Operational Evidence and Outcomes
Volume 1 · Number 1 · 2026

An Average Winner Can Lose on Many Images: A Held-Content Audit of PIPAL Ratings

Trevor Mason[1] · Gabriel Sullivan[2] · Ryan King[3]

[1] Spain · [2] Portugal · [3] Italy

Original publication: November 8, 2026 | Article 579

Abstract

Selecting a restoration method by its mean rating leaves a second question: how often does it win on individual content? We selected the two highest-mean variant codes using one hundred PIPAL reference images, then compared them on another hundred. The selected winner retained a mean Elo advantage of 16.97 points, but had the higher released rating on only 62 references. Its mean-difference interval was wide, from 1.08 to 32.87. A reproducible content partition can therefore preserve an average ordering while revealing many local reversals. The exercise uses existing Elo scores and does not constitute a new observer comparison or identify a universal restoration algorithm.

Public-rating research draft. Variant identities remain coded; authorship, novelty, and final interpretation are pending human assessment.

An Average Winner Can Lose on Many Images: A Held-Content Audit of PIPAL Ratings

Public-rating research draft. Variant identities remain coded; authorship, novelty, and final interpretation are pending human assessment.

Abstract

Selecting a restoration method by its mean rating leaves a second question: how often does it win on individual content? We selected the two highest-mean variant codes using one hundred PIPAL reference images, then compared them on another hundred. The selected winner retained a mean Elo advantage of 16.97 points, but had the higher released rating on only 62 references. Its mean-difference interval was wide, from 1.08 to 32.87. A reproducible content partition can therefore preserve an average ordering while revealing many local reversals. The exercise uses existing Elo scores and does not constitute a new observer comparison or identify a universal restoration algorithm.

1. From a leaderboard choice to a content decision

An image-restoration leaderboard often ranks methods by an average subjective score. A practitioner, however, applies a method to particular content. The method with the best mean can be less desirable on a substantial fraction of images. This distinction matters when the final application has unusual textures, thin structures, or other content poorly represented by an average. We investigate the difference between retaining an aggregate advantage and retaining a per-reference advantage after a content partition.

PIPAL was introduced to assess perceptual image restoration and includes subjective ratings assigned through an Elo-based process (Gu et al., 2020). Its structure allows multiple restoration variants to be compared on common references. Our question does not ask which new architecture should be proposed. It asks whether a choice based on one reference partition remains convincing on another partition when both the mean difference and the frequency of local reversals are reported.

The distinction between semantic information and distortion sensitivity in Meng, 2026 motivates attention to what an aggregate quality result can conceal. Here we do not use prompt labels or evaluate semantic correspondence. We study released restoration ratings, making this a boundary case for the broader argument that a scalar summary can hide different dimensions of performance. The target model is neither trained nor treated as one of the coded restoration variants.

2. Related evaluation choices

Structural similarity (Wang et al., 2004), learned perceptual distance (Zhang et al., 2018), and joint structure-texture comparison (Ding et al., 2020) represent different ways of quantifying image differences. Their relevance is conceptual: a model ranking depends on the quality notion being measured. Our outcome remains the supplied Elo rating rather than any of these objective metrics. Mixing their numerical scales would introduce an additional measurement question that the present paired design does not resolve.

Restoration methods also differ in how they construct missing or degraded content. Deep Image Prior uses network structure as a prior (Ulyanov et al., 2017), while large-mask inpainting (Suvorov et al., 2021) and diffusion-based restoration (Lugmayr et al., 2022) illustrate different routes to plausible output. Partial convolutions (Liu et al., 2018), gated convolutions (Yu et al., 2018b), and contextual attention (Yu et al.,

2018a offer further examples. These works motivate content-sensitive evaluation; they are not labels assigned to our two selected codes without a verified code-to-method mapping.

Edge guidance in Nazeri et al., 2019 and the state-space restoration approach in Guo et al., 2024 reinforce that a visually useful method can be optimized around particular structures. The current metadata cannot identify which local visual property caused a reversal. We therefore avoid attributing score differences to hallucinated texture, edge preservation, or any other mechanism. Those interpretations would require image inspection, explicit features, and a design capable of separating correlated content characteristics.

NIMA models image assessment using rating information (Talebi and Milanfar, 2017), whereas Kirstain et al., 2023 and Wu et al., 2023 involve human preferences for generated images. These examples distinguish a released numerical summary from the individual judgments that produced it. In particular, an Elo difference is not the same as an observed fraction of people preferring one image. Our per-reference win indicator records which released score is larger, and its denominator is reference images rather than human comparisons.

3. A selection rule that excludes test content

We first examine the acquired maintainer metadata rather than assuming its row count matches a single complete grid. It has 24,200 rows. Of these, 23,200 belong to the official training partition, comprising 200 references and 116 variants each. Another 1,000 rows belong to an official validation partition with 25 other references and forty variants. Combining those partitions would produce an irregular design. We restrict analysis to the complete official training grid before constructing our own internal split.

Within that grid, reference identifiers are sorted. The first hundred form the selection partition, and the remaining hundred form the evaluation partition. We calculate each variant's mean raw Elo only on selection references, sort those means with a stable code-based tie rule, and retain the top two codes. This is an internal held-content experiment. It does not use PIPAL's official validation performance or claim to reproduce its published evaluation protocol.

The selected codes are 02_04 and 02_03. We preserve them as identifiers because their algorithm names were not established from an authenticated mapping in this workflow. Guessing an identity from neighboring code values would add unsupported detail. The selection means are 1643.14 and 1613.76, respectively. Their difference describes the information available to the selection rule; the evaluation scores remain excluded until after that choice has been frozen.

Model-selection warnings in Varma and Simon, 2006 apply even when selection is a simple maximum rather than a complex hyperparameter search. Choosing the most favorable test pair would exaggerate the stability of a ranking. Structured-validation guidance in Roberts et al., 2017 similarly supports separating reference content rather than scattering individual distorted versions across both partitions. The partition is deterministic by identifier, however, so it does not establish performance on a random or representative population of future photographs.

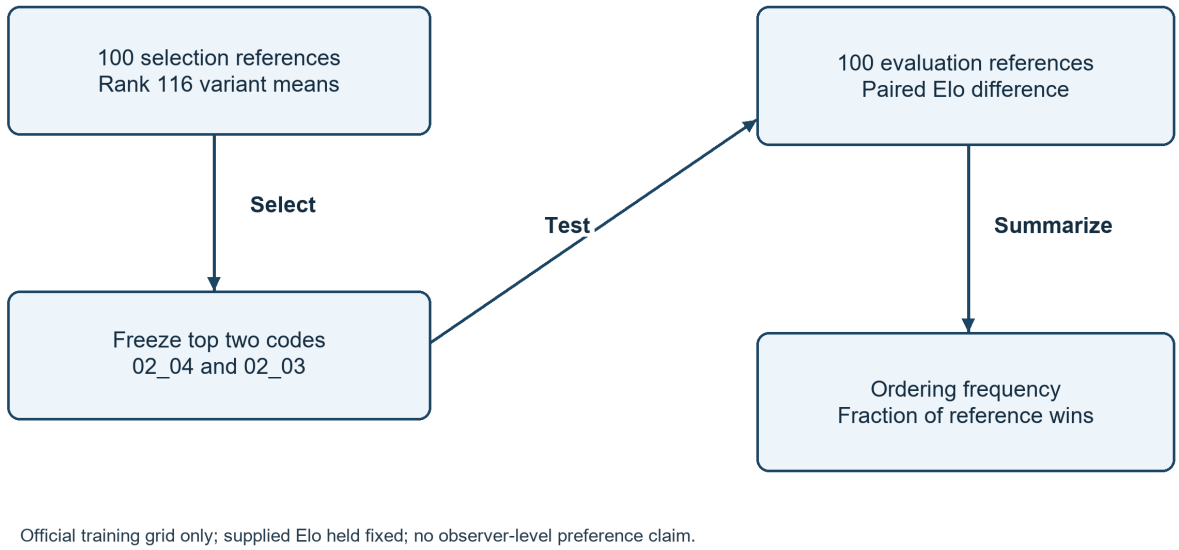

Figure 1. Training-reference means select two variant codes; their frozen identities then feed the mean Elo difference and reference-level ordering frequency on held-out content.

4. Two outcomes from the same paired scores

For each evaluation reference, we subtract the runner-up code's raw Elo from the selected winner's raw Elo. The first estimand is the equal-reference average of these hundred differences. Its intercept-only ordinary least squares model has the usual paired-mean t interval with ninety-nine degrees of freedom. Pairing retains common-reference information and avoids treating two variant ratings of one reference as unrelated observations.

The second outcome is an indicator that the selected winner has the larger Elo for that reference. Its mean is a descriptive fraction. We report an HC3 standard error and t-based interval from an intercept-only linear probability fit. This interval is not a binomial model of observer preference, and no Elo-to-win-probability transformation is assumed. It concerns hypothetical variation across reference content conditional on the already released ratings and the fixed selected pair.

Both models condition on the training selection and the supplied Elo values. They do not propagate uncertainty from the original global rating procedure, possible common observers, or estimation of an entirely new selection set. Gneiting and Raftery, 2007 emphasizes that an evaluation should identify its predictive target. Here the targets are a mean released-score difference and an image-content ordering frequency. Neither is a complete probability forecast of human preference.

Agreement research in Bland and Altman, 1986, Lin, 1989, and Koo and Li, 2016 helps clarify why these outcomes should be kept apart. High average performance, agreement between rating systems, and consistency across individual references answer different questions. We do not estimate an intraclass correlation or concordance coefficient, because neither is needed to define our paired contrast. The references instead provide a measurement vocabulary for resisting an overly broad interpretation of a single leaderboard number.

Table 1. Held-out raw Elo difference and the fraction of references favoring the selected winner. The second row's interval concerns reference-level ordering, not observer votes.

Model	Term	Estimate (SE)	95% CI	N
M1	Mean contrast	16.9719 (8.0112)	[1.0760, 32.8679]	100
M1: fit	R-squared = 0.0000	-	-	100
M2	Fraction of reference wins	0.6200 (0.0490)	[0.5227, 0.7173]	100
M2: fit	R-squared = 0.0000	-	-	100

Model labels: M1 = Held-out selected winner minus runner Elo; M2 = Held-out reference fraction favoring selected winner.

5. A retained mean lead with many reversals

On the evaluation references, code 02_04 has mean Elo 1633.39 and code 02_03 has mean Elo 1616.42. Their equal-reference difference is 16.97 points, compared with the larger separation observed in selection. The paired standard error is 8.01, giving a 95 percent interval of 1.08 to 32.87. The selected mean ordering persists in this partition, but the plausible magnitude under the stated reference-sampling model remains broad.

Code 02_04 has the higher rating on 62 references and the lower rating on 38. There are no ties. The reference-level fraction is 0.620, with HC3 interval 0.523 to 0.717. Thus an average winner is not a reference-by-reference winner. This finding does not establish that a content-adaptive selector would improve user satisfaction: designing such a selector would require features, a separate training rule, and an untouched evaluation set.

The two outcome rows are not independent replications. They summarize the same hundred paired observations in different ways. A large positive difference on a few references can contribute strongly to the mean while counting only once in the ordering fraction. Conversely, numerous small positive differences can produce a high ordering frequency with a modest mean advantage. Reporting both therefore exposes a meaningful property of the comparison without claiming additional data collection.

The mean-only models have zero R-squared by construction. This is unrelated to whether the selected code retains an advantage. White, 1980 motivates attention to heteroskedasticity in general regression, while our ordinary paired-mean interval and separate HC3 fraction interval have explicitly different covariance choices. Neither model introduces reference-image covariates, and neither predicts which of the 38 reversals will occur on a new image.

6. What robustness would mean here

One could ask whether the mean advantage is dominated by a few unusual references. Hampel, 1974 and Rousseeuw, 1984 provide robust-estimation perspectives, and Cook, 1977 addresses influential observations in regression. We cite these as possible extensions rather than implying that an unreported robust fit was conducted. The present result retains all hundred evaluation references. Removing observations because they weaken the selected winner would invalidate the stated equal-reference question.

The interval around the Elo difference only narrowly excludes zero. The interpretive position of Wasserstein and Lazar, 2016 is therefore particularly relevant: scientific meaning does not change

abruptly at an arbitrary p-value cutoff. A small positive lower bound should not be advertised as decisive superiority. We also do not compare all 116 variants on the evaluation partition and report only favorable pairwise tests. Broader simultaneous comparisons would require an explicit multiplicity strategy, such as the false-discovery framework in Benjamini and Hochberg, 1995, and a different inferential claim.

There is an additional dependence problem that reference-level intervals cannot solve. Elo values were estimated within a shared rating process, so their errors may be coupled across variants and references. Holding those scores fixed yields a useful finite-release content audit, not a new psychophysical experiment. To estimate human preference probabilities directly, one would need comparison-level responses, the rating-assignment design, and a model that represents that dependence rather than treating final scores as unmeasured truth.

7. Implications for reporting restoration results

A useful restoration report can state the mean difference, the number of references on each side of zero, and the selection procedure together. That format lets a reader distinguish an average benefit from a reliable per-image guarantee. It also encourages an application team to ask whether its own content resembles the benchmark references. Our results do not specify a minimum fraction needed for deployment; that threshold depends on the consequences of occasional poor output and on available fallback review.

The broader practical lesson is to preserve the separation between choosing a method and describing its uncertainty. Selecting a code on training means is easy to reproduce. Establishing why it wins, whether a user notices the difference, and whether a failure matters requires additional evidence. This manuscript supplies the first kind of evidence and an honest content-level description of the second partition. It does not turn a coded metadata ranking into an undocumented visual demonstration.

Data documentation follows Gebru et al., 2021 by recording the mixed official partitions, the internal split, and the unresolved algorithm mapping. Reuse records follow Wilkinson et al., 2016 through source links, stable local hashes, and separate processed tables. The package includes the selected codes and all hundred paired evaluation scores, allowing the mean and frequency to be recomputed directly. It retains source licensing conditions and identifies the metadata maintainer separately from the original dataset creators.

8. Scope of the conclusion

One deployment response to local reversals might be to let a reviewer choose between two candidate restorations. That decision introduces a new resource budget and a new outcome: reviewer effort and the quality of the chosen image. Our metadata cannot estimate either. Another response would be to train a gate that predicts which code will perform better. Such a gate would need training labels derived without contaminating the evaluation references, and its full selection pipeline would need an additional untouched partition. The observed 38 reversals are motivation for those designs, not evidence that either strategy already works.

The selected pair's modest evaluation advantage also illustrates why reporting only the selection means is insufficient. Selection used a maximum over 116 candidates, whereas evaluation assessed a frozen pair. The smaller held-content gap can arise from content variation and selection optimism without implying any change to the underlying algorithms. We do not estimate the contribution of each explanation. Keeping the two partitions and their respective means visible is enough to avoid confusing a favorable training ranking with demonstrated stability across all future content.

The executed experiment finds a mean advantage that survives one deterministic content split, together with reversals on 38 percent of its evaluation references. Those facts can coexist without contradiction. They caution against presenting the top average rating as a promise for every image. Further work should examine independent content, verified method identities, and observer-level comparisons before drawing claims about an algorithm's universal perceptual quality. For this release, the coded held-content comparison is the full evidential scope.

References

1. Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da Silva Santos, Luiz Bonino; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim; Crosas, Mercè; Dillo, Ingrid; Dumon, Olivier; Edmunds, Scott; Evelo, Chris T.; Finkers, Richard; Gonzalez-Beltran, Alejandra; Gray, Alasdair J.G.; Groth, Paul; Goble, Carole; Grethe, Jeffrey S.; Heringa, Jaap; 't Hoen, Peter A.C; Hooft, Rob; Kuhn, Tobias; Kok, Ruben; Kok, Joost; Lusher, Scott J.; Martone, Maryann E.; Mons, Albert; Packer, Abel L.; Persson, Bengt; Rocca-Serra, Philippe; Roos, Marco; van Schaik, Rene; Sansone, Susanna-Assunta; Schultes, Erik; Sengstag, Thierry; Slater, Ted; Strawn, George; Swertz, Morris A.; Thompson, Mark; van der Lei, Johan; van Mulligen, Erik; Velterop, Jan; Waagmeester, Andra; Wittenburg, Peter; Wolstencroft, Katherine; Zhao, Jun; Mons, Barend. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. https://doi.org/10.1038/sdata.2016.18.
2. Talebi, Hossein and Peyman Milanfar. "NIMA: Neural Image Assessment" *arXiv preprint arXiv:1709.05424*, 2017. https://arxiv.org/abs/1709.05424.
3. Wasserstein, Ronald L.; Lazar, Nicole A.. (2016). The ASA Statement on p -Values: Context, Process, and Purpose. The American Statistician. https://doi.org/10.1080/00031305.2016.1154108.
4. Gu, Jinjin; Cai, Haoming; Chen, Haoyu; Ye, Xiaoxing; Ren, Jimmy; Dong, Chao. (2020). PIPAL: a Large-Scale Image Quality Assessment Dataset for Perceptual Image Restoration. https://arxiv.org/abs/2007.12142.
5. Gneiting, Tilmann; Raftery, Adrian E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association. https://doi.org/10.1198/016214506000001437.
6. Hampel, Frank R.. (1974). The Influence Curve and its Role in Robust Estimation. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.1974.10482962.
7. Ding, Keyan, et al. "Image Quality Assessment: Unifying Structure and Texture Similarity." *arXiv preprint arXiv:2004.07728*, 2020. https://arxiv.org/abs/2004.07728.
8. White, Halbert. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica*, vol. 48, no. 4, 1980, pp. 817-838. https://doi.org/10.2307/1912934.
9. Zhang, Richard, et al. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric" *arXiv preprint arXiv:1801.03924*, 2018. https://arxiv.org/abs/1801.03924.
10. Meng, Zijie. (2026). Decoupling Semantics from Distortions: Multi-Scale Two-Stream Vision-Language Alignment for AI-Generated Image Quality Assessment. arXiv:2606.16799. https://arxiv.org/abs/2606.16799.
11. Yu, Jiahui; Lin, Zhe; Yang, Jimei; Shen, Xiaohui; Lu, Xin; Huang, Thomas S.. (2018a). Generative Image Inpainting with Contextual Attention. arXiv:1801.07892. https://arxiv.org/abs/1801.07892.
12. Kirstain, Yuval, et al. "Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation" *arXiv preprint arXiv:2305.01569*, 2023. https://arxiv.org/abs/2305.01569.
13. Liu, Guilin; Reda, Fitsum A.; Shih, Kevin J.; Wang, Ting-Chun; Tao, Andrew; Catanzaro, Bryan. (2018). Image Inpainting for Irregular Holes Using Partial Convolutions. arXiv:1804.07723. https://arxiv.org/abs/1804.07723.
14. Yu, Jiahui; Lin, Zhe; Yang, Jimei; Shen, Xiaohui; Lu, Xin; Huang, Thomas. (2018b). Free-Form Image Inpainting with Gated Convolution. arXiv:1806.03589. https://arxiv.org/abs/1806.03589.
15. Benjamini, Yoav and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 57, no. 1, 1995, pp. 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
16. Ulyanov, Dmitry; Vedaldi, Andrea; Lempitsky, Victor. (2017). Deep Image Prior. arXiv:1711.10925. https://arxiv.org/abs/1711.10925.
17. Lugmayr, Andreas; Danelljan, Martin; Romero, Andres; Yu, Fisher; Timofte, Radu; Van Gool, Luc. (2022). RePaint: Inpainting using Denoising Diffusion Probabilistic Models. arXiv:2201.09865. https://arxiv.org/abs/2201.09865.

18. Bland, J. Martin, and Douglas G. Altman. "Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement." *The Lancet*, vol. 327, no. 8476, 1986, pp. 307-310. https://doi.org/10.1016/S0140-6736(86)90837-8.

19. Nazeri, Kamyar; Ng, Eric; Joseph, Tony; Qureshi, Faisal Z.; Ebrahimi, Mehran. (2019). EdgeConnect: Generative Image Inpainting with Adversarial Edge Learning. arXiv:1901.00212. https://arxiv.org/abs/1901.00212.

20. Wang, Zhou, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. "Image Quality Assessment: From Error Visibility to Structural Similarity." *IEEE Transactions on Image Processing*, vol. 13, no. 4, 2004, pp. 600-612. https://doi.org/10.1109/TIP.2003.819861.

21. Rousseeuw, Peter J.. (1984). Least Median of Squares Regression. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.1984.10477105.

22. Koo, Terry K., and Mae Y. Li. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." *Journal of Chiropractic Medicine*, vol. 15, no. 2, 2016, pp. 155-163. https://doi.org/10.1016/j.jcm.2016.02.012.

23. Wu, Xiaoshi, et al. "Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis" *arXiv preprint arXiv:2306.09341*, 2023. https://arxiv.org/abs/2306.09341.

24. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.

25. Guo, Hang; Li, Jinmin; Dai, Tao; Ouyang, Zhihao; Ren, Xudong; Xia, Shu-Tao. (2024). MambaIR: A Simple Baseline for Image Restoration with State-Space Model. arXiv:2402.15648. https://arxiv.org/abs/2402.15648.

26. Suvorov, Roman; Logacheva, Elizaveta; Mashikhin, Anton; Remizova, Anastasia; Ashukha, Arsenii; Silvestrov, Aleksei; Kong, Naejin; Goka, Harshith; Park, Kiwoong; Lempitsky, Victor. (2021). Resolution-robust Large Mask Inpainting with Fourier Convolutions. arXiv:2109.07161. https://arxiv.org/abs/2109.07161.

27. Lin, Lawrence I-Kuei. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics. https://doi.org/10.2307/2532051.

28. Roberts, David R., et al. “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure.” *Ecography*, vol. 40, no. 8, 2017, pp. 913-929. https://doi.org/10.1111/ecog.02881.

29. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.

30. Varma, Sudhir and Richard Simon. “Bias in error estimation when using cross-validation for model selection.” *BMC Bioinformatics*, vol. 7, no. 1, 2006. https://doi.org/10.1186/1471-2105-7-91.