**International Journal of Recent Innovations in Academic Research**

E-ISSN: 2635-3040; P-ISSN: 2659-1561
Homepage: https://www.ijriar.com/
Volume-10, Issue-3, July-September-2026: 66-81

**Research Article**

# From Similarity to Plagiarism: Source Attribution in AI-Generated Art

**Byungkil Choi**

Department of Arts, College of Arts and Design, Wonkwang University, Iksan, Republic of Korea
Email: cebhi@naver.com

**Article History**
**Received:** July 28, 2026
**Accepted:** August 19, 2026
**Published:** August 25, 2026

**Abstract**

Existing AI-art plagiarism methods mainly measure visual or semantic similarity between generated and prior images. Similarity can identify suspicious correspondence, but it does not establish that a particular work caused the output or that plagiarism occurred. This treatise therefore separates observable similarity, inferred source dependence, and evaluative plagiarism, integrating perceptual, semantic, compositional, motif-level, stylistic, provenance, and training-data evidence. It distinguishes memorized replication, compositional or fragmentary appropriation, style imitation, and independent convergence; retrieves candidate sources through multi-level correspondence; and evaluates source identifiability, distinctiveness, causal support, transformation, acknowledgment, authorization, and alternatives. Across GAN and diffusion outputs, the framework produces explainable attribution categories and calibrated confidence rather than an automatic copyright judgment, advancing AI-art plagiarism analysis from similarity measurement to explainable source attribution and evidence-based classification.

**Keywords:** AI-Generated Art, Plagiarism, Source Attribution, Training-Data Memorization, Visual Similarity, Compositional Appropriation, Style Imitation, Provenance, Generative Adversarial Networks, Diffusion Models.

## 1. Introduction

Artificial-intelligence image generators have altered both the scale and the evidentiary character of disputes concerning artistic borrowing. The apparent source of a resemblance may therefore be a single artwork, several works contributing different fragments, a general artistic style, a shared cultural reference, or no identifiable source at all. This complexity makes visual resemblance important, but it also makes resemblance an insufficient endpoint of analysis. Most computational approaches begin by measuring how closely a generated image resembles a pre-existing work. Yet they ordinarily answer a descriptive question—how similar are the images?—rather than the more demanding inferential and evaluative questions at the center of plagiarism: whether the generated image is dependent on a particular prior work, which expressive features are attributable to that work, and whether the resulting dependence warrants an adverse classification. Even perceptual metrics designed to reflect holistic human judgments remain measures of correspondence; they do not, without further evidence, establish the historical or causal relationship between two images [1]. The need for a more discriminating framework is not merely theoretical. Empirical studies have also identified content replication in prominent diffusion systems and have associated replication risk with such factors as dataset duplication, dataset size, captions, and model behavior. These findings demonstrate that source-dependent reproduction is technically possible and that some visual correspondences may reflect memorization rather than independent synthesis. A framework that searches only for confirming similarities will therefore tend to confuse suspicion with proof [2].

The central problem is the inferential gap between observable similarity and a reasoned plagiarism finding. Analysis must proceed through three stages: identifying the relevant prior work or works, determining whether distinctive correspondence supports source dependence rather than an innocent explanation, and evaluating established dependence in light of transformation, acknowledgment, authorization, and context.

Recognition of suspicious resemblance and reliable source attribution are therefore related but non-equivalent tasks [3]. This treatise builds on the author's three earlier studies, which developed a multi-metric plagiarism scale, a multimodal semantic and compositional method for GAN-generated abstract art, and a portable severity tool across GAN and diffusion outputs. Those studies did not fully address competing source candidates, region-level multi-source attribution, nonvisual provenance, or uncertainty when the true source is absent [4]. Its principal thesis is that visual similarity is evidence of correspondence, not proof of plagiarism. A defensible plagiarism attribution in AI-generated art requires, at minimum, one or more plausible source works; correspondence in expressive features that are sufficiently distinctive to carry source-identifying weight; evidence supporting source dependence; and a structured examination of alternative explanations. Where the evidence is incomplete or conflicting, it should report possible dependence or an indeterminate result rather than convert numerical similarity into unwarranted certainty.

### 1.1. Research Questions and Scope

The treatise addresses five connected questions: what evidence moves an inquiry from similarity to source dependence; how memorized replication, compositional or fragmentary appropriation, style imitation, and independent convergence can be distinguished; how visual measures should be integrated with training, prompt, reference-image, editing, and provenance records; how single-source, multi-source, and null-source outcomes should be represented; and whether the framework remains reliable across GAN and diffusion systems. The sequence moves from correspondence through source identification and source-dependence assessment to source attribution, and finally to plagiarism classification.

The scope of the inquiry is correspondingly limited. It concerns two-dimensional artistic images generated by GAN- and diffusion-based systems, with human-created artworks serving as the principal candidate sources. The framework is intended to support ethical, scholarly, institutional, and expert assessment of plagiarism; it does not purport to make a definitive judicial determination of copyright infringement. Plagiarism, by contrast, is generally concerned with the unacknowledged appropriation or presentation of another's contribution as one's own, although its precise norms vary by discipline and context [5].

Three different meanings of attribution must also be kept separate. Source-work attribution asks which prior artwork materially explains the expressive features of the disputed output. Evidence that a model generated an image does not identify the artwork on which the image depends; evidence that a training example influenced the output does not by itself establish appropriation; and a visually plausible source candidate may remain unproven when training, reference, or generation records are unavailable [6]. For terminological clarity, source identification refers to selecting the plausible prior work or works; source dependence refers to the supported inference that such work materially explains distinctive expression in the output; source attribution refers to the evidentiary process and report that connect the identified source to that dependence; and plagiarism classification is the subsequent normative evaluation. Keeping these stages distinct prevents evidence relevant to one proposition from being treated as proof of another.

The treatise therefore excludes disputes concerning ownership that are unrelated to source dependence, nonvisual media such as music and literary text, the general detection of whether an image is AI-generated, and attribution solely to a model family or generator. The framework must accordingly permit a null-source outcome and must treat indeterminacy as a scientifically valid conclusion rather than a failure that should be concealed by a forced binary label.

### 1.2. Methodology and Contributions

The first method is conceptual analysis, which distinguishes similarity, expressive correspondence, source identification, source dependence, appropriation, transformation, acknowledgment, authorization, and plagiarism. This ordering avoids circularly treating a high similarity score as both evidence and conclusion and separates work-specific dependence from style resemblance and independent convergence.

The second method is computational analysis. Multiple perceptual, semantic, compositional, and local representations retrieve candidate sources; hard negatives test whether correspondence is source-specific; and a multi-level model determines which regions or features each candidate explains. The system then compares single-source, multi-source, source-cluster, and null-source accounts without assuming universal metric weights.

The third method is evidentiary integration coupled with a proposed empirical validation protocol. Visual correspondence is integrated with dataset membership, duplication, fine-tuning, reference-image use,

prompts, seeds, control inputs, editing history, and provenance. The proposed validation protocol uses controlled replication, appropriation, imitation, convergence, and unrelated cases to separate source retrieval from classification and to evaluate false attribution, source absence, region-to-source assignment, calibration, and stage-specific errors.

The treatise contributes a taxonomy of AI-art source dependence, a multi-level single- and multi-source attribution framework, a plagiarism evidence matrix combining distinctiveness, causal support, transformation, attribution, authorization, and alternatives, and a cross-model evaluation protocol. Together, these contributions make attribution more transparent, contestable, and proportionate to the available evidence.

The chapters follow the same inferential sequence: Chapter 2 separates similarity from plagiarism; Chapter 3 classifies forms of source dependence; Chapters 4 and 5 develop attribution and evidentiary assessment; Chapters 6 and 7 present classification and validation; and Chapter 8 addresses legal and ethical limits. The conclusion returns to explainable inference rather than automatic similarity scoring.

## 2. From Visual Similarity to Plagiarism

### 2.1. The Limits of Similarity-Based Detection

Similarity and plagiarism belong to different analytical categories. Similarity describes a relationship between perceptible or representational features of two images, whereas plagiarism attributes an objectionable form of dependence and presentation. Similarity cannot, however, establish by itself that the alleged source entered the training or generation process, that the model or user had meaningful access to it, that the shared features are distinctive rather than conventional, that transformation was insufficient, that acknowledgment or authorization was absent, or that coincidence and common-source explanations are implausible. Those conclusions require evidence beyond resemblance.

The limitations become clearer when the principal classes of similarity metric are considered separately. Perceptual metrics are more robust because they compare learned visual features rather than raw pixels, yet they may assign excessive weight to texture or local appearance and may fail to capture the significance of a distinctive spatial arrangement. DreamSim and related approaches were developed in part because earlier metrics did not fully track human judgments about layout, pose, semantic content, and holistic appearance, but even these improved measures remain evidence of perceived correspondence rather than evidence of source dependence [7].

A second limitation concerns the comparison set. A pairwise score can appear impressive when the disputed image is compared only with the alleged source, yet lose much of its significance when works by the same artist, works in the same genre, or works depicting the same subject are added. A blue-and-gold palette, a frontal portrait, or a centrally placed figure may be visually salient without being source-specific. Similarity analysis must therefore be comparative and contextual rather than pairwise and absolute.

The same point applies to thresholds. Metrics, genres, and forms of borrowing do not have commensurable meanings. A high pixel score is probative for near-exact replication but relatively uninformative for style imitation, while moderate global similarity may conceal exact local copying. Thresholds must therefore be calibrated to the correspondence type, reference corpus, and error cost, with false attribution weighted heavily in reputationally sensitive settings.

The proper conclusion is not that similarity metrics are unreliable, but that their evidentiary role must be confined. A responsible system reports what the metric detected, where the correspondence occurs, how the result compares with hard negatives, and what transformations were required to align the works. The transition from resemblance to plagiarism requires an intermediate inquiry into source dependence and a subsequent normative assessment of appropriation, transformation, attribution, authorization, and alternative explanation.

This distinction also has consequences for model training and evaluation. A source-attribution benchmark should therefore annotate the intermediate propositions separately: whether a candidate source is present, which features correspond, how distinctive those features are, whether a plausible causal pathway exists, and whether alternatives remain. By supervising the inferential stages rather than only the final label, the system can be evaluated for the reasons that lead to its conclusion and can reveal when a correct classification rests on an unreliable shortcut.

### 2.2. Source Dependence and Competing Explanations

Source dependence provides the conceptual bridge between similarity and plagiarism. The contribution need not preserve every feature of the source, nor must it arise through a single technical pathway. What matters is whether the prior work helps explain the presence and organization of distinctive expression in a way that general learning, conventions, and independent creation do not. Several pathways can produce such dependence. A direct pathway to dependence exists when a source image is supplied as an image-to-image input, editing base, inpainting target, or control image; whether source dependence is ultimately established still depends on whether distinctive source expression materially contributes to the output. Reference-mediated dependence arises when retrieval-augmented generation, adapters, fine-tuning, or other conditioning mechanisms introduce a prior work into the generation process. Dependence may also be diffuse: separate works may account for different regions, motifs, poses, or compositional relationships, so that no single source explains the output as a whole.

These pathways are evidentially and normatively distinct. A verified reference-image input provides direct evidence that the work was available to the generation process, while training-mediated dependence may have to be inferred from dataset membership, duplication, reproducibility, or data-attribution methods. The framework therefore separates the question of how dependence arose from the question of whether the resulting use should be classified as plagiarism. Data-attribution research supports tracing influence but also demonstrates substantial uncertainty. Unlearning, influence-function, and training-data influence methods may identify influential examples, but they do not prove appropriation of distinctive expression and may distribute influence across related images. Influence is therefore causal evidence, not an automatic finding of appropriation [8].

For purposes of this treatise, source dependence is best understood as a supported inference that an identifiable prior work materially contributed to the distinctive visual expression of a generated image beyond what can reasonably be explained by general training, common conventions, shared sources, or independent convergence. It requires an identifiable object of attribution, focuses on expressive rather than merely statistical contribution, and preserves room for alternative explanations. A source-dependent image is not necessarily plagiaristic. Artistic practice has always involved study, influence, quotation, homage, parody, transformation, and participation in shared traditions. The decisive question is not whether a prior work contributed in some sense, but how specifically that work contributed, which elements were preserved, how the contribution was transformed, whether it was acknowledged or authorized, and how the resulting image was represented to others. Plagiarism is therefore an evaluative conclusion about appropriation and attribution, not a synonym for causal influence.

Inspiration ordinarily operates at a level of general theme, technique, mood, or problem-solving strategy. Style imitation is more difficult because a model or user may reproduce an artist-level visual vocabulary without copying a specific work. Such imitation may raise ethical, labor, identity, or market concerns, but work-level plagiarism normally requires more than an artist-level resemblance. Appropriation occupies an intermediate domain in which distinctive expression is taken or reorganized, but the final judgment depends on transformation, acknowledgment, authorization, and context. Coincidence and independent convergence must be treated as genuine hypotheses rather than residual excuses. Similar prompts can narrow the space of possible results, especially when the subject has a conventional iconography or only a few familiar visual representations. The task is therefore not merely to find features consistent with copying but to compare the dependence hypothesis with these alternatives.

Plagiarism must also be distinguished from copyright infringement. Conversely, unattributed borrowing may be ethically treated as plagiarism even when the source is in the public domain, the borrowed element is unprotectable, or the legal claim is otherwise unavailable. The framework developed here therefore informs but does not replace legal analysis. A disciplined attribution system should test competing hypotheses in an explicit order. It should first determine whether a specific source is identifiable, then ask whether the shared features are distinctive and materially preserved, and then evaluate evidence of access or causal contribution. When a common style, shared reference, generic prompt, or independent reconstruction explains the output at least as well as the alleged source, the system should decline an adverse classification.

## 3. Forms of Source Dependence in AI-Generated Art

### 3.1. Memorized and Near-Exact Replication

The clearest form of source dependence is memorized or near-exact replication. At one end of this category lies literal reproduction, in which the generated image preserves essentially the same pixels or visual

structures as a training or reference image. Near-exact replication may additionally involve cropping, recoloring, resizing, mirroring, or other modest transformations. Such changes may reduce a simple similarity score without changing the basic fact that the distinctive expression of the source has been retained. Research on diffusion models demonstrates that individual training examples can sometimes be extracted or closely reproduced, although the frequency and conditions of replication vary. Survey work and studies of memorized-image subspaces further suggest that replication is a particular model behavior capable of analysis and mitigation, rather than an inevitable property of every generated resemblance [9].

A persuasive replication finding depends on converging indicators. Very high global correspondence is important, but it should be accompanied where possible by local alignment of rare details, artifacts, brush patterns, text fragments, watermark remnants, or idiosyncratic errors. Evidence that the source appeared in the training, fine-tuning, or reference set adds causal support, and reproducibility with a particular model, prompt, seed, or generation procedure may transform a visual suspicion into a technically grounded conclusion. The absence of these indicators does not prove independence. Verified replication requires direct or exceptionally strong evidence, such as a known reference-image input or repeatable generation behavior preserving the source. Probable replication may be supported by highly distinctive correspondence and plausible access even when the technical chain is incomplete. A merely high semantic score should not receive either label. Model memorization must also be separated from user-directed reconstruction. A user may deliberately supply a source image, engineer prompts that describe its composition in detail, generate numerous variations, and manually select or edit the closest result. The output may be as source-dependent as a memorized reproduction, but the causal route and allocation of responsibility differ. An evidentiary system should identify the form of dependence before drawing conclusions about human intent or responsibility.

### 3.2. Compositional and Fragmentary Appropriation

Many significant appropriations do not preserve the source as a whole. Compositional appropriation occurs when an output reproduces the organization of a work: the placement and relative scale of objects, the pose and gesture of figures, the viewing angle, perspective, foreground-background relationships, lighting structure, or arrangement of focal points. A system focused on full-image pixel or semantic similarity may therefore miss the very structure that makes the source recognizable. Fragmentary appropriation presents the inverse problem. A generated image may possess a substantially new overall composition while incorporating a source-specific face, character, symbol, motif, texture, distinctive object, area of brushwork, signature, or watermark remnant. Local segmentation, patch-level embeddings, feature matching, and geometric alignment are consequently indispensable.

The assessment of these forms should combine structural and distinctiveness analysis. A reproduced pose is weak evidence when it is common in the genre, but stronger when the pose is anatomically unusual and appears together with the same camera angle and lighting asymmetry. The evaluator should ask not only whether regions match but how likely the same configuration is to arise independently within the relevant comparison corpus. Compositional and fragmentary appropriation also reveal why source attribution may be multi-valued. Region-to-source assignment permits a more accurate account: each source is linked to the feature or area it explains, while confidence is expressed separately for each link. A source cluster may sometimes be more appropriate than an individual work when several closely related images share the relevant feature and the evidence cannot identify one member conclusively. The normative assessment should remain sensitive to transformation. The technical system can provide evidence that a fragment or composition is source-dependent and can measure the extent of preservation, but the strength of that inference must be assessed together with causal, provenance, and alternative-explanation evidence; the system cannot infer the meaning of the use solely from the match. Contextual evidence, acknowledgment, authorization, and human review remain necessary before a final plagiarism classification is made.

### 3.3. Style Imitation and Independent Convergence

Style imitation occupies a boundary position because style-level resemblance often fails to identify a particular source work. An individual artist's style may consist of recurring choices concerning line, palette, texture, composition, subject treatment, or medium simulation, but those choices can also overlap with a studio practice, school, genre, historical movement, or common prompt vocabulary. A generated image may thus appear recognizably "in the style of" an artist without preserving the distinctive expression of any one work. This distinction does not make style imitation ethically irrelevant. At the same time, a framework designed for source-work plagiarism should avoid treating a broad aesthetic resemblance as private ownership of a visual language. The U.S. Copyright Office has similarly distinguished style as such from the protectable elements of

particular works, while acknowledging that a purported style imitation may in practice reproduce work-specific expression and create market concerns [10].

The proper analysis therefore moves from artist-level to work-level evidence. If no work-specific candidate emerges and the shared elements are common within the artist's oeuvre or movement, the result should remain style imitation rather than source-work plagiarism. If a specific work accounts for distinctive expressive features, the label should shift from style similarity to the appropriate form of replication or appropriation. Independent convergence is the corresponding null hypothesis. In some visual problems, the available expressive choices are limited. A prompt requesting a passport-style portrait, a moonlit castle, or a heroic figure on a summit will predictably produce recurring arrangements. The evaluator must compare the disputed work not only with the alleged source but with a representative set of plausible alternatives. A finding of convergence is strongest when the shared elements are conventional, hard negatives achieve comparable scores, provenance supports an independent pathway, or the source was inaccessible to the relevant model or user. A system that is optimized only to detect plagiarism will overfit to resemblance and transform common visual culture into a series of false proprietary claims.

## 4. Source Attribution Framework

### 4.1. Candidate Retrieval and Correspondence Analysis

Source attribution begins with retrieval. The framework therefore searches a large reference collection and constructs a ranked set of candidate sources. The reference collection should be documented for provenance, licensing, coverage, artist and genre distribution, and known gaps, because retrieval confidence cannot exceed the representativeness of the corpus. Broad retrieval should combine complementary representations. CLIP-type representations provide semantic alignment, DINOv2 supplies robust image- and patch-level visual features, while holistic perceptual measures such as DreamSim may better capture layout, pose, and overall human-perceived resemblance. None of these representations should dominate by default; they create overlapping candidate pools that are merged and normalized for further analysis [11].

The broad stage should favor recall. The system can therefore retrieve candidates separately by global perception, semantics, composition, patch correspondence, and style. It should preserve candidates that rank strongly in only one channel, because a fragmentary appropriation may be invisible to global metrics and a transformed composition may be missed by pixel-based retrieval. Fine retrieval then subjects the highest-ranked candidates to more expensive analysis. These procedures should produce interpretable outputs: matched regions, transformation parameters, confidence values, and the features responsible for ranking. A candidate that appears similar only because both images contain a seated figure should be distinguished from one that reproduces the same unusual limb configuration, camera angle, and background geometry.

The final retrieval stage introduces hard negatives. If many works by the same artist achieve similar scores, the evidence may support artist-level style similarity but not attribution to one work. If the alleged source substantially outperforms same-subject and same-style alternatives on rare compositional and local features, its source-identifying weight increases. Retrieval must also permit source absence. The system should estimate whether the leading candidate is strong in absolute and comparative terms and should be capable of returning no source when the ranking is flat or the evidence is weak. Forcing a top-ranked item to function as the source merely because it is first in the list would convert database incompleteness into false attribution.

Once candidates have been retrieved, the framework evaluates correspondence at multiple levels. A convenient representation is a weighted model in which the attribution score for candidate i is expressed as follows:

$$R_i = w_p P_i + w_v V_i + w_c C_i + w_m M_i + w_s S_i + w_d D_i$$

In this expression, P_i represents pixel and perceptual correspondence, V_i semantic correspondence, C_i compositional correspondence, M_i motif or fragment correspondence, S_i stylistic correspondence, and D_i the distinctiveness or rarity of the shared features. Replication analysis may give greater weight to perceptual and local alignment, whereas compositional appropriation requires structural features and style imitation should not be allowed to dominate work-level attribution.

The distinctiveness term performs a particularly important function. A motif that appears in thousands of works contributes little source-identifying weight, while a rare defect, idiosyncratic pattern, or unusual conjunction of common elements may contribute substantially. The system should report the comparison

universe used, because rarity is relational rather than intrinsic. The weights in this representation are therefore task- and evidence-dependent rather than universal constants. They should be estimated or calibrated for the relevant dependence type and validation corpus, while conjunctive gatekeeping requirements prevent a strong score in one channel from compensating for the absence of source-identifying or causal evidence. Although represented additively for convenience, distinctiveness is conceptually also a modifier of the probative value of the other correspondence dimensions: a highly similar but conventional feature may carry less source-identifying weight than a moderately similar but highly distinctive one.

Correspondence should be localized. For each candidate, the report should identify which regions match, which metric detected the match, whether geometric transformation was required, and whether the feature is common or rare. The visual explanation is not an ornamental interface; it allows the reviewer to determine whether the system relied on a copied motif, a generic subject, an irrelevant texture, or an artifact of preprocessing. The framework should also distinguish preservation from transformation. A geometric alignment may reveal that a source has been mirrored or cropped, while object correspondence may show that the same compositional relation remains after substitution. A small transformation that leaves the source readily identifiable may strengthen a finding of disguised replication, whereas a series of changes that reorganizes the expressive relationships may support a more transformative account. Finally, metric convergence should be evaluated qualitatively as well as numerically. Convergence among independent, work-specific indicators strengthens attribution; by contrast, high stylistic and semantic similarity combined with low work-specific correspondence and strong hard negatives may support style imitation or independent convergence. The framework therefore uses the aggregate score to organize evidence, not to replace the evidentiary explanation.

### 4.2. Source Attribution, Confidence, and Alternatives

Single-source attribution is appropriate when one work accounts for most of the distinctive features preserved in the output. The report should identify the scope of attribution, because a source may explain the whole image, only the central figure, or a particular ornamental passage. Multi-source attribution becomes necessary when different works explain different components. A model that forces every disputed output into a single-source category will either select the visually dominant work and ignore other contributions or distribute low-level similarity so broadly that no coherent account emerges.

An attribution graph offers a useful representation. The generated image forms the central node, candidate sources form surrounding nodes, and each edge identifies the region or feature attributed to a source. A content source may differ from a style source, and a source cluster may be used when several variants or related works are indistinguishable on the available evidence. The model should prevent overfragmentation. Region-level attribution should therefore require minimum size, distinctiveness, consistency, and causal plausibility. Neighboring matches should form coherent visual or structural units, and the combined account should explain the output better than a null or common-source hypothesis. Human review is especially important when the graph includes numerous weak edges.

Null-source attribution is equally important. The null outcome may reflect independent convergence, common style, incomplete corpus coverage, or genuinely indeterminate evidence. In a system intended to support public or institutional accusations, the ability to withhold attribution is a central safeguard against the misleading certainty created by rank ordering. Confidence in attribution is not unitary. Retrieval confidence concerns whether the relevant source has been identified within the available reference corpus. Dependence confidence concerns whether the observed correspondence resulted from source influence rather than convention, shared references, or coincidence. Classification confidence concerns whether established dependence satisfies the criteria for the selected plagiarism category.

The framework should report these confidence dimensions separately. Retrieval confidence may be low because the reference corpus is incomplete even when correspondence with the leading candidate is strong; dependence confidence may be low because access or causal evidence is missing despite reliable visual matches; and classification confidence may remain limited when transformation, acknowledgment, authorization, or context is uncertain. Alternative explanations must be tested affirmatively. Each alternative should be connected to observable predictions. If a shared public-domain source explains two works, that third source should account for their overlap. If metadata is inconsistent with the visual record, its reliability should be reduced rather than mechanically accepted. Calibration converts confidence from rhetoric into an empirical property. Source-absent conditions are especially important because retrieval systems often remain highly confident even when the true source is not in the database. Calibration should therefore be performed

separately for source-present and source-absent settings and, where necessary, for different model families. The final report should state both the best-supported account and the principal alternative. For example, it may conclude that Artwork A is the leading source because of rare compositional correspondence and verified reference use, while noting that a shared public-domain photograph explains part of the overlap.

Confidence should also be sensitive to evidentiary dependence among indicators. Treating correlated evidence as independent will inflate confidence. The framework should identify evidentiary families—visual, structural, causal, provenance, and contextual—and give greater weight to convergence across families than to repetition within one family. The objective is not mathematical sophistication for its own sake; it is to prevent apparent quantity of evidence from being mistaken for diversity and strength.

## 5. Evidentiary Assessment of AI-Art Plagiarism

### 5.1. Technical, Visual, and Provenance Evidence

Technical and visual evidence varies in function and limitation. Pixel correspondence is strongest for direct copying but is easily defeated by transformation. Perceptual correspondence detects preserved visual structure across modest changes, yet may overweight texture or local appearance. Semantic correspondence identifies common concepts and objects, but can confuse shared subject matter with copying. Reproduction testing provides powerful causal support, yet requires access to the relevant model and generation conditions.

The evidentiary value of these measures arises through convergence. A moderate semantic match may be weak in isolation, but become persuasive when it coincides with an unusual composition, matching local motifs, a simple geometric alignment, and evidence that the source was supplied as a reference. The framework should record the direction of evidence rather than merely add positive scores; some findings strengthen the source hypothesis, while others support an innocent alternative.

Technical evidence should also be reproducible. Where learned weights or classifiers are employed, the calibration data and validation strategy should be documented. Without this information, an attribution cannot be meaningfully contested or replicated by another evaluator. Proprietary systems that disclose only a final percentage are unsuitable for high-stakes accusations. Visual explanations require human interpretation. Reviewers should examine whether the highlighted correspondence relates to the alleged source in an intelligible way and whether the same feature appears in hard negatives. The appropriate relationship is therefore cooperative: computation screens, aligns, compares, and quantifies, while human review assesses expressive significance and contextual meaning.

Technical evidence must finally be assessed against the possibility of manipulation. Robustness testing should examine how attribution changes under ordinary and adversarial transformations. The objective is not to make the system invulnerable, but to disclose sensitivity and avoid presenting fragile results as established facts. Visual correspondence becomes substantially more persuasive when connected to the generation process. Relevant evidence includes dataset membership and duplication, fine-tuning records, reference-image inputs, prompts, seeds, control images, masks, checkpoints, intermediate outputs, and editing history. These records help distinguish memorized examples from supplied references and show whether source-dependent features entered during generation or later editing.

The strongest evidence consists of verified generation and reference-image records that directly connect a source to the disputed output. Verified membership in a training or fine-tuning set provides a weaker but still important link, especially when combined with duplication and reproducible model behavior. Visual inference alone remains necessary when records are unavailable, but it occupies the lowest position in the causal hierarchy because it cannot by itself identify the route of dependence. C2PA Content Credentials illustrate both the promise and the limit of provenance systems. They can preserve cryptographically verifiable statements concerning an asset's origin, editing history, and use of AI, thereby reducing uncertainty about the technical chain. Provenance evidence must therefore be assessed for issuer, scope, completeness, and consistency with the visual record [12].

Missing provenance should not be treated as evidence of wrongdoing. An adverse inference is justified only when a party had a defined obligation to preserve records or when deliberate removal is supported by additional evidence. Recent work on causal attribution, watermarking, and model fingerprinting may supplement these records. These approaches address different attribution targets and should not be conflated: model or user attribution identifies who or what produced an image, while source-work attribution identifies the prior expression on which the image depends [13].

### 5.2. Human Intervention, Context, and Evidentiary Confidence

Human intervention affects both causation and evaluation. A person may select a source image, construct prompts, use inpainting or masks, choose among numerous generations, composite elements, draw over the output, alter color and lighting, crop or retouch the image, and decide how it is described and published. The framework should reconstruct the workflow rather than classify the final image as though it emerged from an autonomous and uniform process. The causal inquiry asks whether the person introduced or preserved the relevant source-dependent features. A user who uploads a reference image or repeatedly steers the system toward a recognizable composition occupies a different evidentiary position from a user who receives an unexpected output from a general prompt. Intent is not required to establish dependence, but it can affect ethical judgment, represented authorship, and allocation of responsibility.

The transformative inquiry asks whether human intervention materially reorganized the source-dependent expression. Recoloring, filtering, and minor object substitution may require effort while leaving the expressive identity of the source intact. The evidence matrix should therefore focus on the effect of intervention on the borrowed features rather than treating human effort as an automatic defense. Human authorship and plagiarism must not be conflated. A person may contribute protectable or otherwise meaningful expression through prompts, selection, arrangement, and post-generation modification while still retaining source-dependent elements from another work. The presence of human authorship in some aspects does not resolve the attribution of other aspects [14]. Acknowledgment and authorization complete the contextual inquiry. Accurate attribution may make a quotation, homage, or study non-deceptive, although it does not necessarily satisfy copyright law or licensing terms. The framework should report these factors separately rather than allowing one to erase the others. The framework should not borrow legal burdens of proof without qualification. Five levels provide a useful scale: insufficient evidence, possible source dependence, probable source dependence, strong source dependence, and directly verified source dependence. The terms describe support for the causal attribution and should be distinguished from the later normative label of plagiarism.

Insufficient evidence applies when no source is reliably identified or when the shared features are generic. Possible source dependence applies when a plausible source and meaningful correspondence are identified, but causal support remains limited or substantial alternatives remain. Probable source dependence requires a leading source, distinctive correspondence, and convergent causal support that outweighs alternatives. Strong source dependence applies when multiple independent evidentiary families converge on the same source and the principal alternatives are weak, although direct verification remains unavailable. Directly verified source dependence is reserved for authenticated reference-image use, generation records, or reproducible model behavior that connects the source to the output.

Every conclusion should identify the evidence supporting it, evidence weighing against it, missing information, and the principal alternative explanations. The classification must also disclose whether uncertainty arises from corpus coverage, metric reliability, causal access, or normative context. The system should never state merely that an image is a certain percentage plagiarized. A defensible report instead identifies the leading source, the distinctive regions and relationships that correspond, the provenance or access evidence, the performance of hard negatives, and the remaining alternatives. It may then state that the source dependence is probable with high evidentiary confidence and that the contextual factors support a preliminary plagiarism classification. Indeterminacy is a legitimate outcome. Closed training data, missing provenance, unknown sources, contradictory records, and culturally variable plagiarism norms may prevent a stable conclusion. The purpose of evidentiary confidence is to discipline judgment, not to conceal uncertainty behind numerical precision.

## 6. Plagiarism Classification Model

### 6.1. Evidence Matrix and Decision Thresholds

The plagiarism evidence matrix organizes the transition from source dependence to evaluative classification. It contains seven dimensions: source identifiability, expressive correspondence, distinctiveness, causal evidence, transformation, attribution and authorization, and alternative explanations. The matrix is therefore a structure for disciplined judgment rather than a substitute for judgment. Source identifiability asks whether one or more prior works can be reliably named. Expressive correspondence records the shared elements and whether the relationship concerns the whole image, composition, region, motif, or style. Distinctiveness assesses whether those elements are conventional or source-specific, while causal evidence examines access, training influence, reference use, reproduction, and workflow records. Together, these dimensions establish whether the alleged source materially explains the output. Transformation, attribution, and authorization address the normative context. Transformation asks how substantially the source-dependent expression has

been altered in organization, relationship, and communicative function. The final dimension requires the evaluator to state whether common convention, shared source, generic prompting, convergence, or incomplete corpus coverage provides a credible alternative account.

A zero-to-four scale may be used to assist annotation, but the values are ordinal judgments rather than natural measurements. A score of four for source identifiability might represent a verified source, while a score of three represents a clearly leading candidate and a score of two several plausible candidates. The numerical values must be accompanied by evidence and should not be summed mechanically; in particular, a high aggregate score cannot compensate for failure to satisfy a conjunctive gatekeeping requirement. The matrix improves consistency by forcing evaluators to confront negative evidence and missing information. One reviewer may accept source identifiability but disagree about distinctiveness; another may accept dependence but assign greater transformative weight. Such disagreement is more useful than a dispute over a single final score because it reveals the conceptual and evidentiary source of divergence. A two-stage decision structure prevents stylistic or semantic resemblance from being converted directly into plagiarism. The image must satisfy minimum requirements for source identifiability, distinctive expressive correspondence, and causal support. No amount of artist-level resemblance should compensate for the absence of a plausible source work. The thresholds at the first gate should be conjunctive rather than purely additive. A verified source with only generic correspondence does not establish appropriation, and a highly distinctive match without plausible access may remain coincidence or a shared-source case. The evaluator must be able to explain how the source became relevant to the output and why the matching features carry identifying force.

The second gate evaluates plagiarism after dependence has been established. It considers the degree of appropriation, the transformation of the source-dependent features, acknowledgment, authorization, the manner in which authorship is represented, and the strength of remaining alternatives. A disclosed study, licensed adaptation, deceptive substitution, and critical collage may preserve similar source features while differing substantially in plagiarism treatment. Accordingly, failure at the first gate terminates the plagiarism inquiry without an adverse classification, whereas passing the first gate does not predetermine the second. This asymmetry is essential: source dependence is a necessary premise for work-level plagiarism in this framework, but it is not itself sufficient for plagiarism. Thresholds should be calibrated through expert annotation and empirical validation rather than selected solely by the author. Where public accusation or automatic removal is contemplated, the system should privilege precision and require stronger evidence than it would for internal screening or research triage. The two-gate structure also supports review. A creator can contest source identification, demonstrate that the shared elements are conventional, produce provenance showing an independent workflow, or explain transformation and authorization. Contestability is thus built into the classification architecture rather than added after an adverse decision.

### 6.2. Classification Categories

The final reporting categories deliberately combine evidentiary descriptions of source dependence with normative plagiarism classifications. They should therefore not be interpreted as points on a single linear severity scale: a verified replication finding establishes a causal-reproductive relationship, whereas a plagiarism label additionally requires contextual evaluation of transformation, acknowledgment, authorization, and represented authorship. The strongest evidentiary category of source dependence is verified replication. It applies when near-exact or materially preserved source expression is combined with direct or exceptionally strong source evidence and no credible alternative explanation. Verified replication states what the evidence can establish: the output reproduces an identifiable work through a supported causal pathway.

Probable plagiarism applies when an identifiable source accounts for substantial and distinctive expression, the evidence of dependence is strong, and acknowledgment, authorization, or transformation is insufficient to displace the adverse inference. The report should therefore state the basis of probability and invite review rather than presenting the classification as an incontestable fact. Possible appropriation applies when the correspondence is meaningful and source dependence is plausible, but causal or provenance evidence remains incomplete or competing explanations remain viable. It should not be used for public accusation without additional evidence, because the distinction between a plausible hypothesis and a supported finding is precisely what the framework is designed to preserve.

The categories should not be ordered solely by a total score. A work may have extremely high similarity but remain only possible appropriation if the source is uncertain and hard negatives are equally strong. Classification depends on the configuration and quality of evidence, not simply its numerical sum. Reports

should use language proportionate to the result. Verified replication may justify a direct statement that a source was reproduced. Probable plagiarism should be expressed as a preliminary evaluative conclusion supported by specified evidence. Style imitation is assigned when artist- or school-level resemblance is strong but work-specific correspondence is weak and no reliable source work is identified. The system may describe the stylistic features involved and note ethical or market concerns, but it should not characterize the output as source-work plagiarism.

Independent convergence is assigned when the similarity is better explained by common subject matter, convention, prompt, shared public-domain source, or overlapping visual distributions than by dependence on the alleged work. Convergence does not require proof that no influence occurred; it means that source dependence has not been established as the best explanation. Indeterminate cases arise when evidence is conflicting, incomplete, inaccessible, or unreliable. The true source may be absent from the corpus, provenance may be missing or manipulated, and visual evidence may point in different directions. The report should specify what additional information could change the conclusion. These non-adverse categories are essential to system validity. Style imitation, convergence, source absence, and indeterminacy must be represented in training and testing so that the system learns when not to accuse. The false-attribution rate should be treated as a primary performance measure, not a secondary inconvenience.

## 7. Empirical Validation Framework

### 7.1. Dataset Design and Evaluation Metrics

Empirical validation of the proposed framework should use a dataset large enough to represent the principal forms of dependence and the principal innocent alternatives. The distribution should avoid allowing one easily detected category to dominate aggregate accuracy. Because the present treatise proposes an attribution architecture rather than reporting a completed benchmark, the following sections specify a validation protocol. Their purpose is to make the framework falsifiable by defining in advance the datasets, negative controls, metrics, calibration tests, ablations, and case analyses by which its claims should be assessed. Source materials should span multiple artistic genres and include both representational and abstract works. Public-domain and licensed works provide a lawful basis for controlled generation, while naturally occurring disputes can be included only with careful documentation and ethical safeguards. Closed proprietary models may be included as black-box conditions, but their evidentiary limitations should be stated.

A validation study should distinguish source-present and source-absent retrieval. In the source-present condition, the true source is included in the reference corpus and retrieval performance can be measured directly. Hard negatives should include works by the same artist, works in the same style or genre, works depicting the same subject, works with similar compositions, and images generated from similar prompts. Data splitting must occur by artist and source artwork rather than merely by generated image. If variations derived from the same source appear in training and test sets, the model may memorize source-specific features and produce inflated results. The design should also ensure that source clusters and near-duplicate reference works do not leak across splits.

Annotation requires a documented protocol. Experts should identify the source or sources, mark relevant regions, classify the form of dependence, score the evidence-matrix dimensions, and record alternative explanations. Research on plagiarized-painting recognition shows why this separation matters: detecting that a painting is suspicious and retrieving the authentic source are distinct tasks with different error profiles [3]. Controlled and naturalistic data should be analyzed separately. Combining them without distinction can produce misleading performance estimates. The controlled set should test whether the framework detects known forms of dependence under specified transformations, while the naturalistic set should test whether it communicates uncertainty and handles incomplete records responsibly. Results should be reported for each set before any aggregate measure is presented.

Retrieval performance should be evaluated through Recall at 1, 5, and 10, mean reciprocal rank, mean average precision, and source-cluster recall. Region-to-source assignment accuracy is necessary for multi-source cases, because a system may retrieve all relevant sources while attributing the wrong regions to them. Classification performance should include macro-averaged F1, per-class precision and recall, area under the receiver-operating-characteristic curve where appropriate, and confusion matrices. The false-attribution rate should be reported prominently, especially in source-absent and style-imitation conditions, because a system that frequently names an innocent source is unsuitable for reputationally sensitive use. Calibration should be evaluated through Brier scores, expected calibration error, and reliability plots. Calibration should be examined separately for retrieval, dependence, and plagiarism classification, and should be stratified by model

family and dependence type. Confidence estimates should also be tested after distribution shift to unseen generators.

Error analysis should focus on the failures most consequential to the framework. False plagiarism findings in style-imitation cases reveal inadequate separation of artist-level and work-level evidence. Each error should be assigned to the retrieval, correspondence, dependence, or classification stage so that remedies are technically and conceptually targeted. Performance should be compared with expert judgment rather than treated as self-validating. The evaluation should therefore report machine-expert agreement separately for source retrieval, feature correspondence, dependence confidence, and normative classification.

### 7.2. Cross-Model Evaluation, Case Studies, and Error Analysis

Cross-model evaluation should determine whether the attribution framework is portable or merely tuned to one generation architecture. GANs can reproduce training examples or mode-specific structures through mechanisms that differ from diffusion memorization, while diffusion systems may preserve local details, compositions, or textural artifacts under prompt and conditioning regimes. The same correspondence features and thresholds may therefore behave differently across families. The proposed evaluation should test whether thresholds transfer, whether feature weights remain stable, whether unseen generators produce degraded retrieval or calibration, and whether model-specific recalibration is necessary. Additional black-box testing can show whether the system remains useful when checkpoints, training data, or internal activations are unavailable.

The author's earlier portable plagiarism-severity tool should serve as the primary baseline. The comparison isolates the value added by source retrieval, local and compositional analysis, provenance, causal evidence, hard negatives, and alternative-hypothesis testing. An improvement in classification accuracy alone is insufficient; the complete model should also reduce false attribution and improve calibration. Ablation studies should remove one evidentiary component at a time. Removing provenance tests the marginal value of causal records, while removing local analysis reveals its importance for fragmentary appropriation. These ablation studies would connect the theoretical claims of the treatise to measurable system behavior.

Model-specific recalibration should not be treated as a failure if it is disclosed and empirically justified. The goal is not to impose identical thresholds on systems with different visual and memorization properties, but to determine which parts of the framework generalize and which require adaptation. Quantitative evaluation should be supplemented by six to eight detailed case studies. At minimum, the cases should include a near-exact reproduction correctly attributed, a recolored or cropped reproduction, a compositional appropriation with low pixel similarity, a multi-source synthesis, a style imitation falsely flagged by a similarity baseline, an independent-convergence case, a case in which the true source is absent from the corpus, and a case involving manipulated or contradictory provenance.

Each case should present the disputed image, top-ranked source candidates, metric-level scores, corresponding regions, transformation analysis, provenance evidence, alternative explanations, and final classification. This format demonstrates how the framework reasons and allows readers to examine whether the conclusion follows from the evidence. The near-exact and transformed-replication cases should demonstrate the value of alignment and causal records. The multi-source case should demonstrate region-to-source attribution and explain why a single-source baseline fails. The style-imitation and convergence cases should show the safeguards against over-detection.

The source-absent case is especially important. The system should lower retrieval confidence, recognize a flat or inconsistent candidate distribution, and return a null or indeterminate result. The manipulated-provenance case should test whether the framework compares metadata with visual and technical evidence rather than accepting credentials uncritically. Error analysis should explain why the system failed and what form of remedy is appropriate. Dependence failures may reflect missing causal evidence or weak alternative testing. This stage would ensure that the validation framework evaluates the architecture of reasoning rather than merely displaying a table of incorrect predictions.

## 8. Legal and Ethical Implications

### 8.1. Plagiarism and Copyright Infringement

The framework classifies source dependence and plagiarism; it does not automatically determine copyright infringement. A copyright claim may require ownership, copying, similarity in protectable expression, and analysis of limitations, exceptions, and defenses. These requirements vary by jurisdiction and cannot be

encoded fully in a general visual-attribution score. The distinction is particularly important in generative AI. The U.S. Copyright Office has emphasized that memorization becomes legally significant when substantial protectable expression is retained, while the fact of training use alone does not resolve every output- or model-based claim [10].

Transformation likewise has different meanings across ethical and legal analysis. The Supreme Court's decision in Andy Warhol Foundation v. Goldsmith illustrates that alteration and a claimed new meaning do not end the inquiry; the purpose and context of the specific use remain important [15]. The framework can nevertheless support legal analysis by identifying candidate sources, locating corresponding expression, distinguishing conventional from distinctive elements, and documenting causal evidence. Courts, institutions, and parties must apply the governing legal standards independently. The same caution applies to authorship. Copyright guidance distinguishes human-authored selection, arrangement, and modification from material generated without sufficient human control. Those questions do not determine whether preserved elements are attributable to another source. A work may contain human-authored contributions while still retaining source-dependent expression requiring acknowledgment or authorization [14].

### 8.2. Responsibility and Risks of Over-Detection

Responsibility should be differentiated according to control, knowledge, and the conduct that created or preserved the disputed dependence. Model developers control dataset governance, deduplication, memorization testing, training documentation, and many mitigation measures. Platforms control provenance preservation, complaint procedures, disclosure, contestability, and the consequences of automated classifications. Evaluators control thresholds, explanations, bias testing, and human review. A user should not be presumed to know which training image a model may unexpectedly reproduce. Where the output emerges from a general prompt and no source was supplied, responsibility may lie principally in model behavior and the subsequent response after notice. The framework's reconstruction of the causal pathway helps make this distinction. Developers have different responsibilities, which may include dataset documentation, duplicate detection, targeted memorization testing, and mechanisms for responding to verified replication. Research on locating or deleting memorized-image subspaces and retrieval-augmented protection suggests possible mitigation. These techniques remain developing and should be evaluated for effectiveness, side effects, and transparency rather than treated as complete solutions [16]. Platforms should resist automatic removal based solely on a similarity score. They can preserve provenance, provide complainants and creators with the source candidates and relevant regions, permit submission of sketches and generation records, and use graduated responses based on evidentiary confidence. The process should include notice, an opportunity to contest, and review by a person capable of understanding both technical evidence and artistic context.

Evaluators bear responsibility for the language of their conclusions. Reports should identify limitations, disclose database coverage, separate technical dependence from normative judgment, and avoid implying intent where the evidence establishes only resemblance or causal contribution. Responsible attribution is as much a practice of calibrated communication as a practice of measurement. False attribution can damage the reputation of an artist, user, model developer, or platform. Because established artists are more likely to have extensive digitized corpora, a retrieval system may preferentially identify their works and overlook marginalized, anonymous, or non-Western sources. Over-detection is especially likely when the model is trained on positive plagiarism examples without adequate style, convergence, and source-absent controls. Precision, open-set recognition, and calibration must therefore be central design objectives, particularly when results are used publicly. False negatives also matter. A system that protects against accusation by demanding near-exact similarity may miss compositional or fragmentary appropriation. Broad retrieval can remain sensitive, while adverse classification requires stronger source-identifying and causal evidence. This separation permits investigation without premature condemnation. Bias testing should examine performance across genres, cultures, artistic media, levels of digitization, and artist prominence. Hard negatives should include culturally related works and shared traditional motifs so that the model does not mistake collective visual heritage for individual ownership. Where bias cannot be corrected, the limitation should be disclosed and the system's use narrowed. The ethical presumption should be against irreversible action on the basis of weak or opaque evidence. Public labeling, removal, or sanction should require stronger evidence and a meaningful opportunity to contest. This proportionality follows directly from the inferential gap at the center of the treatise.

### 8.3. Transparency and Contestability

Every adverse classification should disclose the candidate sources, relevant image regions, metrics and transformations used, thresholds, provenance assumptions, missing data, alternative explanations, and

confidence level. Without these disclosures, the creator cannot understand or challenge the attribution. A creator should be able to submit original sketches, earlier versions, prompt records, generation histories, independent references, licensing evidence, and proof that metadata is incomplete or misleading. The system should update the conclusion rather than treating its initial output as fixed.

Contestability also requires access to meaningful reasons. Visual overlays, candidate comparisons, and written explanations should show why a feature was considered distinctive and why alternatives were rejected. Where proprietary constraints prevent disclosure of the model, the institution should at least disclose the evidence and decision logic sufficient for an independent expert to evaluate the result. Procedural transparency should include versioning and audit trails. Changes to the reference corpus, model weights, thresholds, or calibration can alter attribution. Appeals should be reviewed by someone other than the initial evaluator when practicable, especially in institutional or platform settings.

Explainability is therefore not merely a desirable interface feature. The central contribution of the framework lies not in producing more accusations, but in making the path from similarity to attribution visible, testable, and reversible. Institutional use should be governed by a documented policy that links evidentiary levels to permissible actions. Possible source dependence may justify a request for process records, while probable dependence may support temporary labeling or a negotiated response. Public accusation, removal, or sanction should require the strongest evidence, independent human review, and an appeal mechanism. In this way, the technical framework becomes part of a proportionate governance process rather than an isolated scoring instrument.

## 9. Conclusion

This treatise has argued that similarity and plagiarism are analytically distinct. The necessary bridge is source dependence: a supported inference that an identifiable prior work materially contributed to the distinctive expression of a generated image beyond what can reasonably be explained by conventions, general training, shared sources, or independent convergence. Reliable attribution requires the convergence of visual, causal, provenance, and contextual evidence. Candidate retrieval, multi-level correspondence, distinctiveness analysis, hard-negative comparison, provenance, alternative-hypothesis testing, and calibrated confidence must operate together. The framework must also permit multi-source, source-cluster, null-source, and indeterminate outcomes.

The theoretical contribution is the separation of three stages that are too often collapsed: correspondence, source dependence, and plagiarism classification. It also distinguishes source-work attribution from training-data and generator attribution, thereby preventing evidence directed to one question from being treated as proof of another. The practical contribution is an explainable system architecture that specifies how candidate sources can be retrieved, relevant regions and relationships identified, provenance and generation evidence incorporated, alternative explanations tested, and calibrated classifications produced. Cross-model evaluation and source-absent testing provide a basis for assessing whether the system is portable and safe enough for its intended use.

The framework remains limited by incomplete training-data access, closed models, missing provenance, unknown sources, cultural variation in plagiarism norms, database bias, difficulty measuring transformation, diffuse multi-source influence, adversarial manipulation, and changing architectures. These limitations define the circumstances in which a conclusion should be weakened or withheld. Future research should extend the framework to video, music, literary and mixed-media works, where temporal and cross-modal dependence create additional attribution problems. It should also develop better open-set retrieval, culturally representative corpora, privacy-preserving provenance, and procedures for real-time platform dispute resolution. The central proposition nevertheless remains stable. Similarity begins the inquiry, source dependence gives it structure, and contextual evaluation determines its conclusion. Any system that skips one of these stages risks replacing analysis with accusation.

## Declarations

**Acknowledgments:** The author would like to acknowledge the independent nature of this research, which was conducted without institutional or external support.

**Artificial Intelligence (AI) Use Statement:** The author used ChatGPT (OpenAI) for language editing, proofreading, and improving the clarity of the manuscript. The author independently reviewed and verified the resulting text and retains full responsibility for the scholarly content, analysis, citations, interpretations, and final manuscript.

**Author Contribution:** Byungkil Choi is the sole author of this manuscript and was responsible for the conceptualization, literature review, analysis and interpretation, preparation of the original draft, review and editing, and final approval of the manuscript.
**Conflict of Interest:** The author declares no conflict of interest.
**Consent to Publish:** The author agrees to publish the paper in International Journal of Recent Innovations in Academic Research.
**Data Availability Statement:** No new datasets were generated or analyzed in this study. All materials supporting the analysis are contained in the article and the cited sources.
**Funding:** This research received no external funding.
**Institutional Review Board Statement:** Not applicable. This study did not involve human participants or animals and therefore did not require institutional review board approval.
**Informed Consent Statement:** Not applicable. This study did not involve human participants.
**Research Content:** The research content of the manuscript is original and has not been published elsewhere.

**References**

1. Fu, S., Tamir, N.Y., Sundaram, S., Chai, L., Zhang, R., Dekel, T. and Isola, P. 2023. DreamSim: Learning new dimensions of human visual similarity using synthetic data. Advances in Neural Information Processing Systems, 36: 50742–50768.
2. Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D. and Wallace, E. 2023. Extracting training data from diffusion models. 32nd USENIX Security Symposium (USENIX Security 23): 5253–5270.
3. Zhou, S. and Kong, S. 2025. Dare to plagiarize? Plagiarized painting recognition and retrieval. 2025 IEEE International Conference on Advanced Visual and Signal-Based Systems (AVSS), 1–4.
4. Choi, B. 2026. Plagiarism detection in GAN-generated abstract art: A multi-modal semantic and compositional approach. American Journal of Art and Design, 11(2): 39–53.
5. Roig, M. 2015. Avoiding plagiarism, self-plagiarism, and other questionable writing practices: A guide to ethical writing. Office of Research Integrity, U.S. Department of Health and Human Services.
6. Wang, S.Y., Efros, A.A., Zhu, J.Y. and Zhang, R. 2023, October. Evaluating data attribution for text-to-image models. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 7158-7169). IEEE.
7. Zhang, R., Isola, P., Efros, A.A., Shechtman, E. and Wang, O. 2018, June. The unreasonable effectiveness of deep features as a perceptual metric. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 586-595). IEEE.
8. Wang, S.Y., Hertzmann, A., Efros, A.A., Zhu, J.Y. and Zhang, R. 2024. Data attribution for text-to-image models by unlearning synthesized images. Advances in Neural Information Processing Systems, 37: 4235-4266.
9. Wang, W., Sun, Y., Yang, Z., Hu, Z., Tan, Z. and Yang, Y. 2026. Replication in visual diffusion models: A survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2026.3713990
10. U.S. Copyright Office. 2025. Copyright and artificial intelligence, Part 3: Generative AI training (pre-publication version). U.S. Copyright Office.
11. Oquab, M., et al. 2023. DINOv2: Learning robust visual features without supervision. arXiv. https://doi.org/10.48550/arXiv.2304.07193
12. Coalition for Content Provenance and Authenticity. 2025. Content credentials: C2PA technical specification (Version 2.2).
13. Asnani, V., Collomosse, J., Bui, T., Liu, X. and Agarwal, S. 2024, June. Promark: Proactive diffusion watermarking for causal attribution. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 10802-10811). IEEE.
14. U.S. Copyright Office. 2025. Copyright and artificial intelligence, Part 2: Copyrightability. U.S. Copyright Office.
15. Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508 (2023).
16. Chavhan, R., Bohdal, O., Zong, Y., Li, D. and Hospedales, T. 2024. Memorized images in diffusion models share a subspace that can be located and deleted. arXiv. https://doi.org/10.48550/arXiv.2406.18566

**Citation:** Byungkil Choi. 2026. From Similarity to Plagiarism: Source Attribution in AI-Generated Art. International Journal of Recent Innovations in Academic Research, 10(3): 66-81.