental gate under the catalogue-class task. It is not claimed as a historical misidentification by a commercial search engine, nor as proof that the two physical sheets have or lack a particular matrix relation.

Discussion

Candidate composition is an epistemic design choice

The results locate the main problem before model interpretation. Pool A and Pool C use the same queries, intended targets and ordinary negatives. For pairwise methods they also use the same existing scores. Adding one relation-confusable candidate is enough to change a perfect-looking result into 53 raw-pixel intrusions, 63 HOG intrusions or 3 ORB intrusions. The benchmark's apparent difficulty was therefore partly produced by what it excluded.

This does not invalidate repeated-design retrieval. It narrows the claim that the score can support. If a project's purpose is to retrieve any representation of a composition, design R@1 may be appropriate and a same-design copy may be relevant rather than erroneous. If the purpose is to help distinguish catalogue classes, states, matrices or copying relations, design R@1 is insufficient. Validity depends on matching the evaluation construct to the research question, not on declaring one metric universally correct.

The distinction also clarifies why more powerful embeddings are not a complete solution. A model trained to be invariant to crop, colour and local damage may become better at collapsing objects that share a design. That invariance is valuable for discovery and potentially harmful for a task that requires material or relation-sensitive differences. The proper response is not simply to seek a higher-capacity model, but to specify what variation should be ignored, what variation must remain visible and what evidence lies outside the image channel.

A relation-aware evidence architecture

The practical response proposed here is a three-layer evidence architecture, summarised in Table 3 and Fig. 1. Each layer has its own record, output and stopping rule.

The *visual-discovery record* stores the query surrogate, canonicalisation and transformation history, model and version, full candidate manifest, scores, tie policy and ranked output. Its permitted claim is deliberately visual: these records are similar under this declared procedure. Preserving the full candidate manifest is essential because rank has meaning only relative to the alternatives presented.

The *relation-confusability record* starts from the historical distinction relevant to the project. It identifies same-design alternatives that cross the operational boundary, documents how their labels were created or reviewed, and reports intended-target loss, strict intrusions, ties, abstention, coverage and accepted-set risk. Its output is not a historical conclusion but an evaluation statement: the retrieval procedure did or did not preserve this declared distinction on these cases.

The *historical-assertion record* identifies the physical objects and states the proposed relation, supporting and conflicting sources, responsible agent, date, uncertainty and revision history. Relevant evidence may include catalogue scholarship, provenance, watermark, paper, colourant, inscriptions, damage, carved line or direct object examination. The record should distinguish assertion from evidence and permit unresolved status. CIDOC CRM can represent cultural-heritage entities and events, while PROV-O can record agents, activities and derivations (Doerr 2003; Lebo et al. 2013). The proposal does not require one database platform; it requires the inferential boundary to survive data exchange.

Table 3
Minimum relation-aware evidence architecture

Record layer	Minimum fields	Permitted output	Escalation rule
Visual discovery	Query and surrogate IDs; preprocessing; model/version; full candidate manifest; scores; tie rule	Visual similarity under the declared procedure	Do not infer a historical relation from rank alone
Relation confusability	Relation of interest; difficult alternatives; label source; intrusion, tie, abstention and coverage outcomes	Whether the procedure preserved the operational distinction on the test cases	Route failed, tied or unsupported cases to review
Historical assertion	Physical object IDs; proposed relation; sources; evidence; agent; date; uncertainty; revision history	A source-bearing, revisable historical assertion or unresolved status	Require documentary, material or specialist review
Audit link	Identifiers connecting all three layers; software and data versions; timestamps and derivations	Traceable movement from discovery to judgement	Preserve the earlier records when an assertion changes

Source: created for this study; compatible in principle with CIDOC CRM entity/event modelling and PROV-O provenance recording.

Four operational rules follow. First, log the candidate set rather than storing only the chosen result. Second, report relation-confusable intrusions and ties separately from design recall. Third, evaluate an abstaining system on both coverage and the errors remaining inside the accepted set. Fourth, do not promote a model's relation hypothesis into catalogue metadata or historical prose without a source-bearing review action. These rules convert a warning into a workflow that can be audited, reproduced and revised.

Consequences for digital cultural heritage

For collection interfaces, the framework suggests that similarity search should expose rather than flatten uncertainty. Results can be grouped by proposed design family while displaying catalogue relation fields, source institution and confidence or review status. A copy should not be hidden because it complicates the ranking; it is often precisely the object a historian needs to see. At the same time, the interface should avoid presenting visual proximity as proof of identity. The useful output is a navigable comparison set with traceable provenance.

For dataset builders, candidate selection becomes a documented scholarly act. A datasheet should state the historical construct, unit of analysis, operational labels, sources of difficult negatives, unresolved cases and known exclusions. Train, development and test designs should be separated before method tuning, and an untouched multi-institution set should be retained for confirmation.

Pairwise relation labels should cite the evidence and allow disagreement rather than forcing every case into a binary original/copy schema.

For computational researchers, relation-confusable evaluation provides a low-cost stress test before expensive model development. Known copies, recuts, later states or reproductions can be inserted into an unchanged candidate pool. If rankings collapse, the result identifies either a model limitation or a mismatch between task and claim. If rankings remain stable, the finding is stronger than an easy design benchmark, although it still requires independent labels and external validation.

For historians and conservators, the architecture protects the value of visual computation without delegating historical authority to it. Similarity can prioritise cases for close looking, reveal dispersed families and direct scarce material analysis. Documentary and physical evidence can then explain why visually close objects diverge. The stages are complementary: discovery expands the field of comparison, confusability testing disciplines the model claim and historical judgement carries the evidential burden.

Transferability and evaluation agenda

The proposed method can be adapted wherever a visually persuasive match may cross a meaningful relation boundary. A photographic archive might test prints from different negatives; a sculpture collection might introduce casts from different moulds; manuscript research might compare exemplars copied from a common source; archaeology might distinguish documented replicas from excavated objects. In each case the operational class must be domain-specific, and a visually similar candidate is an error only relative to the declared task.

A stronger future benchmark for print history should use specialist-reviewed pairwise relation records across institutions. Each pair should identify the physical object and digital surrogate; distinguish same matrix, later state, recut block, replacement block, copy after, reproduction and unresolved; cite the evidence; record reviewer disagreement; and allow revision. Discovery and relation-sensitive evaluation should be reported separately. A subsequent user study could compare a similarity-only interface with the relation-aware interface proposed here, measuring not only retrieval speed but also whether users make better-supported historical judgements. That human-centred test is an important next step, especially given evidence that expert review remains necessary even when automated catalogue output appears plausible (Abele et al. 2026).

Limitations

The primary corpus is small, institution-specific and intentionally constructed. Its effective independent structure is ten challenge designs, not 240 independent historical objects. The transformed queries are correlated stress instances, and severity 4 has no prevalence interpretation. Design-cluster intervals communicate sensitivity to this dependence but cannot turn the sample into a population estimate.

The operational labels are unary catalogue classes. They do not establish pairwise matrix identity, original status, copying direction, attribution or authenticity. Unmarked or differently described copies may remain in the target class. The errors therefore show failure to preserve a declared catalogue distinction, not a connoisseurial determination of the objects' physical history.

The archive also has limits. The initial Met sampling frame, exclusion log, original checker identity, independently timestamped split freeze, original generated query bytes and feature-correspondence masks are unavailable. Current execution reproduces the frozen decisions and reconstructs queries deterministically, but it cannot recover every feature-level mechanism of the historical run. Modern baselines were added after challenge designs were known and should be treated as exploratory.

Finally, the evidence architecture has been specified and applied retrospectively to the benchmark, not evaluated as a working museum interface. Its effects on research quality, workload and user trust require testing with print historians, cataloguers and conservators. The study should therefore be read as a validated benchmark intervention and an implementable design proposal, not as proof that the proposed workflow has already improved professional decisions.

Conclusion

Visual similarity is indispensable for finding repeated designs, but its evidential meaning depends on what the benchmark permits the system to see. In this study, raw pixels, HOG and ORB rank the intended target first on all 240 Pool A queries. Adding one same-design explicit-copy record lowers those counts to 187, 177 and 237 and creates 53, 63 and 3 strict copy intrusions. A gate that can abstain still accepts two copy-class records. Meanwhile, design R@1 remains perfect in the target-versus-copy pool because the metric treats both objects as the same design.

The finding is not that visual retrieval is unusable, nor that an image model should authenticate prints. It is that candidate composition defines the construct being measured. A design-retrieval benchmark cannot warrant a claim about historical sameness unless relation-confusable alternatives are represented and the relevant relation is independently documented. Separating visual discovery, confusability testing and source-bearing historical judgement provides a practical way to preserve both computational usefulness and historical accountability. For digital cultural heritage, the decisive question is no longer only whether the model finds a similar image, but whether the evaluation makes visible the difference the research is trying to understand.

Declarations

Competing interests

The author declares no competing interests.

Supplementary information

Supplementary Information accompanies the article and contains the corpus and candidate-pool documentation, operational-label rules, transformation parameters, complete query-level results, error records, statistical sensitivity analyses, environment details and reproduction instructions.

Funding:

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Author Contribution

J.F. conceived and designed the study, collected and curated the data, developed the methodology and software, performed and verified the analyses, prepared the figures, interpreted the results, and wrote and revised the manuscript. J.F. approved the final manuscript and accepts full responsibility for the work.

Data Availability

The data supporting the findings of this study, including record manifests, operational labels, query definitions, query-level results, gate decisions, parameter files, checksums, environment specifications and analysis code, are provided in the accompanying Supplementary Information and supplementary data-and-code files. Source museum images are not redistributed; the manifest provides object identifiers, source URLs, rights information and checksums, and the included scripts retrieve or reconstruct the analysed inputs.

References

1. Abell C (2015) Printmaking as an art. J Aesthet Art Crit 73(1):23–30. https://doi.org/10.1111/jaac.12137
2. Abele L, Anders G, Aydın T, Buder J, Fischer H, Kimmel D, Huff M (2026) ArchiveGPT: a human-centered evaluation of using a vision language model for image cataloguing. Humanit Soc Sci Commun 13:1241. https://doi.org/10.1057/s41599-026-08367-6
3. Art Institute of Chicago (2026a) API documentation. https://api.artic.edu/docs/. Accessed 17 Aug 2026
4. Art Institute of Chicago (2026b) Open Access Images. https://www.artic.edu/open-access/open-access-images. Accessed 17 Aug 2026

5. Arnold T, Tilton L (2019) Distant viewing: analyzing large visual corpora. Digit Scholarsh Humanit 34(Suppl 1):i3–i16. https://doi.org/10.1093/llc/fqz013
6. Blanke T (2024) Reassembling digital archives-strategies for counter-archiving. Humanit Soc Sci Commun 11:201. https://doi.org/10.1057/s41599-024-02668-4
7. Borgman CL, Scharnhorst A, Golshan MS (2019) Digital data archives as knowledge infrastructures: mediating data sharing and reuse. J Assoc Inf Sci Technol 70(8):888–904. https://doi.org/10.1002/asi.24172
8. Cha J (2026) Automating the past: artificial intelligence and the next frontiers of digital history. Int J Humanit Arts Comput 20(1):1–17. https://doi.org/10.3366/ijhac.2026.0361
9. Chung JS, Yoon K, Arandjelović R, Zisserman A (2015) Re-presentations of art collections. In: Agapito L, Bronstein MM, Rother C (eds) Computer vision-ECCV 2014 workshops. Springer, Cham, pp 85–100. https://doi.org/10.1007/978-3-319-16178-5_6
10. Cook RT (2023) Printmaking: impressions, editions, and reproductions. In: Carroll N, Gilmore J (eds) The Routledge companion to the philosophies of painting and sculpture. Routledge, Abingdon, pp 28–38. https://doi.org/10.4324/9781003312727-5
11. Cronbach LJ, Meehl PE (1955) Construct validity in psychological tests. Psychol Bull 52(4):281–302. https://doi.org/10.1037/h0040957
12. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol 1. IEEE, pp 886–893. https://doi.org/10.1109/CVPR.2005.177
13. DeTone D, Malisiewicz T, Rabinovich A (2018) SuperPoint: self-supervised interest point detection and description. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. IEEE, pp 224–236. https://doi.org/10.1109/CVPRW.2018.00060
14. Doerr M (2003) The CIDOC conceptual reference model: an ontological approach to semantic interoperability of metadata. AI Mag 24(3):75–92. https://doi.org/10.1609/aimag.v24i3.1720
15. Drucker J (2011) Humanities approaches to graphical display. Digit Humanit Q 5(1). https://digitalhumanities.org/dhq/vol/5/1/000091/000091.html. Accessed 17 Aug 2026
16. Fischler MA, Bolles RC (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun ACM 24(6):381–395. https://doi.org/10.1145/358669.358692
17. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé H III, Crawford K (2021) Datasheets for datasets. Commun ACM 64(12):86–92. https://doi.org/10.1145/3458723
18. Hindermann M, Kasper LK, Marti S, Bosse A (2026) From experiments to epistemic practice: the RISE Humanities Data Benchmark. J Open Humanit Data 12:38. https://doi.org/10.5334/johd.470
19. Impett L, Offert F (2022) There is a digital art history. Vis Resour 38(2):186–209. https://doi.org/10.1080/01973762.2024.2362466

20. Jacobs AZ, Wallach H (2021) Measurement and fairness. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. ACM, New York, pp 375–385. https://doi.org/10.1145/3442188.3445901
21. Karjus A (2025) Machine-assisted quantitizing designs: augmenting humanities and social sciences with artificial intelligence. Humanit Soc Sci Commun 12:277. https://doi.org/10.1057/s41599-025-04503-w
22. Korenberg CF, Pereira-Pardo L, McElhinney P, O'Mahony M, Strlič M (2019) Developing a systematic approach to determine the sequence of impressions of Japanese woodblock prints: the case of Hokusai's Red Fuji. Herit Sci 7:9. https://doi.org/10.1186/s40494-019-0250-5
23. Kräutli F, Lockhorst D, Valleriani M (2021) Calculating sameness: identifying early-modern image reuse outside the black box. Digit Scholarsh Humanit 36(Suppl 2):ii165–ii174. https://doi.org/10.1093/llc/fqaa054
24. Lang S, Ommer B (2018) Attesting similarity: supporting the organization and study of art image collections with computer vision. Digit Scholarsh Humanit 33(4):845–856. https://doi.org/10.1093/llc/fqy006
25. Lebo T, Sahoo S, McGuinness D (2013) PROV-O: the PROV ontology. W3C Recommendation. https://www.w3.org/TR/2013/REC-prov-o-20130430/. Accessed 17 Aug 2026
26. Lindenberger P, Sarlin PE, Pollefeys M (2023) LightGlue: local feature matching at light speed. In: 2023 IEEE/CVF International Conference on Computer Vision. IEEE, pp 17627–17638. https://doi.org/10.1109/ICCV51070.2023.01616
27. Metropolitan Museum of Art (2026a) The Met Collection API. https://metmuseum.github.io/. Accessed 17 Aug 2026
28. Metropolitan Museum of Art (2026b) Open Access at The Met. https://www.metmuseum.org/hubs/open-access. Accessed 17 Aug 2026
29. Noehrer L, Gilmore A, Jay C, Yehudi Y (2021) The impact of COVID-19 on digital data practices in museums and art galleries in the UK and the US. Humanit Soc Sci Commun 8:236. https://doi.org/10.1057/s41599-021-00921-8
30. Oquab M et al (2023) DINOv2: learning robust visual features without supervision. https://doi.org/10.48550/arXiv.2304.07193. arXiv
31. Roberts JL (2024) Contact: art and the pull of print. Princeton University Press, Princeton. https://doi.org/10.1515/9780691255866
32. Rublee E, Rabaud V, Konolige K, Bradski G (2011) ORB: an efficient alternative to SIFT or SURF. In: 2011 International Conference on Computer Vision. IEEE, pp 2564–2571. https://doi.org/10.1109/ICCV.2011.6126544
33. Sanderson M (2010) Test collection based evaluation of information retrieval systems. Found Trends Inf Retr 4(4):247–375. https://doi.org/10.1561/1500000009
34. van Noord N (2022) A survey of computational methods for iconic image analysis. Digit Scholarsh Humanit 37(4):1316–1338. https://doi.org/10.1093/llc/fqac003

35. Vermeulen M et al (2020) Beyond the connoisseurship approach: creating a chronology in Hokusai prints using non-invasive techniques and multivariate data analysis. Herit Sci 8:62. https://doi.org/10.1186/s40494-020-00406-y
36. Voorhees EM (2002) The philosophy of information retrieval evaluation. In: Peters C, Braschler M, Gonzalo J, Kluck M (eds) Evaluation of cross-language information retrieval systems. Springer, Berlin, pp 355–370. https://doi.org/10.1007/3-540-45691-0_34

Figures

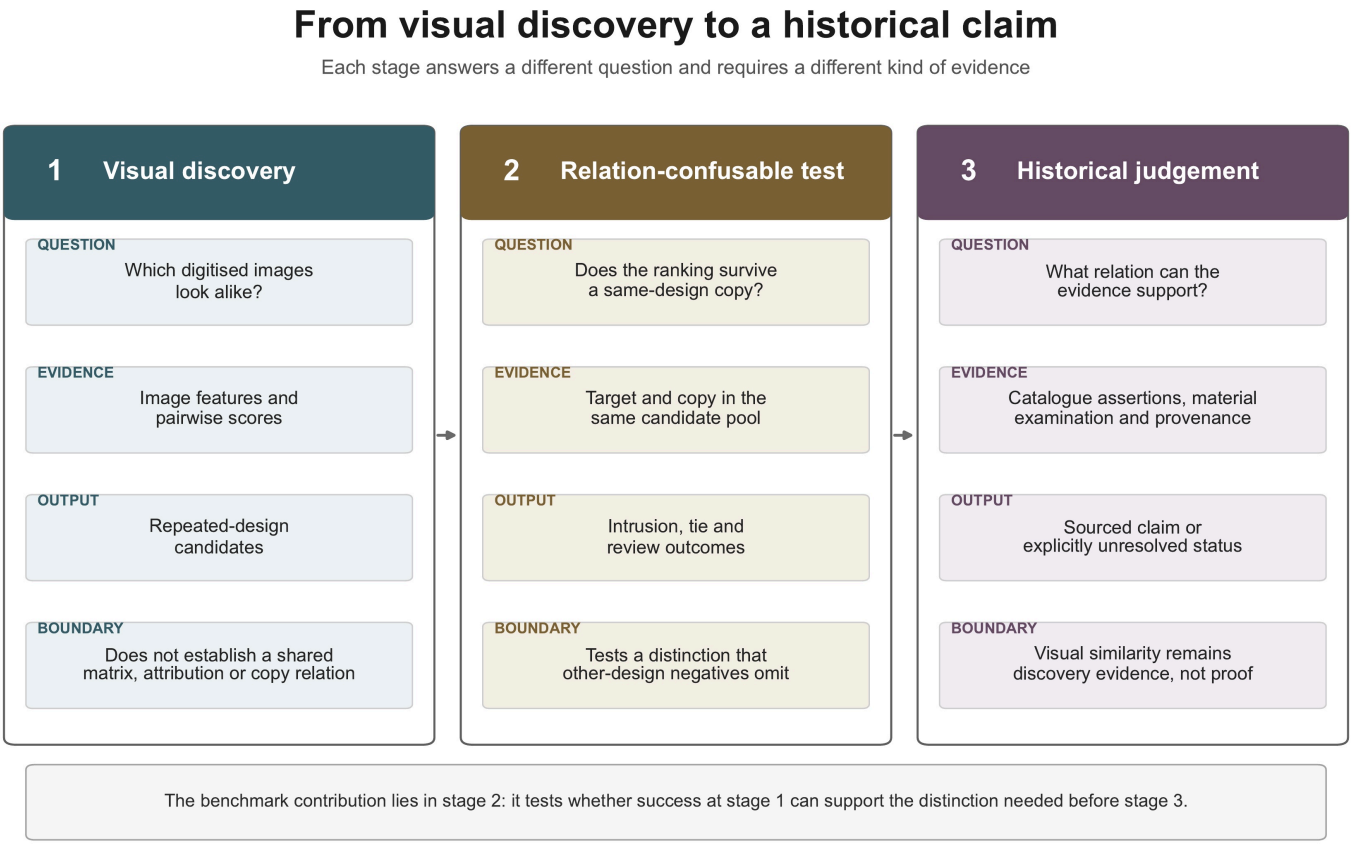


Figure 1

From visual discovery to a source-bearing historical claim. Stage 1 ranks digital surrogates by appearance. Stage 2 introduces relation-confusable candidates and tests whether the ranking preserves the operational distinction. Stage 3 records the relation that documentary, catalogue, material or specialist evidence can support. Source: created for this study.

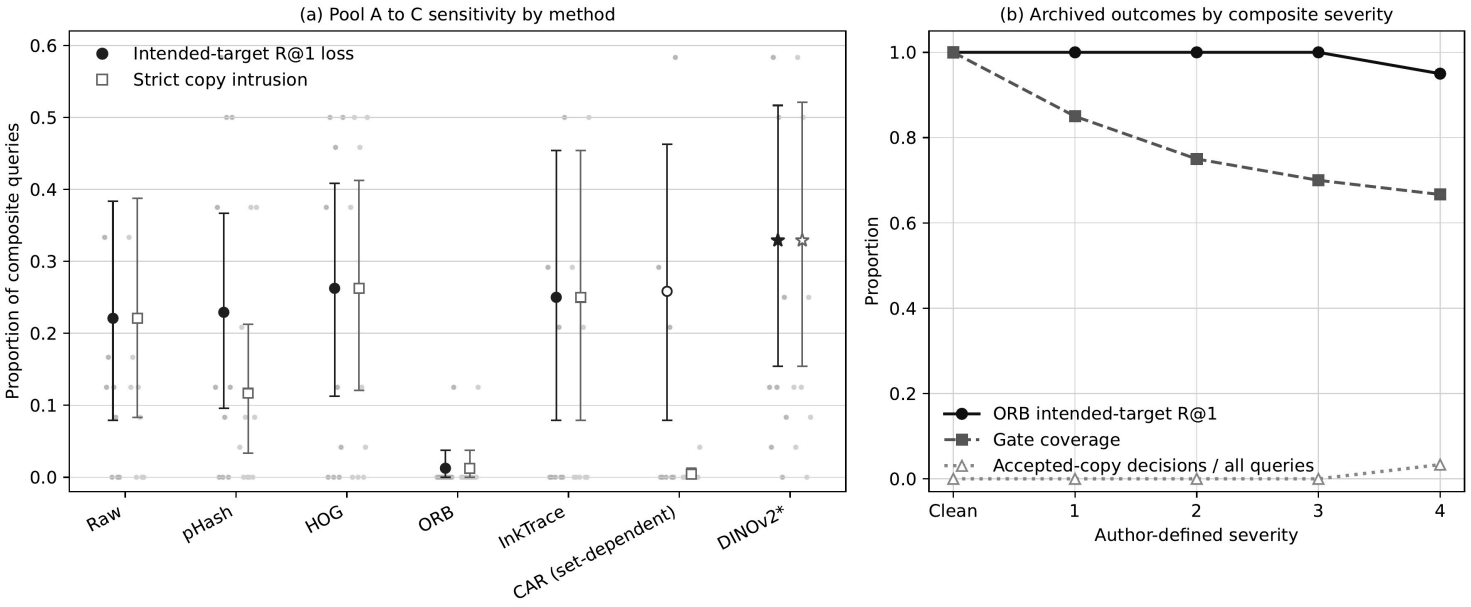
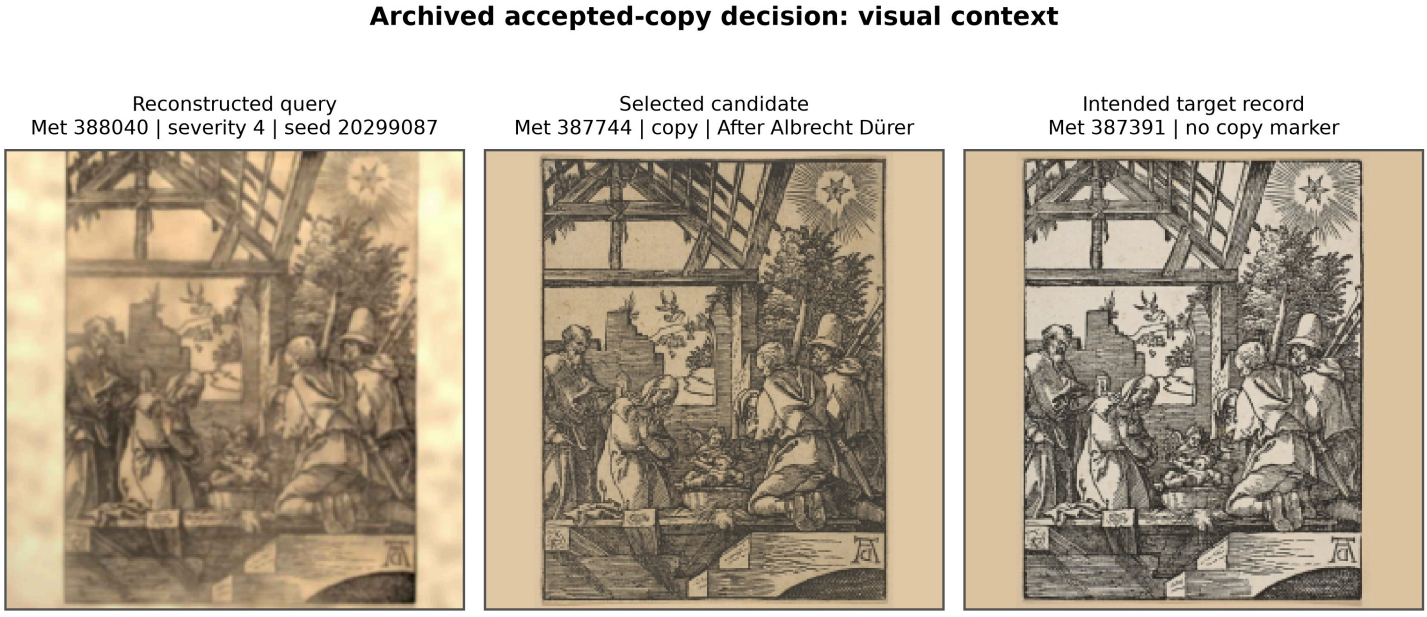


Figure 2

Candidate intrusion and stress severity. Panel a separates intended-target R@1 loss from strict copy intrusion for the Pool A-to-C intervention. Points show the observed proportions; vertical bars show the 2.5th-97.5th percentiles from 10,000 bootstrap resamples of the ten design clusters. These are descriptive sensitivity intervals, not population confidence intervals. Panel b reports frozen ORB and ORB/HOG gate outcomes by author-defined stress severity. Source: created for this study.



Archived scores, copy/target: ORB 0.080000/0.069091; HOG 0.885180/0.883284. The query is a deterministic reconstruction; historical correspondence masks were not archived, so feature-level mechanism analysis is not attempted.

Figure 3

One of two accepted-copy decisions from the same parent and design. The query is a deterministic reconstruction with declared severity-4 blur, colour variation and block-like artefacts; original frozen-run query bytes and feature masks are unavailable. Met 387744 is catalogued as a copy after Albrecht Dürer, whereas Met 387391 has no copy marker under the operational rule. Source images: Metropolitan Museum of Art Open Access object records 388040, 387744 and 387391, made available under the Creative Commons CC0 1.0 Universal Public Domain Dedication. Analytical arrangement and deterministic query transformation: created for this study. No generative AI was used.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- HSSCSupplementaryDataandCode.zip
- SupplementaryInformation.docx