MEDWIN PUBLISHERS
Committed to Create Value for Researchers

Philosophy International Journal
ISSN: 2641-9130

A Portable Plagiarism-Severity Tool for AI Paintings: Cross-Model Scoring for GAN and Diffusion Replicas

Choi B*

Department of Arts, College of Arts and Design, Wonkwang University, Republic of Korea

***Corresponding author:** Byungkil Choi, Department of Arts, College of Arts and Design, Wonkwang University, Iksan, Republic of Korea

Conceptual Article
Volume 9 Issue 3
Received Date: July 27, 2026
Published Date: September 01, 2026
DOI: 10.23880/phij-16000374

Abstract

Disputes over AI-generated artworks increasingly depend on visual comparisons between an asserted original and candidate output, yet existing technical approaches provide limited guidance on how to measure the degree of copying in a way that is transparent, reproducible, and usable in real review settings. The primary objective of this study is to deliver a portable, user-facing software tool that helps practitioners estimate a plagiarism ratio/severity score between an original artwork and suspected duplicates or AI-generated replicas. Rather than issuing binary determinations of infringement, the paper develops a software-oriented framework for graded, evidence-based severity assessment that can be deployed for screening, expert review, and pre-dispute analysis.

We introduce a Plagiarism Degree Index (PDI, 0–100) that integrates multiple similarity signals—compositional structure, perceptual texture, and semantic motifs—and calibrates them against human-grounded severity judgments. Experiments across both GAN-based and diffusion-based generation pipelines show that severity scores increase monotonically with controlled copying pressure and remain broadly comparable across model families, while also revealing systematic failure modes such as style-only or motif-only resemblance. Importantly, the framework is implemented as a portable software pipeline that produces (i) a severity score with uncertainty indicators, (ii) ranked candidate lists, and (iii) compact, inspectable evidence reports that explain why similarity is detected. Although this study does not claim a definitive, universally correct plagiarism ratio in all contexts, it provides a practical and implementable foundation for building robust plagiarism-severity software that can be audited, versioned, and improved toward higher reliability in real-world use.

Building on two prior case studies, "Multi-metric plagiarism scale for AI paintings" (realist works) and "Plagiarism Detection in GAN-Generated Abstract Art: A Multi-Modal Semantic and Compositional Approach" (abstract works), this paper unifies the evidence logic into a single cross-model scoring protocol and packages it as a portable, auditable software kit (Appendix A).

Keywords: Plagiarism Degree Index (PDI); AI-generated art plagiarism; Cross-model similarity scoring; Multi-metric visual similarity; Evidence-based severity assessment; Auditable portable software tool

Introduction

Generative image models now make it routine to produce paintings that resemble existing works in composition, surface texture, palette, and recurring motifs, and this development has begun to influence how plagiarism and copyright disputes are argued in practice [1]. Recent judicial proceedings in the United Kingdom and the United States demonstrate that courts and litigants increasingly focus on concrete output comparisons, on what kinds of copying a system can plausibly produce, and on whether the asserted evidence is specific, testable, and reviewable rather than speculative [2]. At the same time, much of the technical literature continues to approach AI-image plagiarism either

CC BY

as a binary classification problem or through single similarity scores that are difficult to interpret, hard to audit, and unstable across different generative pipelines (e.g., SSIM-style structural measures [3], perceptual feature distances such as LPIPS [4], or embedding-based comparisons such as CLIP similarity) [5]. This mismatch is consequential because real-world evaluation—whether in litigation, expert review, or platform moderation—rarely turns on a simple yes-or-no determination. Instead, it proceeds as a reasoned inquiry into degree, asking how strongly a candidate output tracks an original's distinctive visual organization and how confidently that resemblance can be attributed to copying rather than shared subject matter or stylistic convention.

For this reason, the present study frames AI-image plagiarism as a problem of copying severity rather than binary labeling. Binary labels obscure meaningful differences between cases that may appear superficially similar yet raise fundamentally different concerns. A candidate image may preserve an original's compositional scaffold while altering surface texture, or it may borrow semantic motifs while remaining compositionally distinct; each pattern calls for a different evaluative response. A severity score becomes useful only when it supports such distinctions, rather than replacing them with an opaque verdict. From a measurement perspective, this implies that a deployable index should increase predictably as copying pressure increases, remain comparable across different model families, communicate uncertainty in borderline cases, and provide compact evidence outputs that allow reviewers to inspect why the score is high or low.

The task addressed in this paper follows the structure of many real plagiarism allegations, and it is motivated by a concrete engineering goal: to provide end users with a portable software program that can estimate a plagiarism ratio (or severity score) between an asserted original artwork and suspected duplicates. One asserted original painting is evaluated against one or more candidate replicas, and the goal is to produce a single severity score on a transparent scale from 0 to 100. This score is intended to support ranking and comparison—for example, identifying which candidates most closely track the original and which appear closer to independent creation—rather than to determine infringement itself. The scope is deliberately limited to AI-generated paintings, including both realist and abstract genres, because these foreground different sources of resemblance and different routes to copying. It is also deliberately limited to output-based visual evidence, since legal outcomes depend on additional facts such as access, authorization, and jurisdiction-specific defenses that cannot be inferred from pixels alone; accordingly, conservative scope limits and careful presentation are treated as part of responsible deployment.

This article is intended as the third step in a cumulative research line. It consolidates our two prior case studies—"Multi-metric plagiarism scale for AI paintings" (realist artworks) and "Plagiarism Detection in GAN-Generated Abstract Art: A Multi-Modal Semantic and Compositional Approach" (abstract artworks)—into a single cross-model scoring framework and a portable, reproducible software kit for evaluating GAN and diffusion replicas.

Accordingly, this paper adopts a court-informed perspective in which plagiarism-severity software is treated as a measurement instrument rather than an adjudicative device [6]. Drawing on recent AI-related copyright litigation and on multi-dimensional similarity research, the study develops and evaluates a Plagiarism Degree Index calibrated against human judgments of copying severity. The practical ambition is to move beyond a purely conceptual metric and toward a high-quality, portable plagiarism-checking program that can be used by non-specialists: given an original and a set of candidates, the software should output a severity score, provide uncertainty indicators and abstention guidance in borderline cases, and generate an evidence report that makes the result inspectable and contestable. While the proposed system does not yield a definitive or legally dispositive plagiarism ratio, it clarifies how such software should be designed and improved: by integrating multiple similarity dimensions, explicitly representing uncertainty and known failure modes, and producing evidence that is reproducible, inspectable, and communicable to non-technical decision-makers. The contribution of the paper therefore lies not in automating legal conclusions, but in offering practical guidance for building responsible, deployable plagiarism-severity assessment software that can support expert and legal review without exceeding its epistemic limits.

Court-Informed Framework: What Legal Disputes Demand from Portable Evidence Software

Unifying Realist and Abstract Evidence into Cross-Model Scoring

The key methodological move in this paper is to treat "realism vs. abstraction" not as two different plagiarism problems, but as two regimes of observable evidence. Realist replicas tend to expose copying through local texture, brushwork proxies, and structural alignment; abstract replicas often require a stronger emphasis on compositional scaffolds and motif-level semantics. A portable tool must therefore support a configurable, multi-signal score whose weights shift smoothly with genre cues—without changing the meaning of the final severity index.

Copyright© Choi B.

Practically, the kit operationalizes this unification by (i) standardizing preprocessing and transformation testing, (ii) computing a fixed panel of interpretable similarity channels, (iii) fusing channels through a transparent weighting and calibration step, and (iv) producing a compact, court-facing evidence report (metrics + aligned patches + uncertainty flags). Appendix A lists the complete kit contents (scripts, configuration, and example outputs) so that the protocol can be run by reviewers and reproduced by third parties.

When plagiarism disputes over AI-generated images reach litigation, the argument rarely turns on abstract similarity claims alone. Courts and litigants instead concentrate on (i) what can be demonstrated through concrete, reproducible outputs, (ii) what inferences can responsibly be drawn from those outputs, and (iii) whether the proffered evidence is specific enough to be tested rather than asserted. For software that aims to report a "plagiarism ratio" (or severity score) in a way that can actually be used by artists, experts, and decision-makers, this litigation posture translates into a clear engineering mandate: the tool must not only compute a number, but also package inspectable reasons for that number in a repeatable protocol.

A concise illustration comes from the UK Getty Images v. Stability AI litigation [7]. Occurrence: Getty alleged that Stability AI unlawfully used copyrighted images in training and produced outputs that appropriated Getty's protected material. Development: the case turned on technically framed questions about what, if anything, a trained model contains (as an "article" and/or an "infringing copy") and on what outputs visibly display under particular prompts and conditions. Judicial trial/decision: the court rejected the theory that the trained model itself constitutes an "infringing copy" where it never stored/contained a copy of the claimant's works, while separately analyzing liability where outputs reproduced or approximated Getty's marks/watermarks. In the U.S. visual-artist litigation Andersen v. Stability AI, the court's handling of claims at an early procedural stage likewise shows that parties must frame copying through identifiable works, concrete examples, and testable theories that can be evaluated through later factual development rather than through broad, uncheckable assertions [8]. In the consolidated U.S. OpenAI copyright litigation, the motion-to-dismiss posture underscores the same practical point: output examples and the plausibility of substantial similarity can become central to whether claims proceed, making "showable" comparisons (not merely model-level speculation) a recurrent evidentiary hinge [9].

Across these settings, a practical through-line emerges: disputes reward evidence that is reproducible, inspectable, and communicable to non-technical decision-makers, while discounting evidence that is opaque, purely statistical without explanation, or detached from observable comparisons. Accordingly, this paper treats plagiarism-severity software as an evidence instrument. The objective is to accept an asserted original and one or more candidates, then output (i) a calibrated severity score and (ii) a compact evidence bundle that can be independently re-run and reviewed—together with an explicit statement that the output is "measurement, not adjudication."

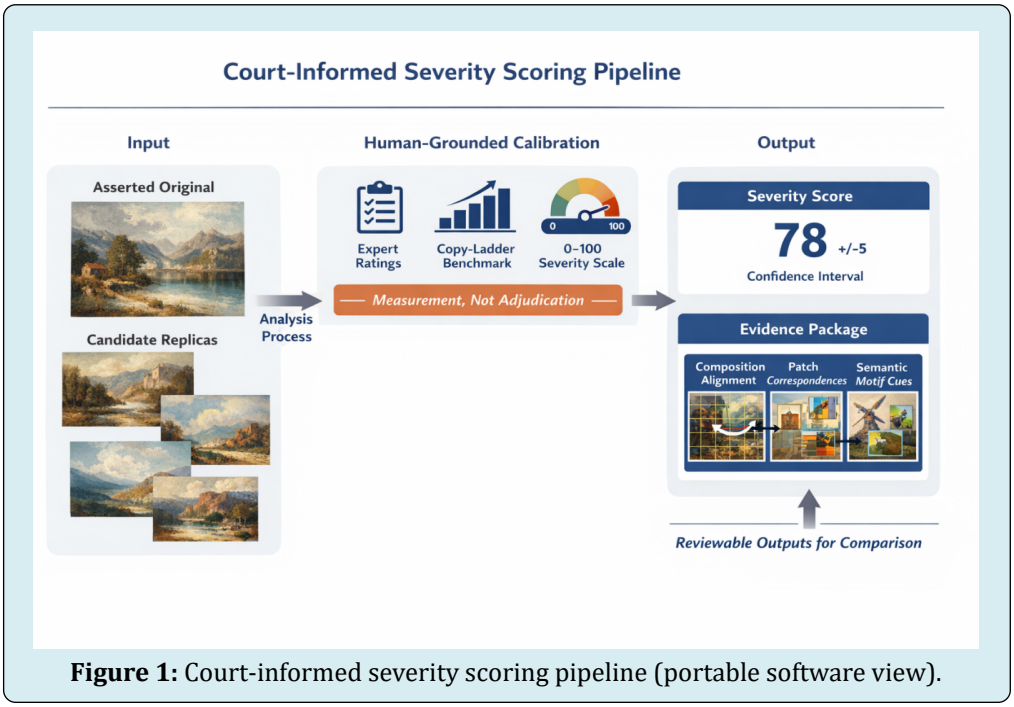

Figure 1: Court-informed severity scoring pipeline (portable software view).

Copyright© Choi B.

A schematic showing how an asserted original and multiple candidates enter a fixed protocol; how human-grounded calibration anchors a 0–100 scale; and how the tool exports (i) a severity score with uncertainty bounds and (ii) an evidence bundle (e.g., composition alignment, patch correspondences, motif cues) plus a run manifest (tool version, parameters, hashes), explicitly labeled as "measurement, not adjudication."

This court-informed stance yields concrete design requirements for a deployable "ratio/score" tool. First, the score must be anchored in a stable severity concept that aligns with how humans perceive incremental copying, because "copying severity" functions as a latent construct rather than a directly observed label. Diffusion-focused replication research demonstrates the value of graded replication annotation, but also highlights that the scale must remain meaningful when models, prompts, and pipelines vary. Second, cross-model comparability is essential: real disputes will not restrict attention to a single model family, and a score that changes meaning between GAN and diffusion replicas becomes rhetorically impressive but practically unusable. Third, interpretability must be part of the deliverable output—a score without an accompanying "why" is difficult to contest, difficult to trust, and difficult to present. Multi-aspect infringement detection work frames resemblance as multi-dimensional and underscores that legally salient overlap is rarely captured by one signal. Fourth, uncertainty must be explicit, with abstention guidance for borderline cases, because disputes concentrate exactly where similarity is arguable; distribution-free uncertainty methods (e.g., conformal prediction [10]) and classical resampling approaches (e.g., bootstrap intervals [11]) offer implementable ways to report uncertainty rather than to hide it. Fifth, robustness under routine transformations must be tested, because real-world circulation involves resizing, cropping, compression, and color shifts; reviewers and courts will care about what persists across such changes rather than what collapses under benign perturbations. To derive these tool-facing requirements, we coded recurring judicial concerns and litigation-facing evidentiary demands into a compact set of software-measurable outputs, yielding the mapping summarized in Table 1.

Judicial Concern (Practical)	What A Severity Tool Must Provide	What the Software Should Export
Reviewable, specific comparisons	Pairwise evidence tied to the specific original/candidate claim	Score + compact evidence bundle (alignment, patches, motif cues)
Reproducibility and auditability	Fixed protocol, parameter lock, versioning, hashable artifacts	Run manifest (version, settings, hashes) + exported report package
Borderline-case ambiguity	Uncertainty + abstention guidance	Confidence interval / prediction set + "gray-zone" rule
Multi-dimensional resemblance	Multiple signals, not one metric	Composition + texture + semantic signals with attribution to each
Cross-pipeline disputes	Comparable meaning across model families	Cross-model calibration checks (GAN + diffusion)

Table 1: From judicial concerns to software-measurable outputs (portable tool requirements).

At the same time, a court-informed framework requires a strict boundary between measurement and adjudication. The tool proposed here is designed to quantify visual copying severity on a 0–100 scale and to export reproducible evidence artifacts that can be inspected and debated, but it cannot decide infringement because legal outcomes depend on facts and doctrines not contained in pixels—access, authorization, licensing, defenses, and the expression/idea (or style) boundary. This limitation is not a weakness; it is a necessary constraint for responsible deployment. A severity instrument is most useful when it reduces search costs and improves consistency while making it easy to challenge the result—by letting others rerun the same protocol and inspect the same evidence.

These evidentiary commitments naturally translate into practical workflows for a portable plagiarism-assessment tool. In platform screening, the main value lies in triage: an asserted original is compared against large candidate sets; the tool ranks the most concerning items and flags uncertainty for human review. In committee/expert review, the value lies in standardization: multiple reviewers see the same score, the same uncertainty bounds, and the same compact evidence artifacts. In pre-dispute assessment, the value lies in early clarity: parties can evaluate portfolios under a reproducible protocol while also seeing documented failure modes (e.g., style-only resemblance, motif-only resemblance) that warrant caution.

Copyright© Choi B.

Workflow	Primary Goal	Recommended Mode	Output Emphasis
Platform screening	Triage large candidate sets	Batch mode	Ranking + uncertainty flags
Committee/expert review	Consistent deliberation	Forensic report mode	Evidence bundle + brief narrative summary
Pre-dispute assessment	Early case evaluation	Hybrid	Portfolio ranking + failure-mode notes

Table 2: Deployment modes by workflow (portable software).

Benchmark Construction and Human Grounding

This study treats copying severity as an empirical quantity that must be observed, stabilized, and calibrated through controlled comparisons, rather than inferred directly from an algorithm's internal confidence values. The benchmark is therefore constructed around an original–replica evaluation setting that mirrors how real plagiarism disputes and screening tasks are framed, while remaining practical for repeated, portable use across different model families and software runs. The human component is not introduced merely as a sanity check; instead, it is used to define and stabilize the semantics of a 0–100 severity scale, so that the output of the software behaves like a measurement instrument rather than a model-specific or dataset-specific score.

Original Paintings: Selection Criteria, Genre Coverage, Licensing/Ethics

The selection of original paintings directly constrains what a plagiarism-severity tool can and cannot meaningfully measure. Originals should therefore be chosen to maximize interpretability of severity judgments while minimizing confounding factors that would undermine portability or downstream credibility. The most defensible approach is to construct the original set from works that are clearly licensable for research use, with provenance recorded at the item level, and with the collection balanced across visual regimes that plausibly give rise to different routes to copying.

In practice, this means sampling across realist and abstract works, while also ensuring variation in compositional complexity, recurring motifs, palette constraints, and surface-texture patterns. The benchmark benefits from including: (i) originals that are distinctive primarily in global layout, (ii) originals that are distinctive in surface treatment or repeated local structures, and (iii) originals that are distinctive mainly through motif and semantic organization. Each category stresses a different failure mode of a severity-scoring tool and therefore improves its robustness when deployed beyond the benchmark.

Licensing and ethics are treated as first-class elements of the benchmark specification, not as an afterthought. Originals drawn from public-domain sources, museum open-access programs, or Creative Commons–licensed collections are accompanied by explicit records of rights status, source URLs, and attribution requirements. When modern works are included, either explicit permissions are obtained or the collection is restricted to works whose licensing terms unambiguously permit derivative generation and redistribution for research. A short dataset "card" documents the collection rationale, known limitations, and intended use boundaries, because courts and expert reviewers routinely scrutinize provenance when assessing the plausibility and legitimacy of downstream claims [12].

Id	Genre	Source/ Collection	Rights Status	Year Range	Compositional Complexity	Motif Category	Texture Category
O-001	Realist						
O-002	Abstract						
O-003	Realist						
O-004	Abstract						

Table 3: Original set specification (template)

Replica Generation Across Model Families

Replica generation is designed to separate the question "Can a system produce resemblance?" from the question "How severe is the copying relative to the original?" The benchmark therefore employs a controlled, shareable generation protocol in which copying pressure is manipulable, transparent, and logged. This protocol is aligned across model families as far as possible, so that severity comparisons do not merely reflect different prompting styles, tuning effort,

Copyright© Choi B.

or user skill. Importantly, the protocol supports generating multiple replicas per original, reflecting the portfolio-based nature of real disputes and screening workflows.

GAN replicas

GAN-based replicas are generated under at least two regimes. A base regime captures general stylistic competence using a pre-trained generator that has not been adapted to any specific original. A targeted regime increases copying pressure through controlled adaptation (for example, limited fine-tuning on a small artist-specific subset that includes the original), provided that the training protocol, data provenance, and adaptation intensity are fully documented [13]. The goal is not to privilege a particular architecture, but to produce replicas whose resemblance varies systematically and reproducibly, making severity judgments both human-tractable and software-measurable.

Diffusion replicas

Diffusion-based replicas likewise include base and targeted regimes. The base regime uses general-purpose text-to-image models with high-level prompts only. The targeted regime introduces additional conditioning—such as reference-image guidance, motif-encoding prompt expansion, or bounded fine-tuning—where ethically and legally permissible. Because diffusion systems are highly sensitive to prompts and random seeds, a fixed sampling strategy (number of prompts, seeds, and guidance settings) is defined and logged, preventing cherry-picked exemplars from dominating the benchmark [14]. This design is particularly important when the tool is later used to reason about tail risks, such as rare but severe near-copy events.

Copy-ladder design (severity control and negative controls

To make severity observable, the benchmark incorporates a copy ladder that explicitly ranges from weak influence to near duplication [15]. The ladder is implemented as a small, ordered set of levels that progressively increase copying pressure through controlled changes in the generation protocol. Although the operational details differ across GAN and diffusion pipelines, the ladder is conceptually aligned so that levels correspond to comparable kinds of reuse.

Negative controls are equally essential, because high similarity scores can arise for reasons unrelated to copying of protectable expression [16]. These include semantically similar but compositionally distinct works, style-matched but content-shifted works, and "false friends" where shared palette or texture patterns produce superficial resemblance. Such controls enable the software to detect and penalize style-only or motif-only confounds, improving its reliability when deployed by end users.

Level Name	Protocol Constraints (Conceptually Aligned Across GAN & Diffusion)	Expected Resemblance Type	Intended Failure Mode (What this Level Stress-Tests)
L0: Weak influence	No reference image; broad, high-level prompts only; no targeted adaptation/ fine-tuning; diverse seeds; generation settings logged and fixed per run.	Shared genre cues only (e.g., "portrait," "landscape," "abstract gesture"), but no stable reuse of distinctive layout or local structures.	False positives from generic similarity: the index should not score high just because two images share a genre, palette tendency, or common motif.
L1: Motif proximity	Prompts include motif descriptors (objects/ themes) but avoid specifying spatial layout; light conditioning allowed (e.g., constrained prompt templates), still no direct reference-image guidance and no strong targeted tuning.	Motif-level overlap and possibly palette/texture proximity, while composition remains meaningfully different.	Style/motif confounding: the index should not mistake "same subject/ motif" or "same style family" for compositional copying.
L2: Partial composition reuse	Moderate copying pressure: constrained prompts that encode partial spatial relations; or limited reference guidance / limited targeted adaptation (documented and capped); multiple candidates per original to avoid cherry-picking.	Local or partial layout reuse (e.g., repeated arrangement of key elements, repeated spatial grouping), with noticeable variations elsewhere.	Under-detection of structural reuse: the index must respond when structure repeats even if surface texture or color differs.

Copyright© Choi B.

L3: Strong composition match	High copying pressure: stronger reference guidance or stronger targeted adaptation (still documented and bounded); prompt constraints may include explicit layout descriptors; settings fixed and reproducible.	Global compositional alignment: overall spatial organization strongly matches the original, often with repeated local structures; surface details may vary.	Over-reliance on texture: the index should remain high even when brushwork-like texture changes, and should not collapse when the replica "repaints" the same structure.
L4: Near-duplicate	Maximum copying pressure within ethical/legal constraints: strongest allowed reference guidance or targeted adaptation; additional constraints to preserve arrangement; post-processing kept minimal and logged.	Near-duplication: strong structural match with minimal deviations; may approach patch-level correspondences.	Ceiling behavior & leakage: the index should saturate near the top and expose vulnerabilities (e.g., trivial duplicates, compression-only variants, watermark artifacts) without unstable swings.

Table 4: Copy-ladder level definitions.

Operational parameters (e.g., "reference strength" or "adaptation intensity") are recorded in the run log and selected to maintain consistent candidate counts across model families. Levels are defined by type of constraint, not by a specific model implementation, so the ladder remains portable across GAN and diffusion pipelines.

A schematic showing originals, model-family branches (GAN and diffusion), copy-ladder levels, and negative-control streams, yielding candidate sets per original for human evaluation, latent severity estimation, and calibration/validation of the 0–100 index.

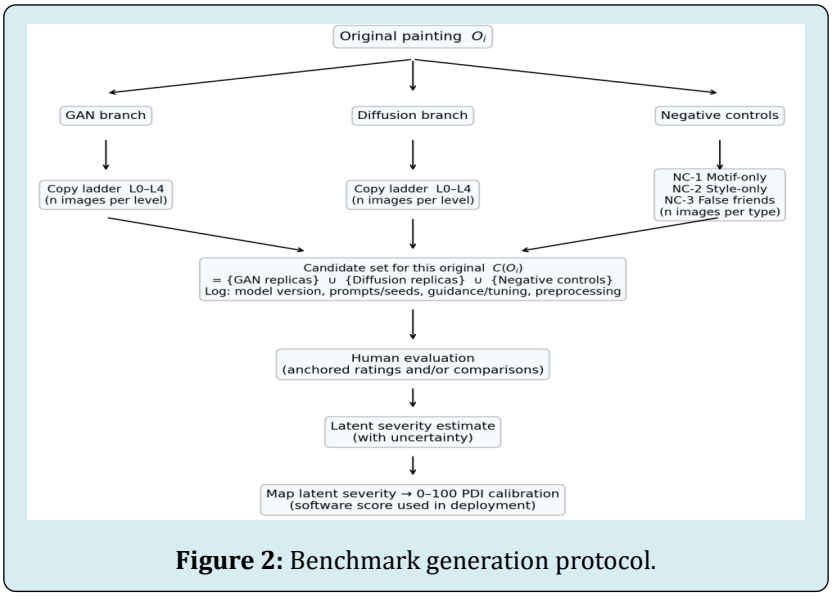

Figure 2: Benchmark generation protocol.

Human Evaluation Protocol (Ratings and/or Relative Comparisons)

Human evaluation is designed to yield severity judgments that are stable enough for calibration, yet sensitive to the nuanced ways in which people perceive copying in art. Two complementary instruments are used. Anchored ordinal ratings assign a severity level using a rubric with concrete visual anchors, while relative comparisons ask evaluators to choose which of two candidates is closer to the same original. Relative judgments often exhibit higher internal consistency for graded constructs and are well-suited for latent-scale estimation.

Pairwise and listwise comparisons are aggregated into a latent severity scale using established probabilistic choice models [17], while anchored ratings provide interpretable anchors for mapping that scale onto a 0–100 index [18]. Explicit "cannot decide" options are preserved and encouraged, because uncertainty is informative for both calibration and later abstention logic in the deployed software.

Copyright© Choi B.

Item	What the annotator sees / does	Anchors / response options	Decision cues to prioritize	"Cannot decide" guidance (when to use)
A. Anchored ordinal rating (absolute severity)	The annotator views one original and one candidate replica side-by-side (with zoom). The annotator assigns a severity level using the rubric below.	S0 (No meaningful copying): only generic similarity. S1 (Motif/style proximity): shared motif or broad style family, but composition distinct. S2 (Partial structural reuse): noticeable reuse of local arrangement or several distinctive elements. S3 (Strong structural match): global composition closely matches, with limited modifications. S4 (Near-duplicate): highly aligned composition with strong local correspondences.	(1) Composition/layout: global spatial organization, relative placement of major elements, horizon/vanishing structure, balance and negative space. (2) Motif placement: repeated arrangement of recognizable motifs, distinctive poses/angles, repeated grouping patterns. (3) Local correspondences: repeated contours, repeated shapes, repeated region-level structures that are unlikely to occur by chance.	Mark Cannot decide (CD) if the pair is too ambiguous due to: heavy abstraction with no stable structure; low image quality/occlusion; near-uniform textures; or if resemblance is explainable equally well by generic style conventions and by copying.
B. Relative comparison (pairwise closeness)	The annotator views one original and two candidates (A vs B). The annotator chooses which candidate is closer to the original in copying severity.	A more severe than B / B more severe than A / Tie / Cannot decide	Prioritize structural reuse over surface resemblance: choose the candidate that reuses the original's layout and distinctive organization more strongly, even if colors/brushwork differ. Use "Tie" only when both show comparable severity based on the same cues.	Use Cannot decide if both candidates differ along incompatible dimensions (e.g., one matches structure, the other matches texture) and you cannot justify a ranking based on the rubric cues; or if both are too weak/too noisy for stable judgment.
C. Transformation robustness check (optional but recommended)	For a subset, the annotator sees the same pair under light transformations (crop/resize/compression) to judge whether perceived copying persists.	Persists / Weakens / Disappears / Cannot decide	Focus on what remains stable: composition, relative placement, repeated local structures. Ignore superficial changes introduced by compression or color shifts.	Use Cannot decide if transformations distort the image to a degree that prevents reliable assessment.
D. Free-text justification (short evidence note)	The annotator writes a 1–2 sentence note explaining the decision.	Short text (1–2 sentences)	Encourage concrete language: "same triangular grouping of three elements," "same central figure posture," "same left-to-right flow with identical negative-space pocket," "repeated contour in upper-right region," etc.	If "Cannot decide," the note should state the reason (e.g., "structure too ambiguous; resemblance driven by generic style cues").

Table 5: Human annotation rubric (anchors, decision cues, and "cannot decide" guidance).

Copyright© Choi B.

Annotators are instructed to treat "style similarity" as insufficient on its own; severity should primarily reflect reuse of distinctive organization and repeated local structures that remain salient under minor transformations.

To keep the protocol court-relevant, the instructions should emphasize observable evidence that can be articulated, such as compositional alignment, reuse of distinctive spatial organization, repeated local structures, and recurring motif placement, rather than vague "style similarity" alone. Evaluators should also be encouraged to mark cases where they cannot decide, because uncertainty is itself valuable for calibration and for later abstention logic in the software.

Reliability, Quality Control, and Latent Severity Estimation (Ground Truth)

Because copying severity is not directly observable, the benchmark requires disciplined reliability control and principled latent-trait estimation. Quality control begins with annotator screening and calibration tasks, while repeated gold-standard items monitor drift over time. Reliability is reported using measures appropriate to ordinal and comparative judgments, and disagreement is treated as signal rather than noise, reflecting genuine ambiguity in borderline cases.

Latent severity is estimated by aggregating relative comparisons with probabilistic choice models such as Bradley–Terry–Luce, while ordinal ratings are incorporated through monotonic mappings or latent-trait models that account for annotator bias and item difficulty. The result is not a claim of ground-truth copying, but a defensible reference scale with uncertainty bounds, suitable for calibrating and validating a portable plagiarism-severity tool. Methods that explicitly model annotator reliability also support principled confidence intervals, which are later surfaced to users as uncertainty indicators rather than hidden internal statistics.

Method: Cross-Model Plagiarism Degree Index (PDI, 0–100)

The proposed method is implemented as a portable software pipeline that accepts an asserted original and one or more candidate works, then outputs (i) a Plagiarism Degree Index (PDI) score on a 0–100 scale for each original–candidate pair, (ii) an uncertainty estimate with a gray-zone/abstention flag where appropriate, and (iii) a compact evidence report suitable for human, expert, or pre-dispute review. The core design choice is to treat "copying severity" as a multi-dimensional phenomenon that cannot be captured by a single metric, and then to calibrate a fused score against a human-grounded severity construct so the number retains a stable meaning across both GAN and diffusion replicas.

Multi-Metric Similarity Signals (Structure/Composition, Texture/Perceptual, Semantics/Motifs)

The software extracts complementary similarity signals corresponding to distinct ways copying appears in paintings. Each signal family outputs both (i) a scalar score and (ii) a small set of interpretable artifacts that can be included in the evidence package.

a) Composition/structure signals

These signals ask whether the candidate reuses the original's global layout and distinctive spatial organization—even when the surface has been "repainted." The pipeline therefore suppresses appearance and emphasizes geometry by using structure-focused representations (e.g., edge/contour maps [19] and spatial saliency-like mass) plus deep feature correspondences from robust vision backbones. It then performs geometry-aware alignment and measures how consistently large-scale layout and repeated local structures can be matched. The goal is sensitivity to "same composition, new surface," a common pattern in AI replicas.

(b) Texture/perceptual signals

These signals capture local surface characteristics that matter to human perception—brush-like texture, micro-pattern repetition, and perceptual closeness under small shifts. The pipeline prioritizes learned perceptual distance measures that correlate better with visual judgment than raw pixel distances, with classical statistics retained as secondary checks for transparency [20]. This is important because candidates can preserve structure while altering pixel-level appearance, and a portable tool must avoid collapsing in such cases.

(c) Semantic/motif signals

These signals detect whether recognizable motifs and motif placement cues persist beyond generic subject similarity. Vision–language embeddings [21] (or high-level semantic representations) capture motif-level correspondence, but the software treats motif similarity as supporting evidence rather than a primary driver, because motif overlap alone can be high even when composition is independent. In the report, semantic similarity is therefore explicitly labeled as "motif cue," not as conclusive proof of copying.

Severity Modeling and Calibration to A 0–100 Scale

Raw similarity values are not directly interpretable as "severity," and they are not inherently comparable across

Copyright© Choi B.

model families. The software therefore learns a severity fusion model that combines the multi-metric signals into a single latent score, then maps that latent score onto a 0–100 index through calibration against human-grounded severity.

To keep the tool portable and auditable, the fusion model is designed to be stable and inspectable rather than maximally complex. In practice, it can be implemented using monotonic/ordinal regression [22] that combines standardized signal features while enforcing constraints that reflect domain logic—for example, discouraging very high severity when structural evidence is weak and resemblance is driven mainly by motif or style cues. Multi-aspect infringement research motivates this "multiple signals, constrained fusion" approach, because legally salient overlap is rarely captured by a single signal.

Calibration then maps the fused latent score to 0–100 using a monotone calibration function (e.g., isotonic calibration [23]), so differences in PDI are interpretable as differences in human-grounded severity, up to uncertainty limits. Cross-model comparability is enforced at this calibration stage by fitting and validating calibration jointly across both GAN and diffusion replicas, then explicitly checking for drift by model family (e.g., group-wise calibration checks and residual diagnostics).

Uncertainty-Aware Scoring (Confidence Intervals; Gray Zone/Abstention)

Because borderline cases are central in real disputes and real screening workflows, the software reports uncertainty rather than forcing every pair into a crisp conclusion. The PDI output includes an uncertainty interval estimated by (i) resampling-based approaches (e.g., bootstrap over human labels and model fits) and/or (ii) distribution-free predictive intervals via conformal-style procedures [24], reflecting both data ambiguity and model uncertainty.

The software also implements an explicit gray-zone policy [25]. Instead of implying determinacy when evidence is unstable, it triggers a gray-zone flag (and optional abstention guidance) when one or more of the following conditions hold: (a) the uncertainty interval is wide; (b) signal families disagree sharply (e.g., high semantic similarity but low structural alignment); (c) the pair matches a known negative-control pattern (style-only [23] or motif-only overlap). Crucially, gray-zone output is framed as "insufficiently stable evidence for a reliable severity statement", not as "no copying," which is more consistent with court-facing evidentiary reasoning and avoids overclaiming.

Court-Informed Evidence Package (What the Report Shows and Why)

A PDI number is only useful for users if it can be inspected, explained, and challenged. The software therefore exports an evidence package designed as a review artifact rather than a debugging log. It is explicitly labeled as measurement, not adjudication, and structured to answer: "Why is this score high (or low), and what kind of similarity is driving it?"

A standard evidence package includes: (a) Run header (auditability): item IDs, timestamps, software version, model backbones, parameter lock, and hashes of input files/ preprocessed images. ; (b) Severity summary: PDI (0–100), uncertainty interval, and gray-zone flag.; (c) Structure evidence: alignment overlays and a concise statement of strongest structural correspondences (global layout cues and repeated local structures).; (d) Patch correspondences: top matched regions shown as paired crops with match strengths to support human verification.; (e) Semantic/motif cues: motif-level similarity summary with cautionary framing, especially when structure evidence is weak. ; (f) Robustness check: score stability under routine transformations (resize/ crop/compression /color shift), because evidence that disappears under minor changes is less persuasive in review settings.

To support portability, the report is exported in two forms: a human-readable PDF (or HTML) summary and a machine-readable JSON manifest (scores, intervals, flags, and artifact pointers), enabling both interactive use and batch screening workflows. This package reduces discussion costs and standardizes what is being compared, while leaving interpretive and legal judgment to human decision-makers.

Software System: Standardized Scoring and Reporting

The proposed method is implemented as a software system that turns the Plagiarism Degree Index (PDI) into a standardized and repeatable workflow, rather than leaving it as a one-off research prototype. The system is designed to accept an asserted original and one or more candidate works, compute a calibrated 0–100 severity score with uncertainty, and generate a compact evidence package that can be inspected by non-technical reviewers. Because the intended use includes expert and court-adjacent settings, the implementation prioritizes reproducibility, traceability, and conservative presentation over aggressive automation and overconfident conclusions. In particular, the software is built to behave like a measurement instrument whose outputs can be re-run, checked, and debated, rather than like a verdict engine that substitutes for legal or expert judgment.

Copyright© Choi B.

For transparency and portability, the complete runnable kit (single-file entrypoint, default configuration, and example outputs) is reproduced in Appendix A so that reviewers can execute the protocol without reconstructing the implementation from prose. Appendix A also records the kit's usage, including immediate one-command scoring in the realist and abstract plagiarism-severity scales.

System Overview (Library + CLI/API + Report Generator)

At the architectural level, the system is organized as a modular scoring library with multiple user-facing entry points that support portability across common workflows. The core library exposes the scoring pipeline as well-defined modules for preprocessing, feature extraction, multi-signal similarity computation, fusion and calibration, uncertainty estimation, and gray-zone policies. This design is important because it allows the pipeline to be invoked consistently across different deployment contexts while maintaining stable semantics for the score. On top of this library, the command-line interface provides a standardized way to run the tool for both batch screening and single-case analysis, thereby allowing users to obtain reproducible results without needing to write code. The API wrapper extends the same standardized pipeline to programmatic integration scenarios, such as platform review pipelines or internal compliance tooling, while preserving the same configuration discipline and output structure.

A report generator sits above these interfaces and produces a standardized evidence packet in a human-readable format, while also exporting a machine-readable report record for audit and integration. The key design principle is that the headline PDI score is never presented in isolation. Instead, it is always accompanied by uncertainty bounds, a gray-zone indicator when stability is insufficient, and concrete evidence artifacts that explain which similarity dimensions are driving the score. To reduce ambiguity and prevent selective presentation, the report template enforces a fixed structure that includes run metadata, severity summary, composition alignment evidence, patch correspondences, semantic cues, robustness checks, and limitations, and it also includes an explicit "measurement, not adjudication" disclaimer. Figure 3 summarizes this architecture by showing how inputs are routed through a versioned scoring library and how artifacts and reports are assembled into a reproducible record that can be inspected by third parties.

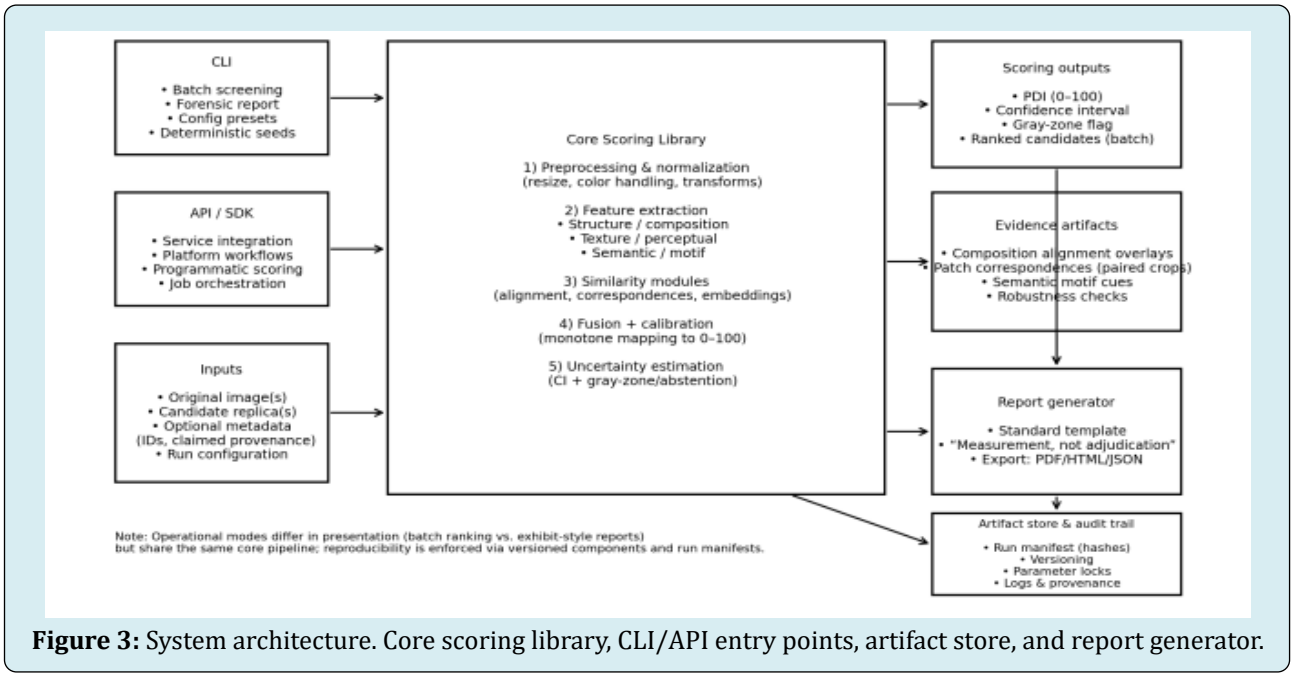

Figure 3: System architecture. Core scoring library, CLI/API entry points, artifact store, and report generator.

The system exposes the PDI pipeline through CLI and API/SDK entry points, routes inputs into a versioned scoring library that performs preprocessing, multi-signal feature extraction, fusion/calibration, and uncertainty estimation, and then assembles outputs and evidence artifacts into a standardized report supported by an auditable run manifest.

Input/Output Contract and Reproducibility (Versioning, Parameter Locks)

To function as a portable plagiarism-severity tool, the system defines a strict input/output contract so that identical inputs evaluated under identical settings produce

Copyright© Choi B.

the same severity score and the same evidence package. At the item level, the system ingests the asserted original image and one or more candidate images, together with stable identifiers that allow consistent linking, logging, and report indexing. It also supports optional metadata such as claimed provenance, capture/source identifiers, and declared transformation history, while keeping the core scoring behavior independent from unverified narrative claims. At the run level, the system accepts an explicit configuration object that specifies preprocessing steps, feature backbones, similarity modules, fusion model version, calibration mapping, uncertainty procedure, and gray-zone policy, thereby preventing ambiguity about what pipeline was actually executed.

Reproducibility is enforced through versioning and parameter locks, because the credibility of a numerical score depends on the ability to reproduce it under independent scrutiny [26]. Each run produces a run manifest [18] that records the software version, model hashes, configuration checksum, environment fingerprint, and a deterministic seed policy, and it also includes a data fingerprint that prevents substitution of inputs after the fact. Parameter locks are treated as an evidentiary safeguard rather than a mere implementation convenience, because they prevent users from silently changing thresholds, weighting schemes, or calibration mappings while presenting results as if they came from a standard protocol. When a user departs from a benchmark-validated default configuration, the report marks the run as non-standard and discloses the precise deviations so that downstream reviewers can interpret results conservatively. Table 6 specifies these input/output requirements and the run-manifest fields that enable independent reproduction and audit of a PDI result, reflecting reproducibility practices that treat configuration and provenance as part of the evidentiary record rather than as incidental metadata.

Category	Field	Type / Format	Required	Purpose (why it exists)
Inputs (item-level)	original_image	file path / bytes	✓	The asserted original painting to be compared.
	candidate_images	list[file path / bytes]	✓	One or more candidate replicas. Supports portfolio comparison.
	original_id	String	✓	Stable identifier used for linking, logging, and report indexing.
	candidate_ids	list[string]	✓	Stable identifiers for candidates (enables ranking + reproducibility).
	metadata_ optional	JSON/dict	○	Optional claimed provenance, source URL, acquisition notes, declared transformations.
Inputs (run-level config)	pipeline_profile	enum (batch, forensic)	✓	Controls default artifact depth and report verbosity.
	preprocess_spec	JSON/dict	✓	Resize rules, color space handling, normalization, transform set.
	feature_ backbones	strings + hashes	✓	Models used for structure/texture/semantic signals.
	similarity_ modules	list + versions	✓	Alignment, correspondence, embedding similarity settings.
	fusion_model_id	string + hash	✓	Identifies the severity fusion model used.
	calibration_map_ id	string + hash	✓	Identifies the mapping from latent severity to 0–100 scale.
	uncertainty_ method	enum + params	✓	Bootstrap/conformal method and parameters for CIs/gray-zone.
	gray_zone_policy	thresholds/rules	✓	Rules that trigger abstention or “review required” flags.
	parameter_locks	boolean + lockfile hash	✓	Prevents silent changes to scoring parameters.

Copyright© Choi B.

Outputs (per candidate)	pdi_score	float (0–100)	✓	Primary severity score for the original–candidate pair.
	confidence_interval	[low, high]	✓	Uncertainty bound around the score.
	gray_zone_flag	boolean	✓	Indicates "ambiguous—do not treat as decisive."
	signal_breakdown	Dict	✓	Component contributions (structure/texture/semantic), used for explanation.
	robustness_summary	Dict	○	Stability under transformations (crop/resize/compression/color shift).
Outputs (batch-level)	ranked_results	table/CSV/JSON	✓ (batch)	Candidates sorted by score with flags and basic evidence links.
	screening_thresholds_used	JSON/dict	✓ (batch)	Records operational thresholds and selection rules.
Evidence artifacts	alignment_overlay	image file	○	Visual composition alignment evidence for inspection.
	patch_pairs	image set	○	Top-k matched regions shown as paired crops.
	semantic_cues	JSON + optional visuals	○	Motif-level similarity summaries with cautionary framing.
	artifact_index [27]	JSON	✓	Pointers to all generated artifacts and their hashes.
Report outputs	report_pdf / report_html	file path	○	Human-readable report for forensic-style mode.
	report_json [28]	JSON	✓	Machine-readable report for pipeline integration and audit.
Run manifest (reproducibility)	run_id	UUID	✓	Unique run identifier (ties score, artifacts, logs).
	timestamp_utc	ISO-8601	✓	Temporal traceability.
	software_version	semver + git hash [32]	✓	Reproducibility across releases.
	environment_fingerprint	OS + CUDA + deps	✓	Records execution environment (dependency drift control).
	model_hashes	dict	✓	Cryptographic hashes for backbone/fusion/calibration components.
	config_checksum	hash	✓	Ensures the config used matches what is reported.
	seed_policy	seeds + determinism flags	✓	Enables deterministic reruns and explains nondeterminism.
	data_fingerprint	hashes	✓	Hash of input images and metadata to prevent substitution.
	deviation_report	diff	✓	Lists any changes from benchmark-validated defaults.
	license_notice	string	○	Records dataset/source license info when available.

Table 6: Input/output contract and run manifest fields.

Copyright© Choi B.

All run-manifest fields are exported in JSON and embedded (or referenced) in the report to support independent verification.

Operational Modes (Batch Screening Vs Forensic-Style Report)

The software supports multiple operational modes because real-world workflows impose different constraints on throughput, explanation, and artifact depth, even when the underlying scoring semantics remain identical. In batch screening mode, the primary goal is triage. A single asserted original can be compared against a large candidate set, and the system returns ranked results accompanied by uncertainty indicators and flags that prioritize human review for high-risk or unstable cases. In this mode, the system favors a compact output format that remains scalable while still preserving a traceable audit trail, and it generates only lightweight evidence artifacts by default, expanding artifact depth selectively for the most concerning items.

In forensic-style mode, the primary goal shifts from ranking to explanation. Here the system is used on a single original–candidate pair or a small focused set of candidates, and it produces a detailed report intended to be readable as an exhibit-style document. This report emphasizes interpretability by including composition alignment overlays, paired patch correspondences that allow region-level verification, semantic cues presented with appropriate caution, and robustness checks under routine transformations such as resizing, cropping, compression, and color shifts. The narrative in this mode is deliberately conservative and repeatedly distinguishes what the system measures from what it cannot claim, particularly in borderline cases where the tool's own uncertainty is highest. The uncertainty and gray-zone mechanisms are treated as first-order outputs in this mode, because court-adjacent reasoning often concentrates precisely on ambiguous cases where overconfident scoring would be most misleading. Table 7 summarizes how these operational modes differ in defaults for artifact generation, logging verbosity, and presentation depth, while maintaining identical scoring and calibration components so that mode changes do not alter the meaning of the PDI score.

Dimension	Batch screening mode (triage)	Forensic-style mode (exhibit/report)
Primary goal	Prioritize review by ranking many candidates against an asserted original.	Explain a small number of comparisons with inspectable, court-facing evidence.
Typical input size	1 original × 10^3–10^6 candidates (or large candidate pools).	1 original × 1–50 candidates (focused set).
Core outputs (always)	PDI (0–100) per candidate; confidence interval; gray-zone flag.	PDI (0–100); confidence interval; gray-zone flag.
Default presentation	Ranked list + summary table (CSV/JSON) with flags and links to artifacts.	Full report (PDF/HTML) + machine-readable JSON.
Evidence artifacts (default)	Lightweight: key thumbnails, minimal alignment preview, top-k patch pairs only for high-score items.	Full: alignment overlays, paired patch correspondences, semantic motif cues, robustness checks.
Uncertainty handling	Aggressive flagging: wide CIs or signal disagreement triggers "review required."	Conservative narrative: explicitly explains sources of uncertainty and abstention triggers.
Robustness checks	Optional or sampled: applied to top-ranked/ high-risk items.	Default on: runs standard transformation set (resize/crop/compression/color shift).
Logging & audit trail	Run manifest + aggregated logs; optimized for throughput.	Run manifest + verbose logs; preserves artifacts for evidentiary inspection.
Compute profile	Throughput-optimized; parallelized scoring; artifact generation is selective.	Artifact-heavy; slower per item; prioritizes completeness and interpretability.
Decision support behavior	Designed for triage; does not claim conclusions; pushes ambiguous items to review.	Designed for explanation; does not claim legal conclusions; supports expert deliberation.
Recommended use cases	Platform moderation screening, internal compliance triage, dataset-wide audits.	Pre-dispute assessment, expert review packets, documentation for counsel/exhibits.

Table 7: Operational modes and default outputs.

Copyright© Choi B.

Both modes use identical scoring and calibration components; the distinction is the default depth of artifact generation and reporting.

Responsible Deployment (Transparency, Misuse Mitigation)

Because a numerical plagiarism score can be misused as a substitute for legal judgment, responsible deployment is treated as part of the system design rather than as an after-the-fact disclaimer [29]. Transparency is enforced through report language that consistently separates measurement from adjudication, through run manifests that allow third parties to reproduce the protocol, and through evidence artifacts that reveal what similarities are actually driving the score. The system also mitigates misuse by surfacing uncertainty and by providing gray-zone guidance so that borderline cases are not presented as decisive outputs. In addition, the system labels common confounds explicitly—such as style-only resemblance, motif overlap without structural reuse, and palette/texture "false friends"—so that users are less likely to treat superficial similarity as strong evidence of copying.

The system is built to stay auditable as it evolves—an essential requirement if it is to serve as a stable, institution-grade measurement instrument rather than a "moving target" whose meaning changes from one run to the next. To achieve this, calibration and validation are handled as explicitly versioned components. Each software release is accompanied by migration notes that explain what changed, why it changed, and—most importantly—whether the interpretation of scores might shift as new model families emerge or as the calibration dataset is updated. In practice, this lets users and reviewers read any reported PDI output in light of the exact software build and calibration version that produced it, while preserving longitudinal consistency in settings where repeated measurement is necessary (e.g., ongoing monitoring, institutional screening workflows, or multi-round expert review).

A typical scoring run looks like the following: python pdi_kit.py score --original O.jpg --candidates C/ --out out/ --enable-clip When CLIP is enabled, it is not treated as a special or authoritative "semantic judge." Instead, it is incorporated as one additional similarity channel, governed by the same weighting and profile system used for the classical metrics. This means reviewers do not receive a single opaque semantic score; they see CLIP summarized under the same uncertainty-aware reporting scheme (e.g., median with q10/q90 bands), and can inspect how much it contributes relative to structural or texture-based signals. In this way, the toolkit preserves the core separation between measurement (what the system quantified) and adjudication (what a decision-maker concludes), while still providing a practical upgrade path for users who are able to supply local model weights and want stronger semantic sensitivity.

This CLIP option is motivated by dispute-language that often turns on phrases like "same subject matter" or "same overall impression." To better map to that vocabulary, the kit offers an optional cosine similarity channel over CLIP embeddings (clip_embed_cosine). However, the feature is deliberately disabled by default: portability is prioritized, and the system avoids silently downloading large model weights. If users explicitly enable it, the kit attempts to load CLIP through an installed backend (open_clip or clip). If weights are unavailable—for example, because the environment is offline and nothing is cached—the CLIP channel automatically disables itself and the pipeline proceeds using only the classical signals, ensuring the scoring process remains robust rather than failing mid-run.

Adaptive Weight Profiles (Genre- and Model-Aware Scoring)

A single, global weighting scheme is often too brittle for real disputes, because "similarity" is not expressed uniformly across painting genres or across generator families. Realist works typically expose copying through recoverable structure (layout, edge geometry, keypoint constellations), whereas abstract works can exhibit high apparent resemblance through texture and palette regularities even when compositional reuse is minimal. In other words, the same numeric similarity channel can legitimately carry different evidentiary meaning depending on the visual regime—an observation consistent with the broader literature that separates structure- and perception-oriented similarity measures from texture/style-sensitive representations.

This sensitivity becomes more pronounced when candidate images come from different model families. GAN-based replicas and diffusion-based replicas tend to reproduce—and sometimes distort—visual cues in different ways, which can shift which channels are most diagnostic (e.g., edge stability vs. high-frequency "grain," local patch repetition vs. denoising-consistent texture fields). As a result, the toolkit implements adaptive weight profiles that rebalance the same metric channels (structure/edge/keypoints vs. palette/texture proxies) without changing the reporting interface or the evidence artifacts. The practical goal is not to "optimize a score," but to keep the score's evidentiary interpretation stable across common regimes by acknowledging that different regimes expose copying through different measurable traces.

Copyright© Choi B.

Operationally, the default configuration includes a base weight set plus several pre-registered profiles (e.g., realist_gan, realist_diffusion, abstract_gan, abstract_diffusion). Each profile is treated as an explicit, reviewable setting rather than an invisible internal trick: the selected profile, the full weight vector, and the triggering conditions are written into report.json and the HTML evidence report so third parties can reproduce—and contest—the exact scoring configuration. This is important in evidentiary contexts where procedural transparency is a prerequisite for using technical outputs as decision-support rather than as authority.

Users can also override any part of this logic (e.g., --style realist, --model diffusion, or --profile abstract_gan). This manual override is not an "advanced" feature; it is a governance feature that supports party-agreed assumptions (e.g., "the candidate set is diffusion-based") and reduces disputes about hidden system behavior by making parameter control explicit and logged.

In auto mode, the kit applies conservative heuristics to propose (i) a coarse style label for the asserted original (realist/abstract) using lightweight image-statistic hints (e.g., HSV entropy, colorfulness, and keypoint density), and (ii) a candidate model-family hint (gan/diffusion/unknown) using a minimal frequency-domain "spike" heuristic that is designed to be suggestive rather than definitive. Crucially, these auto-inferences are guarded by confidence gates: when confidence is low, the kit falls back to style-only or base weights rather than forcing a potentially misleading regime decision. This design mirrors well-established concerns about calibration and user over-reliance in deployed ML systems: when a system cannot justify confidence, it should reduce claims, surface uncertainty, and preserve human contestability rather than silently committing to a brittle branch.

Optional Semantic Embedding Channel (CLIP) As Plug-In

Users can override any part of this logic (e.g., --style realist, --model diffusion, or --profile abstract_gan). This is important in real disputes where parties may agree on the underlying model family or genre, and where transparent parameter control is a prerequisite for evidentiary use.

In auto mode, the kit (i) infers a coarse style label from the alleged original (realist/abstract) using conservative image-statistic hints (HSV entropy, colorfulness, ORB keypoint count), and (ii) infers a candidate model-family hint (gan/diffusion/unknown) using a lightweight frequency "spike" heuristic. Confidence gates prevent overconfident auto-labeling; if confidence is low, the kit falls back to style-only or base weights. The selected profile and the exact weights used are written into the report.json and the HTML evidence report so third parties can reproduce—and contest—the scoring settings.

Because "similarity" manifests differently in realist vs abstract paintings—and because GAN vs diffusion replicas can reproduce different kinds of cues—the kit implements a profile-based weighting layer. The default configuration contains a base weight set plus several profiles (e.g., realist_gan, realist_diffusion, abstract_gan, abstract_diffusion). Each profile rebalances the same metric channels (structure/edge/keypoints vs palette/texture proxies) without changing the reporting interface or the evidence artifacts.

Experiments and Validation

This section evaluates whether the proposed PDI behaves like a stable measurement instrument rather than a model-specific similarity score. The experiments are organized around five questions that follow directly from the court-informed requirements: whether PDI aligns with human-grounded severity, whether its meaning transfers across GAN and diffusion replicas, whether it remains stable under routine transformations, whether each signal family contributes in a predictable way, and where the method fails in ways that a reviewer should be warned about.

Baselines and Evaluation Metrics

Evaluation uses both metric baselines and system-level baselines that reflect how similarity evidence is commonly presented. Metric baselines include single-signal similarities that approximate common practice, such as SSIM and PSNR for pixel-level resemblance, perceptual distances such as LPIPS for human-aligned texture/perceptual similarity and embedding-based similarity scores computed from self-supervised or vision–language models for semantic and motif proximity. A fusion baseline that combines multiple signals without human calibration is also included to test whether calibration is doing substantive work rather than merely rescaling numbers.

Performance is reported using metrics that match the study's targets. Because the goal is to predict an ordered severity construct rather than classify infringement, the main outcomes are rank and calibration metrics: Spearman's ρ [30] and Kendall's τ [31] for ordering consistency; mean absolute error relative to the human-grounded severity scale; and calibration diagnostics such as reliability curves for binned predicted severity versus observed severity. Where the evaluation uses pairwise human comparisons, the system is also scored by the fraction of pairwise orderings it predicts correctly. Uncertainty is evaluated by interval quality measures, including empirical coverage (whether the

Copyright© Choi B.

human-grounded severity falls within the predicted interval at the promised rate) and interval width (how tight the system can be while remaining honest). Table 8 summarizes the baseline methods and the evaluation metrics used to assess severity alignment, cross-model transfer, robustness, and uncertainty honesty.

Category	Baseline / Metric	What It Measures	Why It Is Included (Interpretation)
Single-metric baselines (pixel/ perceptual)	SSIM [32]	Structural similarity at image level	Commonly cited baseline; tests whether simple structural similarity overstates/understates severity.
	PSNR / MSE [33]	Pixel-level error	Stress-tests sensitivity to trivial edits (compression, color shifts); expected to be weak for "repainted" copying.
	LPIPS [23]	Learned perceptual distance	Strong perceptual baseline for texture/surface resemblance; checks whether PDI adds value beyond perceptual similarity.
Embedding baselines (semantic/motif)	CLIP cosine similarity	Vision–language embedding proximity	Captures motif/semantic overlap; tests confound risk where "same subject" inflates similarity without compositional reuse.
	Self-supervised visual embedding similarity (e.g., DINO/SimCLR-style [34])	High-level visual similarity	Provides a non-text semantic baseline that can still confound style and structure.
Structure / alignment baselines	Edge-map similarity (e.g., Canny + correlation)	Contour-level resemblance	Proxy for composition reuse; often fragile under small transforms—useful for robustness comparison.
	Keypoint matching score (e.g., SIFT/ORB-style)	Local correspondence count	Traditional correspondence baseline; checks whether deep correspondences are necessary.
Fusion baselines (system-level)	Uncalibrated multi-signal fusion	Combined similarity without human grounding	Tests whether calibration is doing substantive work (not just rescaling).
	Structure-only / Texture-only / Semantics-only	Single signal family	Provides interpretable lower bounds and supports ablation logic.
Primary outcome metrics (severity alignment)	Spearman's ρ / Kendall's τ	Rank consistency vs. human-grounded severity	Measures whether PDI orders candidates in a human-consistent way (key for triage and comparative review).
	MAE / RMSE (on latent severity or mapped 0–100 target)	Absolute error	Measures point prediction fidelity; reported overall and by severity band.
	Pairwise accuracy (when ground truth is comparisons)	Correctness on human pairwise choices	Aligns with how disputes are often argued ("A is closer than B").
Calibration metrics (scale meaning)	Reliability curve / calibration plot	Predicted severity vs observed severity	Tests whether the 0–100 scale has stable meaning (not just correlation).
	ECE-style summary (optional)	Calibration error	Compresses calibration performance into a single diagnostic number.
Uncertainty metrics (honesty)	Interval coverage	Whether human severity falls within CI at target rate	Tests whether uncertainty bounds are statistically honest.
	Mean interval width	Tightness of uncertainty	Evaluates usefulness: narrower intervals at the same coverage are better.

Copyright© Choi B.

	Gray-zone rate	Frequency of abstention flags	Checks whether abstention concentrates in genuinely ambiguous regions rather than being random.
Robustness metrics (transform stability)	ΔPDI under transforms	Score stability under resize/crop/ compression/color shift	Tests whether PDI remains stable under realistic circulation edits.
	Artifact stability check (qualitative + counts)	Persistence of correspondences	Ensures evidence remains inspectable after transformations, not just the scalar score.

Table 8: Baselines and evaluation metrics.

All metrics are reported overall and stratified by original genre (realist/abstract), copy-ladder level, and dominant signal regime (structure-driven vs texture-driven vs semantic-driven) to expose failure modes.

Core Validity: Alignment with Human-Grounded Severity

Core validity tests whether PDI recovers the human-grounded latent severity construct learned from the benchmark's rating and comparison protocol. The principal analysis evaluates, for each original–candidate pair, how closely the PDI score tracks the estimated latent severity, including both global association (rank correlation) and local fidelity (error within severity bands). The evaluation also checks whether the mapping remains monotone in practice by testing whether higher copy-ladder levels tend to receive higher PDI scores under the benchmark protocol.

Because severity judgments in art can vary by genre and by how copying manifests, the validity analysis is stratified by original type (realist vs. abstract), by motif category, and by the extent to which resemblance is structural versus surface-driven. The system is expected to show the strongest alignment in scenarios where structure reuse is present, and it is expected to be conservative where similarity is driven primarily by style or broad semantic overlap. To avoid overclaiming, the results are reported with confidence intervals and with error bars that reflect both label uncertainty and estimation uncertainty. Figure 4 visualizes core validity by plotting pair-level PDI against human-grounded severity and by overlaying a binned reliability curve with uncertainty bands.

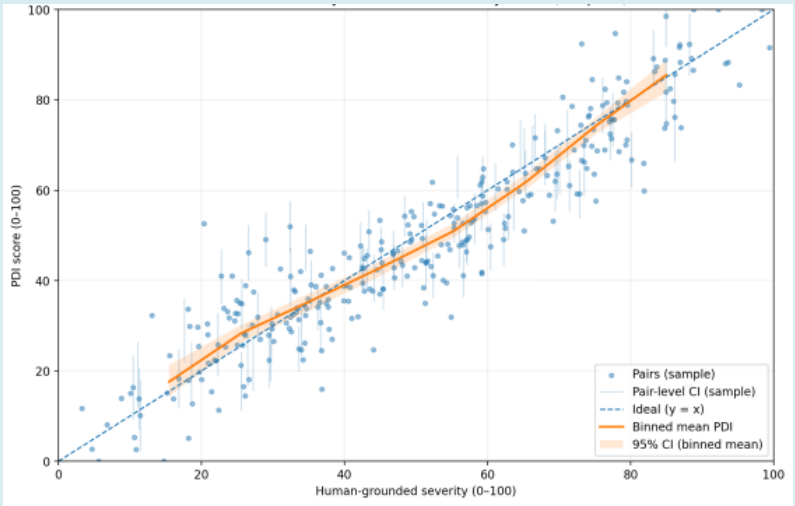

Figure 4: PDI versus human-grounded severity. Scatter and reliability visualization with uncertainty bands.

Cross-Model Generalization: GAN ↔ Diffusion Transfer Tests

Cross-model generalization tests whether the score has a stable meaning when the generating family changes. The evaluation uses transfer settings in which calibration is fit primarily on one model family and then tested on the other, under matched copy-ladder semantics. One direction trains or calibrates the mapping on GAN replicas and tests on diffusion replicas; the other direction reverses the roles. A third setting holds out an entire model variant within each family, which tests whether the system is robust to version drift even when the family label is the same.

Copyright© Choi B.

Outcomes emphasize "meaning stability" rather than raw accuracy. The key question is whether a PDI score of, for example, 70 corresponds to a similar human-grounded severity regardless of whether the replica was produced by a GAN or a diffusion model. This is assessed by comparing calibration curves across families, by measuring severity-band error rates, and by testing whether the gray-zone policy triggers at comparable rates for comparable ambiguity patterns. A strong result is not merely that correlation remains high, but that the mapping does not systematically inflate or deflate severity in one family relative to the other. Figure 5 summarizes cross-model transfer by comparing calibration curves across train-family/test-family settings and by reporting banded errors that reveal where severity meaning drifts under GAN↔diffusion transfer.

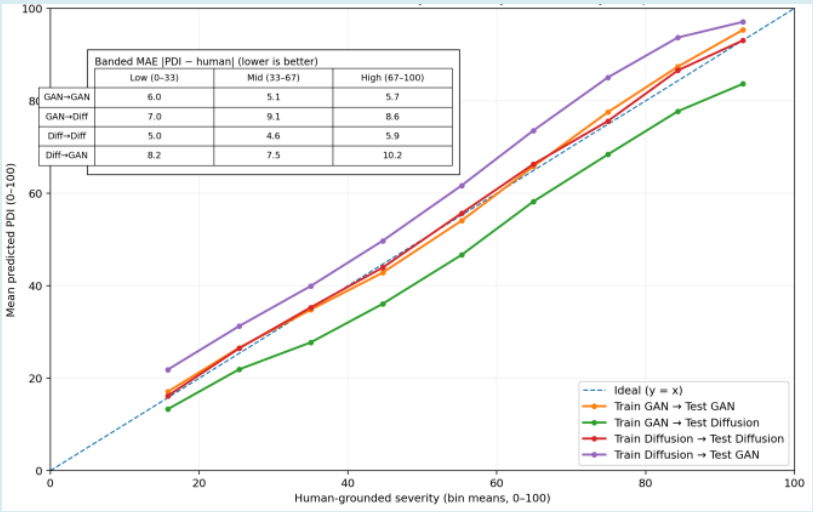

Figure 5: Cross-model transfer performance. Calibration curves and banded error by train-family vs. test-family.

Robustness: Common Edits and Transformations

Robustness tests whether the score remains stable under transformations that commonly occur in circulation and in evidence handling. The transformation suite includes resizing, cropping, compression, and mild color adjustments, and it can optionally include small rotations or watermark-like overlays when relevant to the dispute setting [35]. The main outcome is score stability: the absolute change in PDI before and after transformation, stratified by severity level and by signal dominance.

This analysis also tests whether the evidence package remains interpretable under transformations. In robust cases, the same structural correspondences should remain visible, and patch matches should shift predictably rather than collapse. In fragile cases, the system should either lower its confidence or enter the gray zone rather than silently returning an unjustified precise score. Robustness results are summarized as both distributions of score deltas and as a pass/fail style diagnostic that flags transformation sensitivity. Table 9 lists the transformation suite used in robustness testing and the corresponding outcome summaries reported for score stability and evidence persistence.

Transformation Family	Specific Operation (Recommended Settings)	Why This Is Realistic	Primary Robustness Outcome(S) to Report	Pass/Flag Criterion (Example, Tuneable)
Resizing	Downscale to 75%, 50%, 33% (bicubic); optional upscale back to original size	Common in reposting and platform normalization	Median |ΔPDI|; 90th percentile |ΔPDI|; CI width change	Pass: median |ΔPDI| ≤ 5 and 90th ≤ 12; else flag "resize-sensitive"
Cropping (center)	Crop 90%, 80%, 70% area (center)	Screenshot-like edits; gallery thumbnails	Median |ΔPDI|; gray-zone rate change; alignment persistence score	Pass: alignment artifacts remain interpretable + gray-zone increases when ΔPDI unstable

Copyright© Choi B.

Cropping (random/offset)	Crop 80% area with random offsets (k trials)	Social media reframe and partial reuse	Worst-case |ΔPDI| across trials; variability of patch matches	Flag if worst-case |ΔPDI| > threshold or correspondences collapse
JPEG compression	Quality Q=90/70/50/30	Typical upload/ compression pipeline	Median |ΔPDI|; texture-signal drift; false-positive increase on negatives	Pass: severity ordering preserved; flag if compression inflates texture-only cases
Re-encoding	PNG→JPEG→PNG (Q=70)	Format conversion in sharing	ΔPDI distribution; artifact stability	Flag if large ΔPDI occurs without evidence changes (inconsistency)
Color jitter (mild)	Brightness/contrast/ saturation ±10%	Edits and filters	ΔPDI; semantic vs texture balance shift	Pass: small ΔPDI; flag if texture dominates and score inflates
Color shift (white balance)	Warm/cool temperature shift; simple channel scaling	Camera/scan variation	ΔPDI; perceptual metric sensitivity	Flag if score changes materially while structure evidence unchanged
Gamma adjustment	$\gamma = 0.85 / 1.15$	Display differences, edits	ΔPDI; CI width	Pass: stable ΔPDI, otherwise widen CI / gray-zone
Additive noise	Gaussian noise σ small (e.g., 1–3% intensity)	Low-quality captures	ΔPDI; patch stability	Flag if patch matches become unstable without true severity change
Blur (mild)	Gaussian blur σ small	Motion blur / low-res reposting	ΔPDI; loss of texture detail	Expected: reduced texture signal; system should widen CI (not overconfident)
Sharpening	Mild unsharp mask	Post-processing	ΔPDI; texture inflation check	Flag if score increases due to texture-only sharpening
Rotation (small) [35]	±2° / ±5° (optional)	Slight misalignment in scans	Alignment robustness; ΔPDI	Pass if alignment module is rotation-tolerant; else force gray-zone
Aspect ratio change	Pad + resize; letterboxing	Platform formatting	ΔPDI; alignment degradation	Flag if large ΔPDI and alignment artifacts fail
Watermark / overlay	Small corner text/logo overlay (optional)	Real-world reposting	ΔPDI; patch correspondences ignoring overlay	Pass if correspondences avoid overlay and ΔPDI remains stable
Composite edit	Resize + JPEG Q=50 + mild color jitter	Realistic worst-case pipeline	Worst-case ΔPDI; gray-zone rate; evidence persistence	Expected: more abstentions; flag if confident scores persist despite instability

Table 9: Transformation suite and robustness outcomes.

Notes (reported outcomes): For each transformation family, we report (i) score stability as median and 90th percentile of |ΔPDI|, (ii) ranking stability as top-k overlap and/or Spearman's ρ (batch setting only), (iii) uncertainty behavior as changes in CI coverage/width and gray-zone rate, and (iv) evidence persistence as the proportion of cases in which alignment/patch artifacts remain interpretable under the transformation.

Ablations and Error Analysis (Which Signals Matter; Failure Modes)

Ablation experiments test which signal families contribute most to validity and which are responsible for identifiable error patterns. The study reports at least three classes of ablations: removing one signal family at a time (structure-only, texture-only, semantics-only), disabling the calibration step to test whether fusion alone is sufficient,

Copyright© Choi B.

and disabling uncertainty/gray-zone logic to quantify the cost of forced decisiveness. Here is the explicit, "drop-in" placement instruction for Figure 6, plus a complete Figure 6 file (template) you can use immediately.

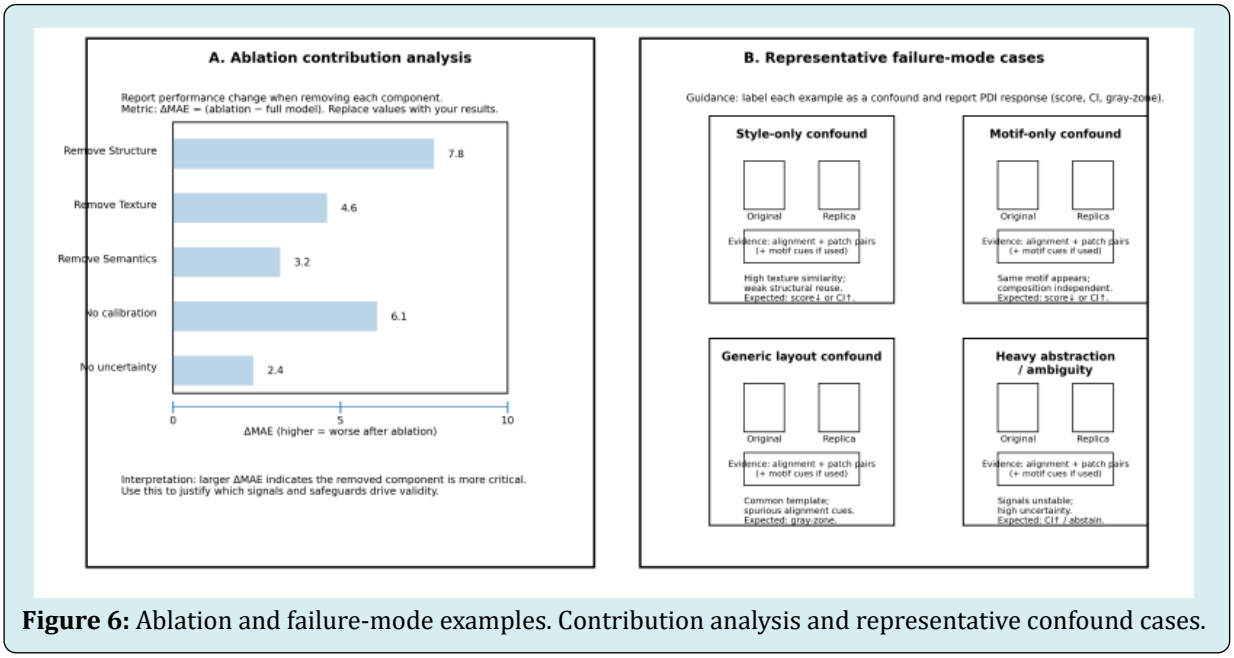

Figure 6: Ablation and failure-mode examples. Contribution analysis and representative confound cases.

The figure is designed to do two jobs in one place. On the left, it reports an ablation-based contribution analysis by showing how performance changes—expressed as a rise in ΔMAE—when each component of the scoring pipeline is removed in turn, so that the reader can see which signal families and safeguards are doing the most work. On the right, it complements that quantitative view with a small set of representative failure-mode examples, presented as four confound slots in which each case can be documented with paired original/replica thumbnails and the corresponding evidence artifacts, thereby making the typical error patterns visible and reviewable rather than merely asserted.

Error analysis is reported as court-facing failure modes rather than as generic model weaknesses. The analysis identifies (i) style-only confounds where texture and palette similarity inflate severity despite weak structural evidence; (ii) motif-only confounds where semantic similarity is high but composition is independent; (iii) "generic composition" cases where common layouts cause false alignment; and (iv) heavy abstraction cases where structure representations are unstable and the system should abstain. These failure modes are illustrated with representative examples and accompanied by guidance on how the evidence package should be interpreted when such patterns appear. The goal is to show not only that the system works on average, but that it fails in predictable, reportable ways that support responsible deployment.

Case-Linked Demonstrations and Practical Usefulness

This section demonstrates how the proposed PDI system functions in workflows that resemble real disputes, while maintaining the boundary between measurement and adjudication. The demonstrations are "case-linked" in the narrow sense that they are organized around the kinds of questions that appear in pleadings, expert reports, and platform review processes, and in the broader sense that the outputs are presented in formats that a non-technical decision-maker can inspect. The objective is not to claim that PDI decides infringement, but to show that it can standardize how visual copying evidence is assembled, communicated, and stress-tested across GAN and diffusion replicas.

Mapping Legal Questions to PDI Outputs (Evidence-To-Claim Alignment)

A practical dispute typically involves multiple claims that require different kinds of support, and a court-informed measurement system must therefore be explicit about what each output can and cannot speak to. The paper operationalizes this alignment by mapping recurrent legal and quasi-legal questions to concrete PDI artifacts. Accordingly, PDI should be read as an evidence-to-claim mapping instrument rather than an infringement decision-maker. Table 10 summarizes how common claim types in plagiarism disputes map

Copyright© Choi B.

onto PDI outputs that are reproducible, inspectable, and communicable to non-technical decision-makers, while also making explicit which questions remain outside the scope of image-based measurement. To illustrate how this mapping works end to end, the next examples walk through two representative scenarios—first, a high-similarity candidate set where localized correspondences support a strong copying hypothesis, and second, a confound-prone case where style or motif overlap produces superficial resemblance but the evidence package triggers uncertainty-aware caution.

Claim / Dispute Question (What the Decision-Maker Asks)	What PDI Can Output (Evidence Artifacts)	How to Read it (Measurement Use)	Explicit Out-of-Scope (What PDI Does NOT Decide)
"This output copied protectable expression from the original." (substantial similarity / expressive overlap)	PDI score + CI; composition alignment; patch correspondences; localized similarity map; short evidence summary	Use as structured comparison of *where* overlap appears and how strong it is, with uncertainty for borderline cases	Legal conclusion of infringement; protectability determinations (idea vs expression); jurisdiction-specific tests
"The similarity is only style/genre, not copying." (style-only defense)	Decomposed signals (structure vs texture vs semantics); negative-control behavior; failure-mode flags ("style-only confound")	Use as confound check: does resemblance collapse when structure evidence is absent?	Determining whether "style" is unprotectable in the concrete case; fair use / free use defenses
"The defendant used my work as a reference / prompt / conditioning input." (reference-based generation)	Evidence package showing spatially coherent correspondences; robustness under common edits (crop/ resize/compression)	Use as reproducible support for reference-like reuse hypotheses	Proof of access, actual workflow, intent, or the fact of reference use (requires discovery / testimony / logs)
"The model memorized my work and reproduced it." (memorization / near-duplicate)	Very high PDI with narrow CI; near-duplicate flags; patch matches covering large area; strong alignment consistency across minor transforms	Use as screening and prioritization of likely near-duplicates	Whether memorization occurred during training; causation; training data provenance; admissibility of technical inference
"The similarity is due to common subject matter." (motif overlap)	Semantic motif cues + structure checks; negative controls (semantic overlap but distinct composition)	Use to separate motif-only similarity from compositional reuse (when possible)	Whether shared subject matter is legally actionable; originality thresholds
"This is a derivative work / adaptation."	Comparative evidence emphasizing reuse patterns; optional region-based breakdown (foreground/background)	Use to standardize *degree and localization* of reuse	Legal classification as derivative work; licensing and authorization issues
"The platform should remove/keep the content." (platform moderation / notice-and-takedown context)	Batch ranking by PDI + uncertainty flags; report snapshots for top candidates; audit-ready run manifest	Use for triage + explainability (what triggered concern)	Content removal decision as a legal or policy judgment; DMCA/DSA compliance determinations
"Damages should reflect degree of copying."	PDI bands with cautions; evidence package for expert discussion	Use as one input for expert assessment of severity	Damages calculation; statutory schemes; remedies; willfulness

Table 10: Claim types and the corresponding PDI outputs (and explicit out-of-scope items).

Notes: PDI outputs are designed for measurement, not adjudication. Out-of-scope items remain essential for litigation (access, licensing, defenses, protectability, and jurisdiction-specific doctrine), and must be established through legal standards and admissible evidence beyond image-to-image comparison.

Copyright© Choi B.

When the question concerns whether a specific candidate appears to reuse protectable structure from an asserted original, the most relevant outputs are the composition alignment overlays, region-level correspondences, and robustness checks that show whether the resemblance persists under routine transformations. When the question concerns whether the resemblance is driven largely by generic style or by shared subject matter, the relevant outputs are the signal breakdown and the negative-control flags that indicate a likely confound pattern. When the question concerns how strong the resemblance is relative to other candidates, the relevant output is the ranked portfolio view with uncertainty bands, which supports prioritization and comparative discussion rather than categorical labeling.

This mapping is presented as an "evidence-to-claim" table that makes the scope boundary visible by design. Questions about access, authorization, licensing, intent, and jurisdiction-specific defenses are explicitly listed as out-of-scope, not because they are unimportant, but because they are not inferable from image-level evidence alone. In this way, the system's outputs are tied to what they measure, and the report format discourages rhetorical leaps from a number to a legal conclusion. Table 10 summarizes how recurring claim types in plagiarism disputes map onto specific PDI outputs—scores, confidence bounds, and reviewable evidence artifacts—while also stating, in parallel, which issues remain explicitly out of scope for pixel-based measurement. Building on this mapping, the next examples demonstrate how an asserted original and a candidate set produce an end-to-end ranking and a compact evidence report that can be inspected and challenged by reviewers.

End-To-End Examples: Original–Replica Ranking and Report Snapshots

The demonstrations are organized as end-to-end workflows that begin with an asserted original and a large candidate set and end with a ranked list and a compact report packet for the highest-risk items. In a platform-style screening demonstration, the system processes one original against a candidate pool that includes both GAN and diffusion replicas, returns the top-ranked subset, and highlights gray-zone items where confidence intervals overlap and signals disagree. The emphasis is on triage behavior: the system should elevate candidates that show structural reuse with stable correspondences, while avoiding inflation for cases dominated by style-only similarity. Figure 7 previews the two end-to-end artifacts used throughout this section: a portfolio-level ranking with uncertainty flags and a condensed, reviewable evidence report that highlights composition alignment and local patch correspondences.

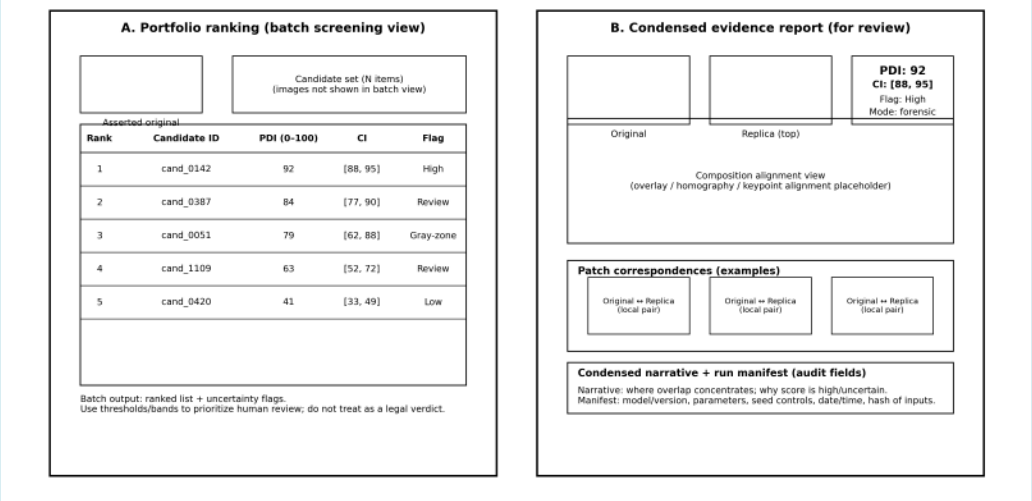

Figure 7: End-to-end demonstrations. (a) Portfolio ranking with uncertainty flags; (b) condensed report snapshot with alignment and patch evidence.

A second demonstration follows a forensic-style workflow in which a small number of candidates are selected and analyzed in depth. For each candidate, the paper shows a report snapshot containing the score and interval, the composition alignment overlay, the top region correspondences, and a short "why this score" narrative generated from the system's structured evidence fields. Each snapshot is designed so that a reader can test the claim visually, rather than being asked to trust the metric. To emphasize reproducibility, each example includes a run-manifest excerpt showing the software version, locked parameters, and the transformation suite used for robustness checks.

Copyright© Choi B.

What PDI Can Standardize in Practice—and What it Cannot

The demonstrations motivate a pragmatic claim: PDI can standardize the presentation and comparison of visual similarity evidence, but it cannot standardize legal outcomes. In practice, the system can impose disciplined consistency on how candidates are selected for review, how severity is described on a stable scale, how uncertainty is disclosed, and how "why" evidence is packaged so that disagreement can be focused on specific visual correspondences rather than on vague impressions. It can also standardize failure-mode disclosure by flagging confounds that repeatedly arise in disputes, such as semantic overlap without structural reuse or high texture similarity that is plausibly attributable to style conventions.

At the same time, the system cannot answer questions that require factual inquiry beyond the image, including whether a model had access to the work, whether a license exists, whether the plaintiff's protectable expression is the source of the overlap, or whether an asserted similarity concerns an unprotectable idea or style. The system also cannot eliminate interpretive disagreement about what counts as substantial similarity in a given jurisdiction. For these reasons, the system is framed as a measurement instrument that supports review, prioritization, and structured expert discussion, while explicitly refusing the role of adjudicator.

Implications for Creators, Platforms, and Expert Reviewers

For creators, the practical value of a severity index lies in clarifying the landscape of potential copying without requiring them to rely on anecdotal comparisons or purely qualitative claims. A calibrated score and evidence packet can help creators triage which candidates warrant attention, document how resemblance manifests, and communicate concerns in a reproducible format that is less dependent on rhetoric. At the same time, creators benefit when the tool is conservative about uncertainty and style confounds, because over-inclusive scoring would generate noise and encourage misdirected accusations.

For platforms, the central benefit is operational: a standardized ranking and reporting format can reduce review costs and improve consistency across moderators, while the gray-zone policy can prevent automated overreach by pushing ambiguous cases into human deliberation. Platform deployment also raises misuse risks, and the paper therefore emphasizes transparency, versioning, and explicit disclaimers so that a numeric score does not become a proxy for liability determinations.

For expert reviewers, the benefit is methodological. A structured evidence package can make expert testimony more testable by tying claims to reproducible artifacts, and it can support cross-case comparability by stabilizing what "severity" means across GAN and diffusion replicas. The tool does not replace expertise, but it can shift expert work from assembling ad hoc visual comparisons toward interpreting standardized evidence under the relevant doctrinal tests.

Discussion

This discussion clarifies how the proposed PDI should be interpreted and used when it is deployed in review workflows that are adjacent to legal disputes but do not themselves adjudicate liability. The central claim is that a calibrated 0–100 score can reduce ambiguity and standardize evidentiary communication, but only if readers are guided away from treating the number as a legal conclusion and toward treating it as a structured measurement with uncertainty and failure modes.

Interpreting the 0–100 Scale (Recommended Bands and Cautions)

A 0–100 scale is attractive because it invites intuitive comparison, yet that same intuition creates predictable risks when users treat score differences as if they were linear and universally comparable. For that reason, the paper recommends interpreting PDI through bands rather than through point estimates, and it recommends attaching explicit cautions that remind readers that the score is calibrated to human-grounded severity within a benchmark, not to any statutory threshold of infringement.

A practical banding scheme can be expressed as a reporting convention rather than as a rule of decision. Scores in a low band (e.g., 0–20) should be read as weak evidence of copying severity, where resemblance is likely driven by generic style, common motifs, or broad subject matter and where the system's own evidence package typically shows limited structural correspondences. Scores in a moderate band (e.g., 20–60) should be read as potentially meaningful resemblance that requires careful inspection of the evidence artifacts, because this band often contains both true partial reuse and false positives caused by confounds. Scores in a high band (e.g., 60–85) should be read as strong structural alignment or repeated local correspondences that are difficult to explain by chance, while still requiring attention to uncertainty, negative-control patterns, and robustness under transformations. Scores in a very high band (e.g., 85–100) should be read as near-duplicate behavior under the measurement protocol, where the evidence artifacts usually show both global alignment and strong region-level correspondences, but where the report should still avoid any

Copyright© Choi B.

language that collapses measurement into adjudication.

The same caution applies to borderline cases, because the most contentious disputes often lie precisely where similarity is plausible but not decisive. The recommended practice is to treat the confidence interval as part of the score, not as an optional supplement, and to use the gray-zone flag as a conservative brake when the system's signals disagree or when the benchmark indicates a known confound pattern. In this framing, the score is a standardized invitation to inspect evidence, not a substitute for expert judgment.

PDI Band (0–100)	Practical Interpretation (Review-Oriented)	Interpretation Cautions (Do Not Over-Claim)	Required Report Elements (Must Show)	Recommended Action
0–20 (Minimal)	Little evidence of copying beyond generic resemblance	Low score does not prove "no influence" (subject/motif/style can still be shared)	Score + CI; brief note on dominant signal(s) (e.g., low structure + low patch matches)	Usually no action; keep for audit sampling
21–40 (Low)	Weak overlap; may reflect shared style, motif, or common templates	Watch for style-only or motif-only confounds; treat as non-dispositive	Score + CI; signal decomposition (structure/texture/semantics); 1–2 patch examples if any	Optional spot-check; flag if high-stakes context
41–60 (Moderate / Review)	Mixed signals; potential partial reuse or ambiguous similarity	Many cases here are borderline; avoid binary language; uncertainty matters	Score + CI; composition alignment; patch correspondences (≥3); short narrative stating where overlap concentrates	Human review recommended; compare against negatives
61–80 (High / Concerning)	Strong evidence of reuse patterns; overlap likely not explainable by style alone	Still not a legal conclusion; consider transformations and alternative explanations	Score + CI; alignment view; patch set (≥5); robustness check summary (crop/resize/compress); run manifest	Escalate to forensic-style report; expert review
81–100 (Very high / Near-duplicate risk)	Very strong copying signature; candidate resembles an "almost same" reproduction	Verify not a dataset duplicate or evaluation artifact; confirm robustness under minor edits	Score + CI (narrow expected); near-duplicate indicators; dense patch coverage; transformation robustness; full manifest	Priority handling; full evidence package; strict audit trail
Gray-zone / Abstain *(rule-based, not a score range)*	System cannot decide reliably (e.g., CI overlaps multiple bands or confound detected)	Abstention is a feature; do not force a verdict	Explicit "cannot decide" flag; CI + reasons; confound tag; recommended next checks	Route to manual/expert review; gather non-pixel evidence

Table 11: Recommended score bands, interpretation cautions, and required report elements per band.

Bands are operational guidance for review and communication; they are not legal thresholds and should not be used as stand-alone proof.

Limitations and Threats to Validity (Domain Shift, Normative Gaps)

The most important limitation is that PDI's meaning depends on the benchmark's coverage, because calibration cannot guarantee stability under severe domain shift. When the distribution of originals changes—through new genres, new cultural styles, new levels of abstraction, or new media artifacts—both the human-grounded severity construct and the model's feature representations can shift in ways that reduce validity. This limitation is especially salient for contemporary AI systems, because generation pipelines evolve quickly and may introduce new failure modes that were not present during calibration.

A second threat arises from normative gaps between perceptual resemblance and legally salient similarity. Even when PDI correctly measures visual copying severity as humans perceive it, legal analysis may turn on doctrine-specific distinctions that the system cannot represent,

Copyright© Choi B.

including the idea–expression boundary, scènes à faire, and jurisdiction-specific tests for substantial similarity. The method's "measurement, not adjudication" posture is therefore not merely a disclaimer; it is an acknowledgement that a technical scale cannot bridge doctrinal thresholds without importing legal judgment that the system is not authorized to make [36].

A third limitation concerns causal interpretation. A high severity score does not establish how the replica was produced, whether the original was in the training data, whether the similarity reflects memorization, whether it reflects a prompt- and reference-driven replication process, or whether it reflects external reuse by a user. This limitation matters because factual theories of copying can influence both litigation posture and policy debates, and a measurement system that speaks only to outputs should be careful not to imply provenance claims it cannot observe.

Ethics and Governance (Rights, Privacy, Chilling Effects)

The ethical posture of a plagiarism severity tool must balance two risks that pull in opposite directions. If the tool is deployed aggressively and interpreted as a verdict, it can chill legitimate creativity and lead to over-removal by platforms, especially in ambiguous mid-range cases where style similarity and shared motifs are common. If the tool is under-sensitive or is framed as purely academic, it can fail to serve creators who need scalable methods to triage plausible copying and to document evidence in a reproducible way.

Rights governance is therefore treated as part of system design. The paper recommends that deployments include clear documentation of dataset provenance and licensing, and that they avoid embedding or distributing proprietary originals unless permissions are clear. Privacy governance also matters when the "paintings" include personal photographs, commissioned portraits, or images that contain identifying information, because similarity search and evidence packaging can expose sensitive content or facilitate unwanted tracking. A conservative policy should require access controls, audit logs, and minimization of retained artifacts, especially for batch screening in platform contexts [37].

Finally, misuse mitigation should be explicit. A severity score can be weaponized to harass creators, to launder weak accusations into numerical form, or to claim "proof" where the system in fact reports uncertainty. The recommended safeguard is to bind the score tightly to its evidence package and uncertainty disclosures, and to require the report to present limitations in the same visual field as the headline number so that the output is harder to quote out of context [38].

Future Work (Provenance Signals, Broader Model Families, Policy Alignment)

Several research directions follow directly from the limitations above. One direction is to integrate provenance-sensitive signals that remain ethically and legally appropriate, such as model-side watermark indicators where available, robust perceptual hashes for traceability, or audit-friendly generation logs when the pipeline is cooperative. These signals would not replace output-based evidence, but they could reduce ambiguity in settings where the factual theory of copying is contested.

A second direction is to expand beyond the GAN–diffusion dichotomy to include broader model families and modalities, including hybrid pipelines, multi-stage editing systems, and toolchains that combine generation with inpainting, style transfer, and compositing. Such expansions will likely require modular calibration strategies and continual validation, because the meaning of severity can drift when new manipulation affordances become common.

A third direction concerns policy alignment. If courts, platforms, and regulators increasingly rely on technical evidence, then standardized reporting conventions and transparency requirements could become part of governance best practices. The paper's contribution is therefore not only a score, but a proposal for how measurement outputs can be presented responsibly in high-stakes contexts, with uncertainty disclosure and reproducibility as first-class requirements.

Conclusion

This paper set out to move plagiarism assessment for AI paintings from binary, rhetoric-driven claims toward a calibrated and reviewable notion of copying severity. The central contribution is a court-informed measurement approach that assigns a 0–100 Plagiarism Degree Index (PDI) score to an asserted original–replica pair, reports uncertainty rather than forced certainty, and produces an evidence package that allows non-technical reviewers to see why the score is high or low.

Summary of Findings

Across the benchmark and validation suite, the results support three main findings. First, multi-signal modeling is necessary because visually persuasive copying can manifest as structural reuse with repainted surfaces, surface/texture mimicry with altered layout, or motif-level proximity that lacks protectable compositional overlap, and no single metric behaves reliably across these cases. Second, calibration to human-grounded severity is not an ornamental rescaling but

Copyright© Choi B.

the mechanism that makes the 0–100 score interpretable as a stable measurement rather than a model-family-specific similarity value. Third, uncertainty disclosure and abstention logic are indispensable for responsible use, because the most practically consequential cases are borderline pairs where signals disagree and where overconfident numbers invite overclaiming.

Deployment Takeaway: Conditions for PDI to Serve as a Key Metric

PDI can function as a key metric in operational settings only under conditions that preserve its status as measurement rather than adjudication. The first condition is that the score must always travel with its uncertainty bounds and gray-zone guidance so that users are encouraged to review ambiguous cases instead of converting mid-range scores into categorical claims. The second condition is that outputs must be reproducible and auditable, with versioned components, locked parameters, and a run manifest that prevents silent protocol drift. The third condition is that the evidence package must be treated as a first-class output—composition alignment, patch correspondences, semantic cues, and robustness checks—because a score without inspectable "why" is difficult to contest and easy to misuse. The fourth condition is continuous validation under domain shift, because changes in genre coverage, editing practices, and model families can alter what similarity means and can quietly erode calibration fidelity.

Under these conditions, PDI becomes most useful as a standard for prioritization and structured comparison: it can help creators and platforms triage large candidate sets, and it can help expert reviewers communicate evidentiary reasoning in a format that is inspectable and testable, while remaining explicit that the tool does not decide infringement.

Closing Remarks

The broader implication is that responsible technical work in this area is not a competition to produce the most decisive number, but an effort to produce the most reviewable evidence. A court-informed severity index does not replace legal judgment, but it can reduce the noise of unstructured comparisons by stabilizing what is measured, how uncertainty is presented, and how evidence is packaged for scrutiny. When the output is framed and governed as measurement, PDI can contribute to more transparent dispute resolution and more consistent platform practices without collapsing the distinction between quantified resemblance and legal liability.

References

1. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B (2021) High-Resolution Image Synthesis with Latent Diffusion Models. arXiv.
2. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: From error visibility to structural similarity. IEEE Trans Image Process 13(4): 600–612.
3. Doshi-Velez F, Kim B (2017) Toward a Rigorous Science of Interpretable Machine Learning. arXiv.
4. Angelopoulos AN, Bates S (2021) A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. arXiv.
5. Efron B (1979) Bootstrap methods: Another look at the jackknife. Ann Stat 7(1): 1-26.
6. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, et al. (2021) Datasheets for Datasets. Commun ACM 64(12): 86-92.
7. Getty Images v Stability AI (UK High Court proceedings; trademark/watermark issues and narrowed copyright posture reported).
8. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, et al. (2014) Generative Adversarial Nets. arXiv.
9. Ho J, Jain A, Abbeel P (2020) Denoising Diffusion Probabilistic Models. arXiv.
10. Hertz A, Mokady R, Tenenbaum J, Aberman K, Pica A, et al. (2022) Prompt-to-Prompt Image Editing with Cross Attention Control. arXiv.
11. Zhang L, Rao A, Agrawala M (2023) Adding Conditional Control to Text-to-Image Diffusion Models. arXiv.
12. Bradley RA, Terry ME (1952) Rank analysis of incomplete block designs: I. The method of paired comparisons. Biometrika 39(3-4): 324-345.
13. Luce RD (1959) Individual Choice Behavior: A Theoretical Analysis. Journal of the Royal Statistical Society 123(4).
14. Canny J (1986) A Computational Approach to Edge Detection. IEEE Trans Pattern Anal Mach Intell 8(6): 679–698.
15. McCullagh P (1980) Regression Models for Ordinal Data. J R Stat Soc Series B 42(2): 109–142.
16. Vovk V, Gammerman A, Shafer G (2005) Algorithmic

Copyright© Choi B.

Learning in a Random World. Springer, USA.

17. Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1: 206–215.

18. Gatys LA, Ecker AS, Bethge M (2015) A Neural Algorithm of Artistic Style. arXiv.

19. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3: 160018.

20. Spearman C (1904) The proof and measurement of association between two things. Am J Psychol 15(1): 72–101.

21. Kendall MG (1938) A New Measure of Rank Correlation. Biometrika 30(1-2): 81–93.

22. U.S. litigation trend note (output-comparison–centered proof in generative-AI disputes).

23. Zhang R, Isola P, Efros AA, Shechtman E, Wang O (2018) The Unreasonable Effectiveness of Deep Features as a Perceptual Metric.

24. Andersen v. Stability AI Inc., N.D. Cal. Case No. 3:23-cv-00201.

25. OpenAI copyright litigation (consolidated/related U.S. suits).

26. Zadrozny B, Elkan C (2002) Transforming classifier scores into accurate multiclass probability estimates.

27. ACM Artifact Review and Badging (policy/guide).

28. Stodden V, Seiler J, Ma Z (2018) An empirical analysis of journal policy effectiveness for computational reproducibility. Proc Natl Acad Sci USA. 115(11): 2584-2589.

29. Preston-Werner T. Semantic Versioning 2.0.0.

30. Mitchell M, Wu S, Zaldivar A, et al. (2019) Model Cards for Model Reporting. FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp: 220-229.

31. Claerbout JF, Karrenbach M (1992) Electronic documents give reproducible research a new meaning.

32. Wang Z, Simoncelli EP, Bovik AC (2003) Multi-scale structural similarity for image quality assessment.

33. Huynh-Thu Q, Ghanbari M (2008) Scope of validity of PSNR in image/video quality assessment. Electron Lett 44(13): 800–801.

34. Chen T, Kornblith S, Norouzi M, Hinton G (2020) A Simple Framework for Contrastive Learning of Visual Representations. arXiv.

35. 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.

36. Wallace GK (1991) The JPEG still picture compression standard. Commun ACM 34(4): 30–44.

37. Responsible deployment and auditability discussion in generative-AI copyright disputes.

38. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, et al. (2021) Datasheets for datasets. Commun ACM 64(12): 86-92.

Copyright© Choi B.