Detecting Metadata-Conditioned Response Drift in AI-Generated Visual Information Systems: A Two-Phase Human–Computer Interaction Framework

Yu Jia[1], Shaojie Zhang[2]*, Na Li[3]
College of Digital Creativity, Henan Technical College of Construction, Zhengzhou, Henan, China[1, 3]
College of Art and Design, Henan Institute of Economics and Trade, Zhengzhou, Henan, China[2]
Faculty of Education, Henan University, Kaifeng, Henan, China[1, 2]

***Abstract*—AI-generated visual information systems often present images with authorship metadata, yet such labels may shape the evaluation data they are intended to contextualize. Existing work has seldom offered a system-level procedure for measuring metadata-conditioned response drift under unchanged visual input. This study develops a two-phase evaluation framework combining blind presentation, attribution-label overlay, trial-level logging, drift computation, and mixed-effects validation. Ninety-two participants rated four AI-generated images first without labels and then with attribution labels, producing 732 valid observations. Linear mixed-effects modeling, supported by ordinal sensitivity analysis and leave-one-image-out checks, identified a significant Phase × Label interaction within the implemented attribution-label assignment scheme. Under this assignment scheme, ratings for stimuli assigned to the human-labeled condition increased after disclosure, whereas ratings for stimuli assigned to the AI-labeled condition decreased. These findings are interpreted as evidence that the proposed framework can detect disclosure-associated response drift, rather than as a fully counterbalanced causal estimate of pure label effects.**

***Keywords*—AI-generated visual information systems; metadata disclosure; response drift; user-response logs; human-computer interaction; authorship labeling; mixed-effects modeling**

I. Introduction

AI-generated visual content is increasingly integrated into interactive visual information systems, including online galleries, recommendation platforms, digital design tools, educational interfaces, and content-evaluation environments [1], [2]. In these systems, images are rarely encountered as isolated visual objects. They are commonly accompanied by metadata indicating human authorship, AI authorship, AI assistance, or the use of a specific generative model [3], [4]. Although these cues are typically introduced to support transparency and provenance disclosure, once embedded in an interface they also become part of the context in which users form ratings, preferences, and quality judgments.

The implications extend beyond individual perception because many AI-mediated visual systems use user responses as operational inputs. Ratings, preference scores, likes, and evaluation logs may inform recommendation, ranking, content moderation, design feedback, or subsequent model evaluation [5]. If authorship metadata alters these responses, ratings collected under different disclosure conditions may not be directly comparable. Recent research on AI-generated media labeling confirms that disclosure cues can affect how users interpret and respond to visual content [6]. When blind and labeled ratings are pooled without retaining their disclosure state, downstream systems may learn patterns that partly reflect attribution framing rather than visual preference alone. Authorship metadata is therefore relevant not only to provenance communication but also to the interpretation of user-response data.

Evidence from AI-art evaluation and authorship attribution research suggests that source attribution can shape user judgment, yet the implications of these attribution effects for system-generated response logs remain insufficiently developed [7-9]. Existing studies have generally examined whether evaluation outcomes differ across attribution conditions rather than how metadata-conditioned changes should be recorded, quantified, and validated when visual input remains unchanged. This leaves a measurement problem: a rating may represent both an evaluation of the image and a response to the interface context in which the image was presented.

We develop and evaluate a two-phase framework for detecting metadata-conditioned response drift in AI-generated visual information systems. Participants first evaluate AI-generated images without authorship information and subsequently re-evaluate the same images with experimentally assigned attribution labels. Trial-level records are matched by participant and image, allowing disclosure-associated rating changes to be computed and statistically evaluated. Because attribution labels were assigned to fixed stimuli, the implementation is used to verify the detection procedure under the specified assignment scheme rather than to estimate a fully counterbalanced, stimulus-independent causal effect. The main contributions of this study are as follows:

- A two-phase evaluation procedure is introduced for examining response changes while holding the visual input constant across blind and disclosure conditions.
- Disclosure-associated change is formalized as a participant–image-level drift measure supported by explicit logging of phase, image identity, attribution condition, and metadata visibility.

*Corresponding author

- The response pattern is assessed using a linear mixed-effects model, a cumulative link mixed model for ordinal sensitivity, and leave-one-image-out robustness analyses.
- Analysis of 92 participants and 732 valid observations identifies a disclosure-associated pattern that remains consistent across the primary mixed-effects model, the ordinal sensitivity analysis, and the leave-one-image-out checks, illustrating the value of retaining metadata-display conditions when ratings are compared or reused as system inputs.

II. Background and Related Work

Research on AI-generated art has shown that visual judgments are shaped by more than the perceptual properties of an image. Early work on computer-generated art examined whether viewers evaluated computer-generated and human-made artworks differently [10]. Subsequent studies demonstrated that knowledge or prior assignment of AI authorship could alter aesthetic appreciation and evaluations of artistic value [7], [11]. More recent experiments have linked differences between human- and AI-attributed works to perceived authenticity, creative involvement, and anthropocentric preferences [8],[9],[12]. The direction of these effects is not necessarily uniform: broader research on algorithm aversion and algorithm appreciation indicates that responses to algorithmic systems depend on the task, prior experience, and evaluative context [13],[14]. Taken together, these studies establish authorship attribution as a consequential source cue, but their primary object of analysis remains the judgment outcome—whether a work receives higher ratings for aesthetic quality, creativity, authenticity, or artistic value under a particular attribution.

Research on synthetic-content disclosure has shifted attention from authorship beliefs alone to the design and placement of labels within digital interfaces. Wittenberg et al. found that label effects depended partly on whether the label described the production process or warned of potentially misleading content; labels affected belief in claims associated with AI-generated images, whereas a simple AI-generated label had more limited effects on intended engagement [6]. Gamage et al. further showed that responses to synthetic-content warnings vary with wording, positioning, visual presentation, and level of detail [15]. Recent findings also caution against assuming that disclosure produces a uniform penalty. Li et al. reported no overall main effect of AIGC labeling on perceived credibility, although the effect varied with prior experience and content category [16]. Gallegos et al. similarly found that labeling messages as AI-generated did not significantly reduce their persuasive effects [17]. Recent evidence further suggests that disclosure timing and the level of AI involvement may influence user engagement with labeled content [18]. Disclosure is therefore neither uniformly beneficial nor uniformly detrimental; its consequences depend on the labeled content, the presentation of the label, the outcome being measured, and the characteristics of the audience. This context dependence is consistent with human–AI interaction research, which treats interface presentation as part of interaction design rather than as a neutral wrapper around the underlying system [19].

Less well developed is an operational account of how attribution metadata conditions the generation and subsequent interpretation of user-response logs. Randomized between-subject experiments have examined the effects of AI-related labels on belief, trust, engagement intentions, and persuasion [6],[15],[17]. A recent mixed experiment examined perceived credibility under variations in label presence, content category, and prior experience [16]. These studies show that labels can shape evaluation, but they seldom formalize the resulting change as a measurable property of the response record itself. This distinction matters for systems that reuse ratings as recommendation, ranking, or evaluation signals: responses collected under different metadata conditions may be valid within their original experimental settings yet remain non-equivalent when pooled as system inputs. The unresolved issue is therefore not simply whether disclosure affects judgment, but how disclosure-associated variation can be retained and quantified within user-response records. Accordingly, metadata exposure is treated here as a logged evaluation condition, with response drift defined at the participant–image level under unchanged visual input.

III. Proposed Metadata-Conditioned Response Drift Evaluation Framework

A. Framework Overview

This study proposes a two-phase evaluation framework for measuring metadata-conditioned response drift in AI-generated visual information systems. The framework is designed to quantify changes in user-response data associated with authorship-related interface cues while holding the visual stimulus constant. Rather than treating the study as a general comparison between human-created and AI-generated art, the proposed framework examines assigned authorship information as a configurable interface variable that may influence rating behavior and evaluation-log interpretation.

The framework consists of three functional layers: the visual stimulus layer, the interface control and metadata layer, and the response drift analysis layer. The visual stimulus layer provides a controlled set of AI-generated images and maintains stimulus consistency across evaluation phases. The interface control and metadata layer manages phase transition, stimulus presentation order, attribution-label display, and response collection. The response drift analysis layer links blind-phase and disclosure-phase ratings at the participant–image level and estimates the direction and magnitude of rating change after metadata disclosure.

The evaluation procedure contains two sequential phases completed by the same participants. In the blind phase, participants rate AI-generated images without any authorship information. In the disclosure phase, the same images are re-presented with assigned authorship labels, either human-labeled or AI-labeled. The attribution condition is implemented at the stimulus/trial level through the interface rather than embedded in the image file, caption, or visual content. This design preserves a clear distinction between the visual stimulus and the attribution cue while enabling response changes to be

measured without altering the image content. The implementation is therefore treated as a framework-verification setting: it detects disclosure-associated response drift under the implemented label-assignment scheme, while causal claims about pure authorship-label effects independent of image identity are intentionally limited. The overall workflow of the proposed framework is shown in Fig. 1.

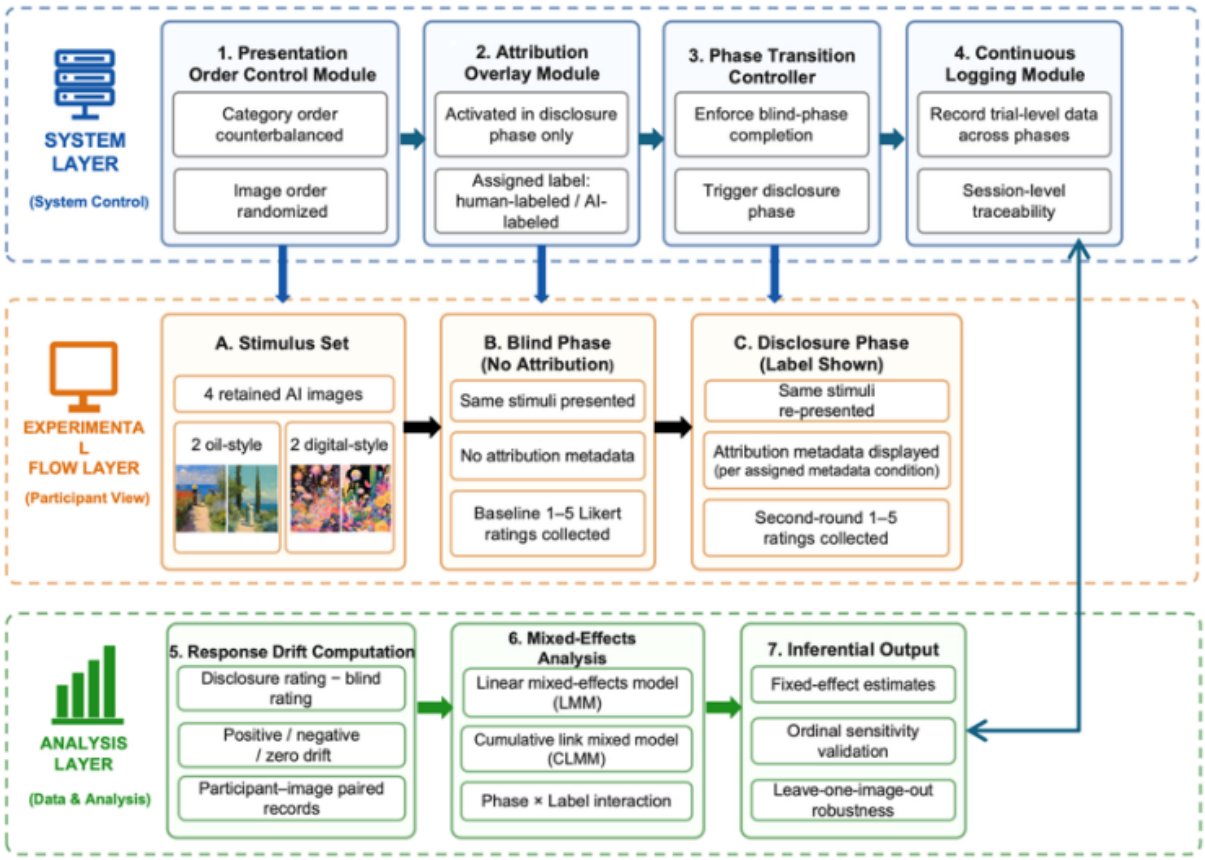

Fig. 1. Overall workflow of the proposed framework, including stimulus presentation, attribution metadata display, trial-level response logging, and response drift validation.

The primary output of the framework is response drift, defined as the difference between the disclosure-phase rating and the corresponding blind-phase rating for the same participant and image. Positive values indicate an upward shift after metadata disclosure, whereas negative values indicate a downward shift. The confirmatory effect of interest is the Phase × Label interaction, which tests whether rating changes from the blind phase to the disclosure phase differ between images assigned to the human-labeled and AI-labeled conditions.

This positioning is consistent with the objective of evaluating how attribution cues enter user-response logs as measurement-context variables in AI-generated visual information systems. The framework is intended to support disclosure-aware response logging and interface evaluation, rather than to generalize user preference for human- or AI-attributed images.

B. AI-Generated Visual Stimulus Module

The visual stimulus module was designed to provide controlled AI-generated image inputs for evaluating metadata-conditioned response drift. All stimuli were generated using the Midjourney text-to-image model through the Niji-Journey interface. Two rendering categories were constructed: oil-style compositions and digital-style compositions. These two categories were selected to represent visually distinct rendering modes while preserving a manageable experimental structure for repeated evaluation.

For each rendering category, a standardized prompt template was used to constrain the main visual structure, compositional layout, color distribution, subject arrangement, and rendering style. The oil-style template specified a painterly outdoor scene with a clear central visual element, balanced composition, soft color transitions, and an oil-painting rendering style. The digital-style template specified a colorful digital visual scene with a clear central visual element, balanced composition, vivid color distribution, and a digital-illustration rendering style. The templates were used to generate four candidate images for each category. Two images from each category were then retained, resulting in a final stimulus set of four AI-generated images. The selected stimuli were newly generated for the present study and had not been previously shown to participants.

The screening procedure was used to improve comparability across stimuli before the attribution-label conditions were applied. Candidate images were evaluated according to the following criteria: comparable visual complexity, clear thematic legibility, absence of salient rendering errors, similar prominence of the central visual element, and adequate interpretability at the display size used in the evaluation interface. The retained images were presented in the same interface layout across both phases. Image size, screen position, rating scale format, and background layout were kept constant so that the only interface difference between the blind and disclosure phases was the presence or absence of authorship metadata. This procedure was not intended to make the retained images perceptually identical or to replace full label counterbalancing. Rather, it reduced obvious stimulus-level imbalance and provided a controlled stimulus set for framework verification. Remaining participant-level heterogeneity and stimulus-related variation were addressed through mixed-effects modeling and robustness checks.

Fig. 2. Retained AI-generated visual stimuli used in the two-phase evaluation. O1 and O2 represent oil-style stimuli, whereas D1 and D2 represent digital-style stimuli. All four images were generated by AI and were re-presented across blind and disclosure phases.

The retained oil-style images were assigned image identifiers O1 and O2, and the retained digital-style images were assigned image identifiers D1 and D2 (see Fig. 2). During the disclosure phase, O1 and D1 were displayed with the assigned label Human-labeled, whereas O2 and D2 were displayed with the assigned label AI-labeled. Because all four stimuli were AI-generated, these labels functioned only as experimentally assigned attribution cues within the evaluation interface rather than as indicators of actual authorship.

C. Attribution Metadata Overlay Module

The attribution metadata overlay module was developed to separate visual content from authorship information. In the blind phase, the metadata overlay was inactive, and participants viewed the AI-generated images without any authorship cue. In the disclosure phase, the overlay was activated and displayed the assigned attribution label for each image. The displayed label used one of two interface texts: Human-labeled or AI-labeled.

The authorship label was presented as an interface text label. It was placed in a fixed position adjacent to the image display area and remained visible on the same screen as the image and rating scale throughout the disclosure-phase trial. The label position, font size, and display format were kept constant across labeled trials. This design introduced attribution as a controlled interface variable while preserving the visual stimulus itself.

The metadata overlay was persistent during the labeled evaluation trial, meaning that the assigned label remained visible until participants submitted the second-round rating. This procedural control ensured that attribution information was available during the disclosure-phase response. No post-task manipulation-check item was collected; therefore, label exposure was controlled through the interface procedure rather than through a separate self-report verification item.

D. Phase Transition and Trial-Level Logging Module

The phase transition and trial-level logging module controlled the evaluation sequence and recorded structured response data. The evaluation followed a two-phase repeated-measures procedure and involved 92 participants (mean age = 38.30 years; SD = 13.54; range = 20–72 years). All participants reported normal or corrected-to-normal vision and no known color-vision deficiency. Thirty-eight participants reported at least two years of prior art-related learning or practice experience, whereas fifty-four participants reported no comparable art-training background. In the first phase, participants completed the blind evaluation without authorship information. In the second phase, the same images were re-presented under disclosure conditions with assigned attribution labels.

The interface managed the transition from the blind phase to the disclosure phase and prevented participants from entering the disclosure phase before completing the blind evaluation. A rating response was required before participants could proceed to the next trial. No time limit was imposed on image viewing or rating submission. The order of stimulus categories was counterbalanced across participants, and the presentation order of individual images within each category was randomized. Specifically, participants were assigned to alternating category-order sequences, and image order within each category was randomly shuffled at the session level. This procedure reduced the influence of fixed presentation order while maintaining the intended two-phase structure.

For each trial, the system recorded a structured response entry containing the anonymized participant/session identifier,

image identifier, stimulus category, presentation phase, attribution condition, metadata visibility state, and rating response. These trial-level records provided the data structure required to match blind-phase and disclosure-phase ratings for the same participant and image. Responses were stored without personally identifiable information. The logging module therefore served as the operational basis for computing response drift and for evaluating whether assigned attribution cues were associated with systematic changes in user ratings.

The complete evaluation consisted of eight trials for each participant: four blind-phase trials and four disclosure-phase trials. Because each participant evaluated the same four images twice, observations were repeatedly measured within participants across fixed stimuli. This structure motivated the use of a participant-level mixed-effects model in the validation layer. The complete procedure is summarized in Table I and Algorithm 1.

TABLE I. OPERATIONAL PROCEDURE OF THE PROPOSED EVALUATION FRAMEWORK.

Stage	Module	Operation	Output
1	Visual stimulus module	Generate and screen AI-generated images	Retained image set
2	Blind evaluation	Present images without metadata	Baseline ratings
3	Metadata overlay	Activate human-labeled / AI-labeled display	Disclosure condition
4	Disclosure evaluation	Re-present the same images with metadata	Labeled ratings
5	Logging layer	Match records by participant and image ID	Trial-level paired data
6	Drift computation	Apply Eq. (1)	Response drift scores
7	Validation layer	Fit LMM, CLMM, and robustness checks	Effect estimates and CIs

Algorithm 1: Metadata-Conditioned Response Drift Detection Procedure

Require: AI-generated stimulus set S={O1,O2,D1,D2} ; attribution label set L={Human-labeled, AI-labeled}; metadata visibility states V={Hidden,Visible}; participant set **P**; 5-point rating scale.

Ensure: Paired trial-level response logs, response drift scores, and statistical validation outputs.

start
for each participant $p \in P$ **do**
initialize anonymized participant session SID_p
load the retained AI-generated stimulus set S
assign participant p to one counterbalanced category-order sequence
randomize the presentation order of images within each category
for each image i in the randomized blind-phase sequence **do**
present image i with attribution metadata hidden
collect blind-phase rating $R_{p,i}^{\text{blind}}$
record log entry $\langle SID_p, i, Category_i, Blind, None, Hidden, R_{p,i}^{\text{blind}} \rangle$
end for
activate the attribution metadata overlay module
for each image i in the corresponding disclosure-phase sequence **do**
re-present the same visual stimulus i
display the assigned attribution label $Label_i$
keep $Label_i$ visible until rating submission
collect disclosure-phase rating $R_{p,i}^{\text{disclosure}}$
record log entry $\langle SID_p, i, Category_i, Disclosure, Label_i, Visible, R_{p,i}^{\text{disclosure}} \rangle$
if both blind-phase and disclosure-phase ratings are available **then**
compute response drift $Drift_{p,i} = R_{p,i}^{\text{disclosure}} - R_{p,i}^{\text{blind}}$
end if
end for
end for
aggregate matched participant–image records
fit the primary linear mixed-effects model
conduct ordinal sensitivity using a cumulative link mixed model
perform leave-one-image-out robustness checks
output drift estimates, confidence intervals, and validation results
end

E. Response Drift Computation

The primary output of the framework was response drift, defined as the rating change between the disclosure phase and the blind phase for the same participant and image. Let $R_{p,i}^{\text{blind}}$ denote the blind-phase rating provided by participant p for image i, and let $R_{p,i}^{\text{disclosure}}$ denote the corresponding disclosure-phase rating after authorship metadata was displayed. The response drift score was computed as:

$$\text{Drift}_{p,i} = R_{p,i}^{\text{disclosure}} - R_{p,i}^{\text{blind}} \quad (1)$$

where, $\text{Drift}_{p,i}$ represents the metadata-associated rating shift for participant p and image i. A positive drift value indicates that the participant assigned a higher rating after metadata disclosure than under blind conditions. A negative drift value indicates that the participant assigned a lower rating after disclosure. A drift value of zero indicates no change between the two phases. This computation allowed the framework to transform raw trial-level ratings into an interpretable measure of metadata-associated response change.

The response drift measure was calculated at the participant – image level and then examined by attribution condition and stimulus category. This structure enabled the framework to evaluate whether rating shifts differed between human-labeled and AI-labeled conditions, and whether the observed pattern was stable across oil-style and digital-style images.

F. Statistical and Robustness Validation Layer

The statistical validation layer was designed to estimate disclosure-associated changes in ratings while accounting for repeated observations within participants. Because each participant provided ratings across multiple images and phases, linear mixed-effects modeling was used as the primary inferential approach [20]. Ratings were modeled as a function of presentation phase, attribution label, stimulus category, and art experience.

The primary model was specified as:

$$\text{Rating}_{p,i,t}=\beta_0+\beta_1\text{Phase}_t+\beta_2\text{Label}_i+\beta_3\text{Category}_i$$
$$+\beta_4\text{ArtExperience}_p+\beta_5(\text{Phase}_t\times\text{Label}_i)$$
$$+\beta_6(\text{Phase}_t\times\text{Category}_i)+\beta_7(\text{Label}_i\times\text{Category}_i)$$
$$+\beta_8(\text{Phase}_t\times\text{Label}_i\times\text{Category}_i)+u_p+\varepsilon_{p,i,t} \quad (2)$$

where, $\text{Rating}_{p,i,t}$ refers to the 5-point evaluation score provided by participant p for image i at phase t. Phase_t distinguishes blind and disclosure evaluations, Label_i distinguishes human-labeled and AI-labeled metadata conditions, and Category_i distinguishes oil-style and digital-style stimuli. ArtExperience_p was entered as a participant-level covariate. In this specification, u_p represents the participant-level random intercept, and $\varepsilon_{p,i,t}$ denotes the residual error term. Dummy coding was used, with the blind phase, AI-labeled condition, and oil-style category serving as reference levels.

A more complex candidate model allowing participant-level random slopes for phase and label was also examined:

$$\text{Rating}_{p,i,t}=\beta_0+\beta_1\text{Phase}_t+\beta_2\text{Label}_i+\beta_3\text{Category}_i$$
$$+\beta_4\text{ArtExperience}_p+\beta_5(\text{Phase}_t\times\text{Label}_i)$$
$$+\beta_6(\text{Phase}_t\times\text{Category}_i)+\beta_7(\text{Label}_i\times\text{Category}_i)$$
$$+\beta_8(\text{Phase}_t\times\text{Label}_i\times\text{Category}_i)+u_{0p}$$
$$+u_{1p}\text{Phase}_t+u_{2p}\text{Label}_i+\varepsilon_{p,i,t} \quad (3)$$

When this model produced convergence warnings or singular estimates, the random-effects structure was reduced to a participant-level random intercept. The confirmatory effect of interest was the Phase × Label interaction, which tested whether rating changes from the blind phase to the disclosure phase differed between human-labeled and AI-labeled conditions. Because attribution labels were not fully counterbalanced across specific images, the Label coefficient was interpreted conditionally at the reference levels and was not treated as evidence of a general preference for either label independent of stimulus identity.

Because ratings were collected using a 5-point Likert scale, an ordinal sensitivity analysis was conducted using a cumulative link mixed model with the same fixed-effects structure and a participant-level random intercept [21],[22]. This analysis tested whether the main inferential pattern remained stable when the ordinal nature of the outcome variable was explicitly modeled. Planned contrasts were used to estimate phase-related rating changes within each label and category condition, and follow-up comparisons were adjusted using the Holm method.

Robustness was further examined through four leave-one-image-out analyses. Because omitting one stimulus left one Label × Category combination unobserved, each subset was analyzed using a reduced mixed-effects specification retaining the Phase × Label interaction, with stimulus category and art experience included as covariates and a participant-level random intercept. This analysis assessed whether the interaction was sensitive to any individual stimulus.

IV. Results and Framework Evaluation

A. Data Completeness and Logging Reliability

The proposed framework generated a maximum of 736 possible rating records, corresponding to 92 participants, four images, and two evaluation phases. A total of 732 valid rating observations were recorded and retained for analysis, yielding a valid logging rate of 99.46% and an item-level missing rate of 0.54%. The four missing observations came from one participant and did not alter the repeated-measures structure of the dataset. Analyses were therefore conducted using available observations without imputation.

The high valid logging rate confirmed that the framework produced sufficiently complete participant–image paired records for subsequent response-drift computation and mixed-effects validation.

B. Detection of Disclosure-Associated Response Drift

The primary validation objective was to determine whether the proposed framework could detect a disclosure-associated response-drift pattern under fixed visual input. The confirmatory linear mixed-effects model revealed a significant Phase × Label interaction in the oil-style reference category, β = 0.780, SE = 0.153, 95% CI [0.481, 1.080], p < 0.001. This interaction indicates that rating changes from the blind phase to the disclosure phase differed systematically between human-labeled and AI-labeled conditions.

TABLE II. Key Coefficient Estimates from the Primary and Ordinal Sensitivity Models.

Model	Term	β	SE	95% CI	p
Linear mixed-effects model	Phase × Label	0.780	0.153	[0.481, 1.080]	<0.001
	Label coefficient	-0.022	0.108	[-0.234, 0.190]	0.839
	Phase × Label × Category	0.220	0.216	[-0.203, 0.642]	0.309
	Art Experience	0.018	0.088	[-0.155, 0.191]	0.841
Cumulative link mixed model	Phase × Label	2.195	0.407	[1.398, 2.992]	<0.001
	Label coefficient	-0.051	0.275	[-0.589, 0.488]	0.854
	Phase × Label × Category	0.548	0.569	[-0.567, 1.664]	0.335
	Art Experience	0.102	0.238	[-0.365, 0.569]	0.668

Note. β coefficients from the cumulative link mixed model are estimated on the log-odds scale and are therefore not directly comparable in magnitude to coefficients from the linear mixed-effects model.

Specifically, ratings decreased after disclosure for images assigned to the AI-labeled condition and increased for images assigned to the human-labeled condition. For the oil-style reference category, the blind-phase difference between images assigned to the two label conditions was not significant, β = −0.022, SE = 0.108, 95% CI [−0.234, 0.190], p = 0.839,

indicating no detectable baseline difference within that category. Art experience also showed no reliable association with ratings, β = 0.018, SE = 0.088, 95% CI [−0.155, 0.191], p = 0.841. Together, these results indicate that the framework detected a disclosure-associated response-drift pattern after accounting for participant-level heterogeneity, stimulus category, and art experience, as summarized in Table II and illustrated in Fig. 3.

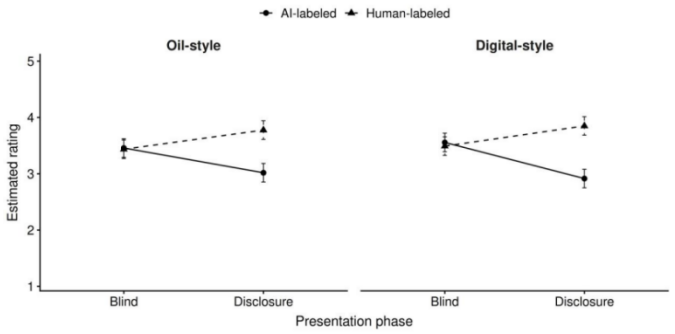

Fig. 3. Disclosure-associated response drift across stimulus categories.

C. Category-Level Stability of the Drift Pattern

Category-level stability was evaluated to determine whether the *detected* response-drift pattern was limited to a single rendering category. The Phase × Label × Category interaction was not significant, β = 0.220, SE = 0.216, 95% CI [-0.203, 0.642], p = 0.309. This result indicates no reliable evidence that the disclosure-associated drift pattern differed between oil-style and digital-style stimuli.

Model-derived follow-up contrasts were consistent with the directional pattern observed in the primary model. In the AI-labeled condition, ratings decreased significantly from the blind phase to the disclosure phase for both the oil-style category, β = −0.440, SE = 0.108, 95% CI [−0.652, −0.227], Holm-adjusted p < 0.001, and the digital-style category, β = −0.641, SE = 0.108, 95% CI [−0.852, −0.430], Holm-adjusted p < 0.001. In contrast, in the human-labeled condition, ratings increased significantly for both the oil-style category, β = 0.341, SE = 0.108, 95% CI [0.128, 0.553], Holm-adjusted p = 0.002, and the digital-style category, β = 0.359, SE = 0.108, 95% CI [0.148, 0.570], Holm-adjusted p = 0.002. These contrasts are shown in Fig. 4.

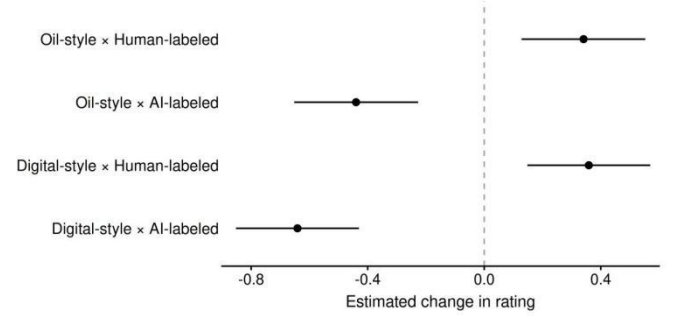

Fig. 4. Model-derived rating change from the blind phase to the disclosure phase. Points indicate estimated changes in rating within each Label × Category condition; horizontal bars indicate 95% confidence intervals. The dashed vertical line marks zero change.

D. Ordinal Sensitivity Validation

Because ratings were collected using a 5-point Likert scale, an ordinal sensitivity validation was conducted using a cumulative link mixed model. The ordinal model produced the same inferential pattern as the primary linear mixed-effects model. The Phase × Label interaction remained significant, β = 2.195, SE = 0.407, 95% CI [1.398, 2.992], p < 0.001. The Label coefficient at the blind-phase, oil-style reference levels was not significant, β = −0.051, SE = 0.275, 95% CI [−0.589, 0.488], p = 0.854. The Phase × Label × Category interaction also did not reach significance, β = 0.548, SE = 0.569, 95% CI [−0.567, 1.664], p = 0.335. Art Experience was also not significant, β = 0.102, SE = 0.238, 95% CI [-0.365, 0.569], p = 0.668. The consistency between the linear mixed-effects model and the cumulative link mixed model indicates that the detected response-drift pattern was not attributable to treating Likert-scale ratings as approximately continuous.

E. Leave-One-Image-Out Robustness Check

Robustness was assessed through four leave-one-image-out analyses. Across the reduced models, the Phase × Label interaction remained significant, with estimated coefficients ranging from 0.789 to 0.991 (all p<0.001). The direction and significance of the interaction were unchanged when any one stimulus was omitted, indicating that the observed response-drift pattern was not driven by a single image.

Participant-level drift estimates showed the same directional pattern. In the AI-labeled condition, mean drift was negative, M = -0.543, SD = 0.623, median = -0.500, with 67 of 92 participants showing negative drift. In the human-labeled condition, mean drift was positive, M = 0.353, SD = 0.557, median = 0.500, with 56 of 92 participants showing positive drift, as shown in Fig. 5.

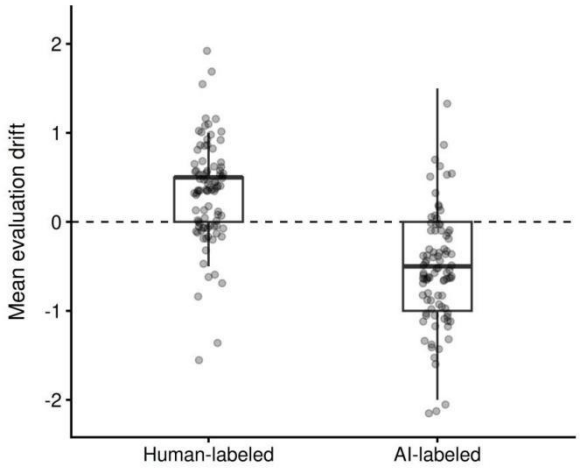

Fig. 5. Participant-level response drift by assigned authorship label. Each point represents one participant's mean response drift within a label condition, computed as the disclosure-phase rating minus the corresponding blind-phase rating. Positive values indicate higher ratings in the disclosure phase than in the blind phase, whereas negative values indicate lower ratings. The dashed horizontal line denotes zero response drift.

Together, the leave-one-image-out and participant-level analyses provide additional support for the stability of the

detected response-drift pattern across the retained stimuli and participants.

V. Discussion and System Implications

A. Metadata as a Response-Shaping Interface Variable

Across the primary mixed-effects model, ordinal sensitivity analysis, planned contrasts, and leave-one-image-out checks, the results converged on the same directional pattern: ratings shifted after authorship metadata was disclosed, while the visual stimuli remained unchanged. This convergence supports the interpretation that metadata disclosure was associated with measurable response drift in the implemented two-phase evaluation framework.

The consistency across analyses follows from complementary features of the design and validation strategy. Matching ratings within the same participant–image pair limits the influence of stable individual differences in scale use, while the participant-level random intercept accounts for dependence across repeated ratings from the same individual. The ordinal model tests whether the inference depends on treating the five-point rating scale as approximately continuous, and the leave-one-image-out analyses examine whether the pattern is driven by any single retained stimulus. Agreement across these analyses therefore indicates that the observed pattern is not tied to one statistical specification or one image, although it does not eliminate the confounding introduced by fixed label–stimulus assignment.

These findings do not establish a general preference for human-created content or a stimulus-independent causal effect of authorship labels. They instead indicate that authorship metadata forms part of the conditions under which user-response data are produced. Ratings collected under different metadata states should therefore not be treated as automatically equivalent when they are compared or reused as system inputs; stronger causal claims require fully counterbalanced label assignments.

B. Implications for User-Response Logs and Downstream Visual Systems

The findings have direct implications for AI-generated visual platforms that use ratings, preference scores, or interaction logs as system inputs. In recommendation, ranking, content evaluation, and design-feedback settings, these responses are often treated as indicators of content quality, relevance, or user preference. The present results show that this assumption is incomplete: a rating collected after authorship metadata disclosure may not be equivalent to a rating collected under blind conditions, even when the visual content remains unchanged.

For this reason, metadata-display conditions should be retained as contextual variables in response logs. Ratings collected under blind, AI-labeled, human-labeled, or other disclosure conditions should not be pooled as fully interchangeable signals. Otherwise, downstream models may learn patterns that partly reflect attribution framing rather than visual preference alone.

The proposed framework provides a practical way to identify this source of variation. By matching blind-phase and disclosure-phase ratings at the participant–image level, it separates baseline visual response from metadata-conditioned response change. At the implementation level, a minimal response-log schema should include the participant or session identifier, image identifier, presentation phase, label condition, label visibility state, rating value, and rating time. These fields allow system designers to determine the interface context under which each rating was produced.

These results also suggest that future AI-generated visual systems should account for disclosure state during response modeling. Possible implementations include disclosure-aware feature engineering, condition-specific rating calibration, or drift-weighted response modeling. Such approaches would help recommendation, ranking, and evaluation pipelines distinguish visual preference from interface-conditioned response variation.

C. Disclosure Design Considerations

The findings should not be interpreted as an argument against AI disclosure. Transparency remains important for provenance communication, user autonomy, and responsible deployment of AI-generated content [6]. However, the present study shows that disclosure is not behaviorally neutral. Once authorship metadata is placed inside an evaluation interface, it becomes part of the rating environment and may influence user-response data [6, 8]. Because the present implementation used a persistent text label displayed during the rating trial, the results should be interpreted most directly as evidence for this specific disclosure format rather than for all possible forms of AI-content disclosure.

Disclosure should therefore be treated as an interface design variable rather than as a fixed label added after system development. The timing, wording, placement, and visual salience of disclosure may all affect how users interpret and evaluate content [15]. A label displayed continuously during rating may produce a different response from post-evaluation disclosure, a hover-based provenance marker, a low-salience label, or a progressive disclosure interface. These design choices should be empirically tested rather than assumed to have no effect on user judgment.

For AI-generated visual information systems, the practical question is not simply whether disclosure should exist, but how disclosure should be implemented when user ratings are used as system input. Platforms may need to distinguish between disclosures used for compliance, disclosures used for provenance explanation, and disclosures shown during evaluative tasks. When ratings are used for recommendation, ranking, or quality assessment, disclosure design should be evaluated not only for informational accuracy but also for its effect on response drift.

VI. Limitations, Boundary Conditions and Future Work

The findings should be interpreted within the scope of the implemented framework. First, the study adopted a repeated-evaluation design in which the same participants rated the same images before and after metadata disclosure. This design was necessary for computing participant – image level response drift, but it may also introduce familiarity or carryover effects

between the blind and disclosure phases. The results therefore reflect response changes under a controlled two-phase interface setting rather than responses in a fully naturalistic browsing environment.

Second, attribution labels were not fully counterbalanced across stimuli, which limits claims about pure label effects independent of stimulus identity. To address this limitation, the primary model included stimulus category as a fixed effect and participant identity as a random intercept, while ordinal sensitivity analysis and leave-one-image-out analyses were used to assess the robustness of the detected pattern across model specifications and individual stimuli. These procedures provide robustness evidence but do not replace full label counterbalancing or support population-level generalization across images. The study is therefore interpreted as framework verification rather than as a causal estimate of pure authorship-label effects.

Third, the stimulus set was intentionally compact. The use of four AI-generated images made it possible to maintain a controlled repeated-measures structure and to keep the evaluation task manageable for participants. However, this design limits the range of visual styles, image topics, and aesthetic variation covered by the study. Future work should test the framework with larger and more diverse stimulus sets, including different genres of AI-generated visual content, multiple levels of visual quality, and broader interface contexts.

Fourth, label exposure was controlled through a persistent on-screen attribution overlay, but no separate post-task manipulation-check item was collected. The findings therefore rely on procedural exposure rather than self-reported verification of label awareness. Future studies should include manipulation checks or attention measures to examine whether participants noticed and interpreted the displayed attribution labels as intended.

Future research can extend the framework in three main directions. A fully counterbalanced design should assign each image to different metadata conditions across participants in order to separate label effects from stimulus identity more strictly. Alternative disclosure formats should also be compared, including pre-evaluation disclosure, post-evaluation disclosure, persistent labels, hover-based provenance markers, low-salience labels, and progressive disclosure interfaces. Finally, future AI-generated visual systems should record metadata-display conditions as part of the evaluation log, so that recommendation, ranking, and content assessment models can support disclosure-aware feature engineering, condition-specific calibration, or drift-weighted response modeling instead of treating all ratings as equivalent indicators of visual preference.

VII. Conclusion

This study proposed a two-phase evaluation framework for detecting metadata-conditioned response drift in AI-generated visual information systems. By integrating fixed visual input, assigned attribution-label presentation, trial-level logging, drift computation, and mixed-effects validation, the framework provides a structured procedure for examining whether attribution cues are associated with measurable changes in user-evaluation data. Based on 92 participants and 732 valid observations, the results showed a disclosure-associated drift pattern under the implemented label-assignment scheme: ratings increased for images assigned to the human-labeled condition and decreased for images assigned to the AI-labeled condition. This pattern was consistent across oil-style and digital-style stimuli and remained stable in ordinal sensitivity analysis and leave-one-image-out checks. Because the present implementation did not fully counterbalance labels across images and used a fixed blind-to-disclosure phase order, the results should be interpreted as framework-verification evidence rather than as a fully isolated causal estimate of pure authorship-label effects. Even within this boundary, the framework offers practical support for disclosure-aware response logging, interface evaluation, and the interpretation of user-rating data in AI-generated visual information systems.

Ethical Declarations

This study was a minimal-risk anonymous human–computer interaction evaluation involving adult participants who viewed and rated AI-generated images. Participation was voluntary, and informed consent was obtained before the task. Responses were recorded without personally identifiable information and analyzed only in aggregate form. According to institutional guidelines for minimal-risk anonymous evaluations, formal ethics review was not required. After completing the task, participants were informed that the displayed authorship labels were experimentally assigned attribution cues and that all stimuli were AI-generated.

References

[1] Z. Epstein et al., "Art and the science of generative AI," Science, vol. 380, no. 6650, pp. 1110-1111, 2023.

[2] E. Cetinic and J. She, "Understanding and creating art with AI: Review and outlook," ACM transactions on multimedia computing, communications, and applications (TOMM), vol. 18, no. 2, pp. 1-22, 2022.

[3] J. K. Eshraghian, "Human ownership of artificial creativity," Nature Machine Intelligence, vol. 2, no. 3, pp. 157-160, 2020.

[4] Z. Epstein, S. Levine, D. G. Rand, and I. Rahwan, "Who gets credit for AI-generated art?," Iscience, vol. 23, no. 9, 2020.

[5] S. Gheewala, S. Xu, and S. Yeom, "In-depth survey: deep learning in recommender systems—exploring prediction and ranking models, datasets, feature analysis, and emerging trends," Neural Computing and Applications, vol. 37, no. 17, pp. 10875-10947, 2025.

[6] C. Wittenberg, Z. Epstein, G. Péloquin-Skulski, A. J. Berinsky, and D. G. Rand, "Labeling AI-generated media online," PNAS nexus, vol. 4, no. 6, p. pgaf170, 2025.

[7] H. Gangadharbatla, "The role of AI attribution knowledge in the evaluation of artwork," Empirical Studies of the Arts, vol. 40, no. 2, pp. 125-142, 2022.

[8] L. Bellaiche et al., "Humans versus AI: whether and why we prefer human-created compared to AI-created artwork," Cognitive research: principles and implications, vol. 8, no. 1, p. 42, 2023.

[9] K. Millet, F. Buehler, G. Du, and M. D. Kokkoris, "Defending humankind: Anthropocentric bias in the appreciation of AI art," Computers in Human Behavior, vol. 143, p. 107707, 2023.

[10] R. Chamberlain, C. Mullin, B. Scheerlinck, and J. Wagemans, "Putting the art in artificial: Aesthetic responses to computer-generated art," Psychology of Aesthetics, Creativity, and the Arts, vol. 12, no. 2, p. 177, 2018.

[11] S. G. Chiarella, G. Torromino, D. M. Gagliardi, D. Rossi, F. Babiloni, and G. Cartocci, "Investigating the negative bias towards artificial intelligence: Effects of prior assignment of AI-authorship on the

aesthetic appreciation of abstract paintings," Computers in Human Behavior, vol. 137, p. 107406, 2022.

[12] C. B. Horton Jr, M. W. White, and S. S. Iyengar, "Bias against AI art can enhance perceptions of human creativity," Scientific reports, vol. 13, no. 1, p. 19001, 2023.

[13] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Algorithm aversion: people erroneously avoid algorithms after seeing them err," Journal of experimental psychology: General, vol. 144, no. 1, p. 114, 2015.

[14] J. M. Logg, J. A. Minson, and D. A. Moore, "Algorithm appreciation: People prefer algorithmic to human judgment," Organizational behavior and human decision processes, vol. 151, pp. 90-103, 2019.

[15] D. Gamage, D. Sewwandi, M. Zhang, and A. K. Bandara, "Labeling synthetic content: User perceptions of label designs for AI-generated content on social media," in Proceedings of the 2025 CHI conference on human factors in computing systems, 2025, pp. 1-29.

[16] F. Li, Y. Yang, and G. Yu, "Nudging perceived credibility: The impact of AIGC labeling on user distinction of AI-generated content," Emerging Media, vol. 3, no. 2, pp. 275-304, 2025.

[17] I. O. Gallegos et al., "Labeling messages as AI-generated does not reduce their persuasive effects," PNAS nexus, vol. 5, no. 2, p. pgag008, 2026.

[18] F. Seeger, M. Wessel, and C. Lehrer, "AI content labeling and user engagement on social media: The role of AI level, content type, and disclosure timing," Electronic Markets, vol. 36, no. 1, p. 31, 2026.

[19] S. Amershi et al., "Guidelines for human-AI interaction," in Proceedings of the 2019 chi conference on human factors in computing systems, 2019, pp. 1-13.

[20] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," Journal of statistical software, vol. 67, pp. 1-48, 2015.

[21] R. H. B. Christensen, "ordinal—regression models for ordinal data," R package version, vol. 10, no. 2019, p. 54, 2019.

[22] G. Norman, "Likert scales, levels of measurement and the “laws” of statistics," Advances in health sciences education, vol. 15, no. 5, pp. 625-632, 2010.