Operationalising Journalistic and Indexical Dimensions of AI Imagery

Bruno N. Sotic[1], Jaap Kamps[1]

[1]University of Amsterdam, Amsterdam, the Netherlands

Abstract

Current evaluations of news images typically reduce article - image fit to a single score, implicitly treating fit as a unified construct. In this Quest for Insight, we examine whether different dimensions of evaluation become visible when editorial and AI-generated news images are assessed separately in terms of content, function, authenticity, and trust.

Drawing on Westman's theory of journalistic image attributes and Benjamin's notion of indexical authenticity, we develop a five-dimensional framework comprising visual-content match, thematic appropriateness, functional suitability, indexical authenticity, and editorial trustworthiness. We pilot the framework on ten NewsImages 2026 articles paired with four image variants each: the editorial original, an AI restoration, an AI photorealistic generation, and an AI illustration.

The pilot reveals three recurring patterns. First, original photographs and AI restorations are largely indistinguishable on visual, thematic, and functional dimensions, yet diverge substantially on authenticity and trustworthiness. Second, AI illustrations and photorealistic generations receive similar trust judgments despite occupying opposite positions on authenticity, suggesting that editorial evaluation depends not only on realism but also on whether an image implicitly claims to document a real event. Third, differences in authenticity largely disappear for abstract topics where no photograph could plausibly function as evidence, indicating that the importance of authenticity is itself context dependent.

As a small pilot, the study is not intended to establish how readers generally evaluate AI-generated news imagery. Rather, it demonstrates that dimensions commonly collapsed into a single notion of fit can diverge in theoretically meaningful ways, particularly around questions of authenticity and trust. The observed patterns suggest that evaluating AI-generated news images through a single fit measure may obscure distinctions that become visible when content, function, authenticity, and editorial appropriateness are considered separately.

1. Introduction

Prior iterations of the NewsImages task at MediaEval report that AI-generated images are frequently rated as a better fit than the editorial original [1] and that prompt-aligned generation can outperform retrieval over large image collections [2]. Our two companion baselines extend this to photorealistic restoration of degraded historical scans [3] and to non-photorealistic illustration [4]. Both report on a single-Likert measure of perceived article - image fit.

In this *Quest for Insight* we step back from that metric and ask what "fit" is doing for news imagery. A single rating collapses several judgments journalism actively separates: whether the image shows the article's named subject, whether it captures the article's broader theme, whether it functions appropriately in an editorial role, and whether it presents itself as the right *kind* of image given the article's content. Westman's empirical work on journalistic image access [5] organises these judgments at three levels of image attributes: *non-visual* (biographical, contextual, and production metadata), *syntactic* (visual characteristics such as composition,

MediaEval'26: Multimedia Evaluation Workshop, June 15–16, 2026, Amsterdam, Netherlands and Online
*Corresponding author.
nadalic.sotic@uva.nl (B. N. Sotic); kamps@uva.nl (J. Kamps)
0009-0003-2122-2235 (B. N. Sotic); 0000-0002-6614-0087 (J. Kamps)

colour, and texture), and *semantic* (what the image depicts and what it is about, including a thematic sub-layer). Two of her empirical findings are particularly relevant here. Domain experts construct *functionally oriented* categories (categories defined by what an image does in a publication rather than by what it shows) substantially more than non-experts, indicating that Function is a real expert-recognised access point alongside content. And the availability of page context promotes *thematic* rather than object-based categorization, meaning that what counts as a good image for an article is not reducible to a visual match between image content and article entities. Standard AI image evaluation (e.g., CLIPScore, FID, learned no-reference perceptual quality metrics, and the single-Likert fit scale) operates almost entirely at the syntactic and semantic-content layers; the Function and thematic dimensions Westman identifies are invisible to it.

Generative AI introduces a further dimension that pre-AI evaluation did not need to register. Benjamin's account of mechanical reproduction [6] is best known for the claim that the *aura* of the unique artwork erodes under reproduction, and at first glance this argument does not transfer to journalism, where images have always been mass-reproduced and have never had aura in that classical sense. The argument that does transfer is the inverse case that photojournalism developed in the twentieth century: the press photograph's authority rests not on uniqueness but on its *indexical* bond to a real event [7]. A news photograph functions as a witness; its meaning includes the implicit claim "this happened, this person was there," and editorial conventions, captioning practices, and reader expectations all sit on top of that indexical guarantee. AI-generated photorealistic imagery ruptures this bond without disturbing the visual surface that existing metrics measure: it depicts events whose visual specifics no camera recorded, while satisfying every test of photographic plausibility. Recent work has begun to articulate the consequences for photographic authenticity in journalism. Park's notion of *semi-aura* [8] captures precisely the in-between status of AI-generated work that is neither uniquely authored nor straightforwardly mechanical, and recent qualitative research with practitioners reports a generalised distrust that bleeds back into the reception of authentic photography [9]. The implication for evaluation is that indexical authenticity, and the editorial trustworthiness that follows from it, are dimensions on which AI-generated and editorial images occupy qualitatively different positions invisible to existing instruments.

We propose a five-dimension instrument that operationalises Westman's typology (visual-content match, thematic appropriateness, functional suitability) together with indexical authenticity and editorial trustworthiness, and we pilot it on a small, deliberately stratified sample of ten NewsImages 2026 articles paired with four image variants each (original, AI restoration, AI photorealistic generation, AI illustration).

2. Method

Stimuli. We selected ten articles from the NewsImages 2026 test set through purposive stratified sampling across four content types: historical-scan articles with high indexical stakes ($n = 3$), contemporary articles with politically charged scenes ($n = 3$), abstract or scientific articles whose subject has no canonical photographic referent ($n = 2$), and light-news / human-interest articles where we expect image-type differences to flatten ($n = 2$). The full list of articles is provided in Appendix. For each article, we assembled four image variants: the editorial original from the test set, an AI photorealistic restoration was produced with Qwen-Image-Edit-2509 from the restoration companion paper, an AI photorealistic generation produced with Z-Image-Turbo from the article title under a photorealistic prompt, and an AI illustration produced with Flux.2 Klein from the illustration companion paper. This yields 40 stimuli.

Table 1
The five-dimensional evaluation instrument. Each dimension is rated on a 7-point Likert scale from 1 (not at all) to 7 (completely).

Dimension	Wording shown to coder	Theoretical home
Visual-content match	The image visually depicts what the article describes (people, objects, scene named in the article).	Westman, semantic content
Thematic appropriateness	Independent of whether the image shows the article's specific subject, the image captures the article's broader theme or mood.	Westman, thematic semantic
Functional suitability	This image works appropriately as the visual accompaniment to this article in a news publication.	Westman, Function
Indexical authenticity	This image presents itself as a photograph of a real event that actually occurred.	Benjamin, indexical bond
Editorial trustworthiness	Using this image with this article would be editorially appropriate in a reputable news publication.	Editorial concern

Instrument. Each stimulus is rated on five 7-point Likert dimensions, anchored at 1 ("not at all") and 7 ("completely"). Three dimensions operationalize Westman's typology of journalistic image attributes, one operationalizes the indexical reading of Benjamin, and one captures editorial trustworthiness as the downstream practical consequence. The full instrument is shown in Table 1.

ICR and protocol. Two Media and Communication graduates and the first author independently coded the full set of 40 stimuli. Prior to the main session the coders calibrated on five articles separate from the final ten, yielding Cohen's κ per dimension in the 0.64 to 0.78 range (lowest on thematic appropriateness, highest on indexical authenticity).

3. Results and Analysis

Given the pilot nature and small sample size, the following analysis remains exploratory; however, formal statistical tests (Friedman tests and Kendall's W coefficients of concordance) are provided in the Appendix. Figure 1 shows how the four image variants score on each of the five dimensions, averaged across the ten articles. Three patterns in the data carry the argument we made in the introduction.

The original and the restoration look the same on the Westman dimensions, but not on the Benjamin dimension On visual content, thematic appropriateness, and functional suitability, the original photograph and the AI restoration are essentially indistinguishable. They diverge on indexical authenticity, where the original sits at the top of the scale and the restoration falls more than a point below it, and on editorial trustworthiness, where the restoration falls further still. A reader rating "fit" on a single scale would treat these two images as broadly equivalent. The multi-dimensional view shows that restoration improves the visual surface of the original without rebuilding what makes the original *evidentiary* in a news context.

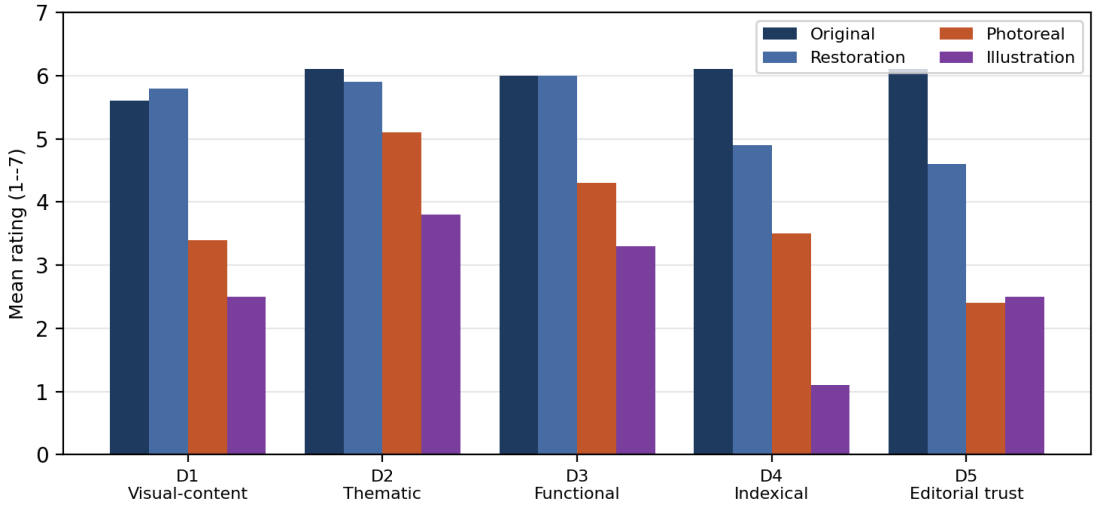

Figure 1: Mean ratings (1–7) by image variant and dimension across the ten pilot articles.

The restored image presents itself as a clearer version of the same scene, but the rater does not extend the same confidence in its authenticity, and the editorial judgment follows.

Illustration is trusted as much as photorealistic AI generation, and for the opposite reason The most striking pattern in Figure 1 is the bottom of the editorial trustworthiness bar. AI photorealistic generation and AI illustration land at the same point on this dimension, even though they sit at opposite ends of indexical authenticity — the illustration at the floor of the scale, the photorealistic generation in the middle. In nine of the ten articles, illustration is trusted as much as or more than photorealistic generation. The framework reads this cleanly. A photorealistic depiction of an event that did not occur in the form depicted is editorially compromising because it makes a photographic claim without the indexical bond to back it. An illustration of the same event makes no such claim and asks the reader to interpret it as an illustration; raters appear to penalise the failed photographic claim more than they penalise the absence of one. Editorial trust does not require indexicality. It requires that the image does not pretend to have it.

The dimensions decompose into two layers Across the forty stimuli, ratings on visual content, thematic appropriateness, and functional suitability move together tightly: a stimulus that scores high on one tends to score high on the others, and the three dimensions are barely distinguishable when read jointly. Indexical authenticity sits noticeably apart from this cluster. Editorial trustworthiness sits between them, pulled toward the syntactic cluster on stimuli that look like coherent photographs and pulled toward indexicality on stimuli that raise an authenticity question. The framework's split between Westman's typology of journalistic image access and Benjamin's indexical bond is not just a theoretical convenience here. The two layers behave differently in the data.

Where the framework predicts no difference The patterns above describe what happens on articles where a press photograph would normally function as a witness to the event being reported - the historical, contemporary, and human-interest articles in our sample. Table 5 shows what happens on the two abstract articles, where the article subject (a Mercury transit, a

Table 2
Mean ratings on indexical authenticity by image variant and content stratum.

Stratum	Original	Restoration	Photoreal	Illustration	Range
Historical ($n = 3$)	7.0	5.0	3.0	1.0	6.0
Contemporary ($n = 3$)	7.0	6.0	4.0	1.0	6.0
Human interest ($n = 2$)	6.5	5.0	4.0	1.0	5.5
Abstract ($n = 2$)	3.0	3.0	3.0	1.5	1.5

paleontological reconstruction) has no canonical photographic referent at all. The four image variants line up on indexical authenticity across a six-point range on the other strata; on the abstract stratum the same range collapses to two points, and the original photograph itself receives a low score. The indexical contract is not in force here. There is no event for the photograph to bear witness to, no claim for the generative image to fail to keep. Image type stops mattering on this dimension, which is exactly what the framework predicts.

4. Discussion and Conclusion

We set out from a concern that the standard evaluation of news images, which reduces the question of fit to a single rating, fails to register what journalism actually decides when it picks an image for an article. The framework we proposed says that fit decomposes into three Westman dimensions describing what an image does in a publication, plus an indexical dimension describing what kind of relationship the image has to the world, plus the editorial judgment that follows. The pilot above offers an early look at whether that decomposition does any empirical work, on ten articles, four image variants each, with three coders.

The answer the data gives is: the decomposition does work, and the Benjamin layer is where most of the work happens. AI restoration and the original are visually equivalent - the Westman dimensions cannot tell them apart, but rate differently as evidence. AI illustration and AI photorealistic generation are trusted equally despite looking nothing alike, because the failed photographic claim of the generation matters more to raters than the explicit non-claim of the illustration. And the indexical separation across image variants collapses on articles where no photograph could have been taken in the first place, which is the boundary condition the framework predicts. None of these patterns would have been visible if the rating instrument had asked for one number.

We are aware of what this pilot is and what it is not. Ten articles and three coders cannot establish how readers in general weight these dimensions, and we have not made that claim. What we have shown is that the dimensions are codeable, that they capture distinct things, and that the framework's prediction about where indexicality should matter and where it should not is borne out in this sample. The natural next step is to apply the same instrument retrospectively to existing NewsImages submission data, where pairwise reader preferences across many image types and articles could be reread through this lens. The longer step is a larger user study with non-expert readers, where the open question is not whether the dimensions separate — this pilot suggests they do — but how readers weight them against each other when AI-generated imagery is the image they encounter in their feeds.

Acknowledgments

We thank Bram Bakker for providing the dataset of historical news articles, and Qi Bi and Lucien

Heitz for providing the generation pipelines. Bruno N. Sotic, and Jaap Kamps are partly funded by the Netherlands Organization for Scientific Research (NWO NWA #1518.22.105). Jaap Kamps is further supported by the University of Amsterdam (AI4FinTech program) and ICAI (AI for Open Government Lab).

Declaration on Generative AI

The authors used GPT-5.5 and Grammarly for spelling checks. The authors have reviewed and edited the content as needed. They take full responsibility for the publication's content.

A. Appendix

Table 3
The ten articles selected for the pilot study, by stratum.

ID	Stratum	Title
8511	Historical	Landslide French elections: De Gaulle's great triumph
8519	Historical	Departure of Apollo 16 went smoothly
8598	Historical	Hutu refugees flee Rwandan capital; Tutsi rebels surround Kigali
13739	Contemporary	Merkel marks anniversary of Berlin Wall
17280	Contemporary	Hundreds protest outside Supreme Court during DACA case
19378	Contemporary	Jacksonville monument to U.S. Marines killed in 1983 Beirut bombing
9181	Abstract/scientific	Mercury transit is coming to a sky near you Monday
13570	Abstract/scientific	Fluffy Dinosaurs Existed and Lived at the South Pole
9621	Human interest	Cincinnati Zoo: rescue dog Remus and baby cheetah Kris "BFF sleepover"
22554	Human interest	Student volunteers help the elderly as winter approaches

Table 4
Friedman test per dimension. Within-subjects unit: ten articles; repeated measures: four image variants. W is Kendall's coefficient of concordance, bounded $[0, 1]$, with larger values indicating stronger image-type effects.

Dimension	χ^2	p	Kendall's W
D1 Visual-content match	25.76	< 0.001	0.86
D2 Thematic appropriateness	24.39	< 0.001	0.81
D3 Functional suitability	26.44	< 0.001	0.88
D4 Indexical authenticity	26.20	< 0.001	0.87
D5 Editorial trustworthiness	28.24	< 0.001	0.94

Table 6 reports median ratings by image type and dimension across the ten articles. Three features of the pattern are worth naming.

Table 5
Median indexical-authenticity (D4) rating by image type and content stratum.

Stratum	Original	Restoration	Photoreal	Illustration	Range
Historical ($n = 3$)	7	5	3	1	6
Contemporary ($n = 3$)	7	6	4	1	6
Human interest ($n = 2$)	7	5	4	1	6
Abstract ($n = 2$)	3	3	3	1	2

Table 6
Median ratings (1–7) by image type and dimension across the ten articles. Interquartile ranges in brackets.

Image type	D1 Visual	D2 Theme	D3 Function	D4 Indexical	D5 Editorial
Original	6 [5–6]	6 [6–7]	6 [6–6]	7 [6–7]	6 [6–7]
AI restoration	6 [5–6]	6 [6–6]	6 [6–6]	5 [5–6]	5 [4–5]
AI photoreal	3 [3–4]	5 [5–6]	4 [4–5]	3 [3–4]	2 [2–3]
AI illustration	2 [2–4]	3 [3–5]	3 [2–4]	1 [1–1]	2 [2–3]

References

[1] L. Heitz, A. Bernstein, L. Rossetto, An empirical exploration of perceived similarity between news article texts and images, in: Working Notes Proceedings of the MediaEval 2023 Workshop, 2024.

[2] L. Heitz, Y. K. Chan, H. Li, K. Zeng, A. Bernstein, L. Rossetto, Prompt-based alignment of headlines and images using openclip, in: Working Notes Proceedings of the MediaEval 2023 Workshop, 2024.

[3] L. Heitz, B. N. Sotic, Q. Bi, J. Kamps, Assessing the viability of image-to-image models for restoring historical photographs of news articles, in: Proceedings of the MediaEval'26: Multimedia Evaluation Workshop, Amsterdam, Netherlands, 2026. To appear.

[4] B. N. Sotic, J. Kamps, Illustration baseline for the newsimages 2026 challenge, in: Proceedings of the MediaEval'26: Multimedia Evaluation Workshop, Amsterdam, Netherlands, 2026. To appear.

[5] S. Westman, Journalistic image access: Description, categorization and searching, Ph.D. thesis, Aalto University, Finland, 2011.

[6] W. Benjamin, The work of art in the age of mechanical reproduction, in: H. Arendt (Ed.), Illuminations, Schocken Books, New York, 1969, pp. 217–251.

[7] J. Huxford, Beyond the referential: Uses of visual symbolism in the press, Journalism 2 (2001) 45–71. doi:10.1177/146488490100200102.

[8] S. Park, The work of art in the age of generative ai: aura, liberation, and democratization, AI & SOCIETY 40 (2024) 1807 – 1816. URL: https://api.semanticscholar.org/CorpusID:269570297.

[9] D. Harff, Detection and spill-over effects of ai-generated images in political messages: Evidence from two pre-registered experiments, Computers in Human Behavior (2026). URL: https://www.sciencedirect.com/science/article/pii/S0747563226000865. doi:10.1016/j.chb.2026.108989.