Population Variation and Equity
Volume 1 · Number 1 · 2026

A Common Sharpness Association Hides Modest Scene Variation in Smartphone Quality Ratings

Austin Reynolds[1] · Jacob Bennett[2] · Evan Anderson[3]

[1] Canada · [2] Germany · [3] United Kingdom

Original publication: November 3, 2026 | Article 694

Abstract

Sharpness ratings accompany overall quality strongly, but their conditional association need not be identical across photographic scenes. Among 8,354 SPAQ images with an unambiguous dominant scene label, the common association was 6.15 quality-score units per ten sharpness-score units after adjustment for four other rated attributes. Scene interactions rejected a common-slope restriction, yet increased in-sample R-squared by only 0.00018. The small fit change limits the practical interpretation of that statistical heterogeneity. Ambiguous scene assignments were excluded rather than resolved by an arbitrary label order.

Exploratory observational draft. Associations among released human-score summaries; no causal optical effect or evaluated quality-assessment network.

A Common Sharpness Association Hides Modest Scene Variation in Smartphone Quality Ratings

> Exploratory observational draft. Associations among released human-score summaries; no causal optical effect or evaluated quality-assessment network.

Abstract

Sharpness ratings accompany overall quality strongly, but their conditional association need not be identical across photographic scenes. Among 8,354 SPAQ images with an unambiguous dominant scene label, the common association was 6.15 quality-score units per ten sharpness-score units after adjustment for four other rated attributes. Scene interactions rejected a common-slope restriction, yet increased in-sample R-squared by only 0.00018. The small fit change limits the practical interpretation of that statistical heterogeneity. Ambiguous scene assignments were excluded rather than resolved by an arbitrary label order.

1. A scene-dependent relationship between recorded judgments

Sharpness is one recognizable component of photographic quality, but a numerical sharpness score is not the same object as a physical change in an image. It is a judgment recorded under a particular rating procedure. When it is related to an overall quality judgment, the association can reflect common visual properties, shared evaluators, correlated attributes, and scene composition. Understanding that relationship requires naming both the measurement level and the scope of any subgroup comparison.

Our research question asks whether the conditional association between released sharpness ratings and overall quality MOS differs across unambiguous dominant scene groups. The Smartphone Photography Attribute and Quality database Fang et al., 2020 supplies overall scores, five attribute scores, and fractional scene annotations for 11,125 photographs. We use its metadata rather than pixels or newly collected judgments. The analysis describes how recorded quality dimensions move together within a declared cohort; it does not estimate the effect of sharpening an image.

Meng, 2026 proposes a two-stream vision-language framework that separates semantics from distortions in generated-image quality assessment. That distinction motivates examining whether scene composition qualifies a low-level attribute-quality relationship. We do not implement the framework, evaluate generated images, or use these regressions to rank its architecture. The present evidence concerns smartphone photographs and aggregate human ratings. It can motivate a measurement question without supplying an experiment on a different model and image population.

The analytical proposition is that a common sharpness coefficient may conceal scene-specific variation, even after other rated attributes are represented. A second proposition concerns interpretation: detectable interaction need not add much explanatory fit. We therefore report both the joint equal-slope test and the change in in-sample R-squared. A result can support some statistical heterogeneity while offering little evidence that a more complex model would meaningfully improve future quality prediction.

2. Keeping the attribute and the optical intervention separate

The source distinguishes human opinions about brightness, colorfulness, contrast, noisiness, and sharpness from EXIF fields. We use the first Brightness score column, not the separately named EXIF brightness field. All five predictors are released rating summaries. The Noisiness field is retained under its source label; we do not reinterpret its coefficient as the causal effect of adding a measured quantity of sensor noise. Numerical rating units and physical acquisition units have different meanings.

Measurement-scale discussion Stevens, 1946 motivates stating that a ten-unit score difference is treated as a linear numerical increment. Agreement methodology Bland and Altman, 1986 also helps distinguish a strong association from interchangeability of two measurements. A high sharpness-quality slope would not make one rating a complete substitute for the other, especially after controlling several correlated scores. Measurement-error reasoning Carroll et al., 2006 further cautions that aggregate opinions can contain shared method structure, which the available image-level table does not identify separately.

Nor does adjustment automatically produce a causal interpretation. Causal-diagram reasoning Greenland et al., 1999 clarifies why including several observed variables is not the same as identifying an intervention. Attributes can reflect the same latent image properties, and overall quality may be judged through several overlapping impressions. We have no randomized sharpening treatment, image-editing history, or observer-level joint response data. Transparent cohort and variable reporting, consistent with the observational-reporting motivation of von Elm et al., 2007, is more appropriate than a causal account of what sharpness contributes to perception.

The distinction also applies to automated quality assessment. Natural-image quality estimation Mittal et al., 2013 and BRISQUE Mittal et al., 2012 derive model-based image descriptors, while feature similarity Lin Zhang et al., 2011 and learned perceptual similarity Zhang et al., 2018 define other evaluative constructions. Our sharpness predictor is none of those metrics. It is already a human-rated attribute. A strong conditional association in this table cannot certify a pixel-based predictor, and using the recorded attribute to explain MOS is not an operational method for evaluating an unrated new image.

3. A cohort with unambiguous scene assignment

The source has nine scene-share fields: animal, cityscape, human, indoor scene, landscape, night scene, plant, still life, and others. Their row sums equal one within the checked numerical tolerance. Several images receive the same maximum share for more than one scene. A first-in-column-order rule would assign those records deterministically, but its result would depend on the spreadsheet's ordering rather than a unique dominant annotation. That is an undesirable source of artificial subgroup composition.

The primary analysis therefore excludes 2,771 images with tied maximum shares, leaving 8,354 images. An initial source-order assignment was calculated during the metadata audit and is retained as exploratory provenance; the restricted analysis was adopted before manuscript drafting to avoid arbitrary primary labels. This was an exploratory analytical decision, not a preregistered protocol. The excluded images are not described as defective photographs. They simply do not satisfy the present single-dominant-scene definition.

Group sizes are unequal. The cohort has eighty-one animal images, 1,586 cityscape, 615 human, 495 indoor, 965 landscape, 662 night, 985 plant, 2,519 still-life, and 446 others. These counts are properties of the declared dominance rule, not prevalence estimates for smartphone photography. Images with mixed scene interpretations are deliberately outside the main estimand. The subgroup contrast is consequently conditional on a relatively unambiguous dominant annotation, rather than a universal semantic partition of all images.

All included images have complete overall MOS and attribute-score fields. No EXIF plausibility restriction, missing-value imputation, or image-content filtering is applied. This differs from a physical acquisition analysis that requires valid exposure metadata. The source image names are unique, and the model matrices have full numerical rank. Each attribute is divided by ten so its regression coefficient is expressed per ten released score units; overall MOS remains in its original numerical units.

4. Two models for the sharpness association

Model M1 relates overall MOS to sharpness, brightness, colorfulness, contrast, and noisiness ratings, with scene-specific intercepts. It imposes one common sharpness slope across the nine groups. Model M2 adds all eight scene-by-sharpness interactions, retaining common slopes for the other four attributes. Animal is the reference scene. The second model therefore allows the sharpness association to vary while keeping another part of the specification deliberately constrained. It is not a fully separate regression for every attribute in every scene.

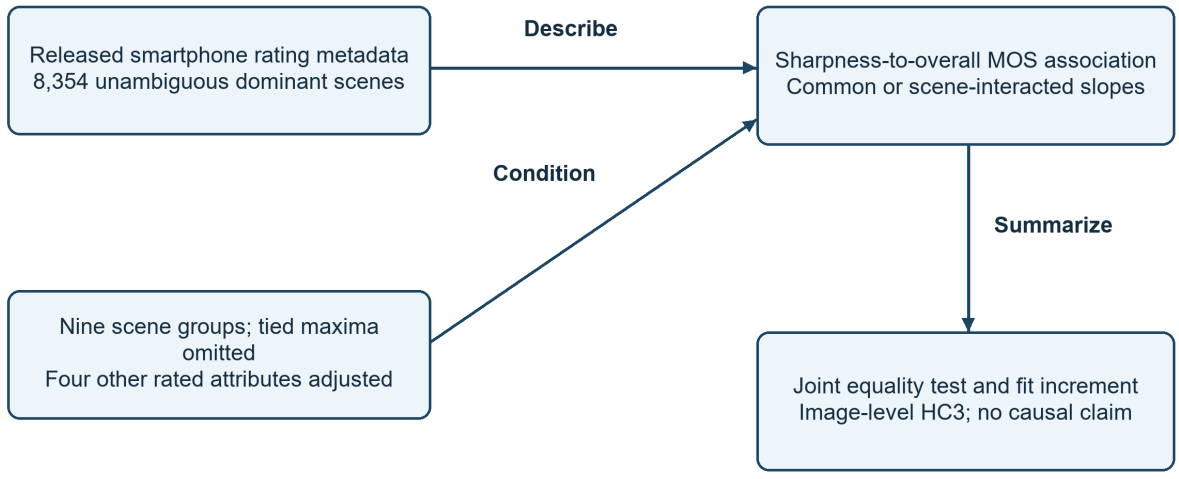


Figure 1. Scene grouping as a modifier of a conditional rating association.

Figure 1 connects the recorded sharpness score with overall MOS and shows the dominant-scene grouping as the source of slope variation. Other rated attributes enter as covariates. Arrows denote the fitted association structure rather than a physical intervention pathway. A joint Wald test evaluates whether the eight interaction coefficients are all zero. The corresponding null is common sharpness slope under the stated linear model, not equality of every visual experience or equivalence of all scene populations.

Both regressions use HC3 standard errors and normal-reference confidence intervals, following robust covariance principles White, 1980. The reported N is 8,354 images. Guidance on clustered inference Cameron and Miller, 2015 motivates acknowledging what cannot be modeled: the merged table used here does not identify the full observer, photographer, or phone dependence structure needed for corresponding cluster estimates. Image-level HC3 is therefore a working conditional summary. It does not make unknown repeated-rater or device effects disappear.

The group slopes and interaction intervals are pointwise. We do not claim that they form a simultaneous confidence set over every scene comparison. An influence analysis following Cook, 1977 could examine unusually consequential images, but no such diagnostic is used to remove observations here. The joint test

and R-squared comparison use the same complete primary cohort, avoiding a change of sample being mistaken for an effect of adding interactions. No validation holdout is used, so fit is explicitly in sample.

Full numerical rank is only a basic identification check for the fitted coefficient vector. It does not establish that sharpness and the other attribute scores vary independently, or that every scene occupies the same covariate range. The phrase holding other attributes constant describes a regression comparison, not an intervention that an observer actually experienced. If combinations of scores are rare in a subgroup, interpreting a large hypothetical change can require extrapolation. We therefore keep the reported slopes in their numerical association role instead of turning them into recommended weights for manipulating an image.

The common slopes imposed on the four remaining attributes are another boundary. They reduce the number of interactions and define a focused question about sharpness, but a scene-specific relationship in brightness or noisiness could alter the fitted sharpness contrast. A model allowing every attribute to vary by scene would answer a broader question and would need separate scrutiny of support, conditioning, and generalization. It is not described as a completed robustness check here. The present conclusion remains attached to the two formulas that were actually fitted.

5. Statistical heterogeneity with little additional fit

M1 estimates a common sharpness slope of 6.1478256 MOS units per ten sharpness units, with standard error 0.0535950 and interval from 6.0427812 to 6.2528699. The strong association is conditional on scene intercepts and four other rated attributes. It should not be read as the improvement from an image-editing operation. All predictors and the outcome are score summaries from the same source setting, which can make numerical explanation easier than predicting quality from independently measured image features.

The joint equal-slope restriction in M2 gives Wald chi-square 32.4535262 with eight degrees of freedom and p-value approximately 0.0000772. Under the model's covariance assumptions, the common-slope restriction is not supported by that test. However, R-squared rises only from 0.9566453 to 0.9568220, an increase of about 0.0001767. The large sample makes modest deviations detectable. The fit increment is essential context for interpreting what the interaction result adds.

Table 1. Common and scene-interacted sharpness associations with overall MOS

Model	Term	Estimate (SE)	95% CI	N
M1	Sharpness score, per ten units	6.1478 (0.0536)	[6.0428, 6.2529]	8354
M1: fit	R-squared = 0.9566	-	-	8354
M2	Sharpness score, per ten units	6.0051 (0.2656)	[5.4845, 6.5258]	8354
M2	Cityscape slope minus Animal	0.0892 (0.2638)	[-0.4278, 0.6062]	8354
M2	Human slope minus Animal	0.0979 (0.2711)	[-0.4334, 0.6291]	8354
M2	Indoor scene slope minus Animal	0.1641 (0.2747)	[-0.3744, 0.7026]	8354
M2	Landscape slope minus Animal	0.0715 (0.2680)	[-0.4537, 0.5967]	8354

Model	Term	Estimate (SE)	95% CI	N
M2	Night scene slope minus Animal	-0.2613 (0.2769)	[-0.8039, 0.2814]	8354
M2	Plant slope minus Animal	0.2771 (0.2657)	[-0.2436, 0.7978]	8354
M2	Still-life slope minus Animal	0.1987 (0.2618)	[-0.3145, 0.7119]	8354
M2	Others slope minus Animal	-0.2269 (0.3226)	[-0.8592, 0.4054]	8354
M2: fit	R-squared = 0.9568	-	-	8354

Model labels: M1 = Common sharpness association adjusted for scene and other attributes; M2 = Scene-specific sharpness association with common other-attribute slopes.

Table 1 reports OLS coefficients, HC3 standard errors, normal-reference 95% intervals, and N8354. Both models include scene intercepts plus brightness, colorfulness, contrast, and noisiness scores. M1 has a common sharpness slope; M2 adds eight sharpness-by-scene interactions. M2's sharpness main effect is the animal-group slope, and each interaction is a difference from that reference. Displayed fit is in-sample R-squared. No causal design, neural predictor, held-out score, or observer-level reliability estimate is included.

The conditional scene slopes range from about 5.744 in the night group to 6.282 in the plant group. Animal is 6.005, cityscape 6.094, human 6.103, indoor 6.169, landscape 6.077, still life 6.204, and others 5.778. These derived slopes combine the main effect with the relevant interaction. They are not separate unadjusted correlations. Their values describe the declared linear association after retaining the common covariate structure, and their ordering is not a quality ranking of the scene categories.

Every displayed interaction interval against animal includes zero, even though the joint test rejects the common-slope restriction. This is not a numerical inconsistency. The reference group has only eighty-one images and a relatively imprecise slope; its uncertainty is shared across the reference-based contrasts. A joint test uses the covariance among those contrasts and can identify variation among other groups more precisely. We avoid selecting a favorable pair after looking at all slopes and presenting it as a prespecified independent discovery.

6. Boundaries of the quality construct

The scene labels are a coarse numerical representation of human annotations. Excluding tied maxima helps avoid an arbitrary assignment but narrows the cohort toward clearer dominance. It does not make semantics perfectly measured. A fractional-membership model could retain the ambiguous records, but would change the interpretation of an interaction and require another specification. The present result should not be generalized to all mixed-content photographs or treated as evidence that the nine groups form mutually exclusive natural categories.

Other quality databases also differ in how they construct observations. KonIQ-10k Hosu et al., 2019 and AVA Murray et al., 2012 concern different natural-image evaluation settings. KADID-10k Lin et al., 2019 and TID2013 Ponomarenko et al., 2015 organize distortion-based evidence. AGIQA-3K Li et al., 2023 and AIGIQA-20K Li et al., 2024 concern generated-image scores. None is added to the regression sample here. Their different targets make it inappropriate to transport a smartphone-attribute slope into a universal semantic-distortion weighting without a new analysis.

Likewise, NIMA Talebi and Milanfar, 2017, MUSIQ Ke et al., 2021, MANIQA Yang et al., 2022, CLIP-based quality assessment Wang et al., 2022, and structure-texture similarity Ding et al., 2020 provide other modeling constructions. They might be evaluated against subgroup-specific outcomes in future work, but no implementation or prediction from them appears in this study. Text-faithfulness evaluation Hu et al., 2023, Human Preference Score Wu et al., 2023, and Pick-a-Pic Kirstain et al., 2023 also address related evaluative concerns whose decision rules are not recoverable from the five SPAQ attribute means alone.

The practical implication is consequently restrained. A single sharpness association is a useful summary of this cohort, and a richer interaction specification detects some deviation from it. The tiny fit increment cautions against describing that deviation as a major predictive improvement. Common-method structure, unknown evaluator dependence, unequal group counts, and the exclusion of ambiguous scenes remain central limitations. An external predictive comparison or an intervention on actual image sharpness would be needed for stronger claims.

7. A traceable subgroup analysis

The analysis retains the source file identity, group-assignment rule, excluded tie count, all image-level predictors, complete model formulas, derived scene slopes, and the joint restriction test. Dataset documentation Gebru et al., 2021 supports recording those definitions, while FAIR stewardship Wilkinson et al., 2016 motivates preserving the identifiers and structured results. No source pixels are redistributed by this metadata calculation, and original attribution and source-specific reuse conditions remain attached to the record.

The answer to the research question is therefore twofold. Sharpness-score association varies across the declared unambiguous scene groups under a robust linear model, but allowing that variation adds very little in-sample explanatory fit. Reporting both facts avoids equating statistical detectability with substantive prediction value. The distinction is especially useful when ratings about visual attributes and scene semantics are later combined into a single quality-assessment narrative.

References

1. Murray, Naila; Marchesotti, Luca; Perronnin, Florent. (2012). AVA: A large-scale database for aesthetic visual analysis. 2012 IEEE Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/cvpr.2012.6247954.
2. 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.
3. von Elm, Erik, et al. "The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies." *PLoS Medicine*, vol. 4, no. 10, 2007, pp. e296. https://doi.org/10.1371/journal.pmed.0040296.
4. Wang, Jianyi, et al. "Exploring CLIP for Assessing the Look and Feel of Images." *arXiv preprint arXiv:2207.12396*, 2022. https://arxiv.org/abs/2207.12396.
5. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.
6. Lin, Hanhe, et al. "KADID-10k: A Large-scale Artificially Distorted IQA Database" *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, 2019. https://doi.org/10.1109/qomex.2019.8743252.
7. Fang, Yuming; Zhu, Hanwei; Zeng, Yan; Ma, Kede; Wang, Zhou. (2020). Perceptual Quality Assessment of Smartphone Photography. CVPR 2020, 3677-3686. https://openaccess.thecvf.com/content_CVPR_2020/html/Fang_Perceptual_Quality_Assessment_of_Smartphone_Photography_CVPR_2020_paper.html.
8. 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.
9. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.

10. 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.

11. Hosu, Vlad, et al. "KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment" *arXiv preprint arXiv:1910.06180*, 2019. https://arxiv.org/abs/1910.06180.

12. Li, Chunyi; Kou, Tengchuan; Gao, Yixuan; Cao, Yuqin; Sun, Wei; Zhang, Zicheng; Zhou, Yingjie; Zhang, Zhichao; Zhang, Weixia; Wu, Haoning; Liu, Xiaohong; Min, Xiongkuo; Zhai, Guangtao. (2024). AIGIQA-20K: A Large Database for AI-Generated Image Quality Assessment. arXiv:2404.03407. https://arxiv.org/abs/2404.03407.

13. Yang, Sidi, et al. "MANIQA: Multi-dimension Attention Network for No-Reference Image Quality Assessment." *arXiv preprint arXiv:2204.08958*, 2022. https://arxiv.org/abs/2204.08958.

14. 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.

15. Ding, Keyan, et al. "Image Quality Assessment: Unifying Structure and Texture Similarity." *arXiv preprint arXiv:2004.07728*, 2020. https://arxiv.org/abs/2004.07728.

16. 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.

17. Greenland, Sander, Judea Pearl, and James M. Robins. “Causal Diagrams for Epidemiologic Research.” *Epidemiology*, vol. 10, no. 1, 1999, pp. 37-48. https://doi.org/10.1097/00001648-199901000-00008.

18. Ponomarenko, Nikolay; Jin, Lina; Ieremeiev, Oleg; Lukin, Vladimir; Egiazarian, Karen; Astola, Jaakko; Vozel, Benoit; Chehdi, Kacem; Carli, Marco; Battisti, Federica; Jay Kuo, C.-C.. (2015). Image database TID2013: Peculiarities, results and perspectives. Signal Processing: Image Communication. https://doi.org/10.1016/j.image.2014.10.009.

19. Cameron, A. Colin, and Douglas L. Miller. “A Practitioner’s Guide to Cluster-Robust Inference.” *Journal of Human Resources*, vol. 50, no. 2, 2015, pp. 317-372. https://doi.org/10.3368/jhr.50.2.317.

20. Stevens, S. S.. (1946). On the Theory of Scales of Measurement. Science. https://doi.org/10.1126/science.103.2684.677.

21. Mittal, A.; Moorthy, A. K.; Bovik, A. C.. (2012). No-Reference Image Quality Assessment in the Spatial Domain. IEEE Transactions on Image Processing. https://doi.org/10.1109/tip.2012.2214050.

22. 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.

23. Ke, Junjie, et al. "MUSIQ: Multi-scale Image Quality Transformer" *arXiv preprint arXiv:2108.05997*, 2021. https://arxiv.org/abs/2108.05997.

24. Mittal, A., et al. "Making a “Completely Blind” Image Quality Analyzer" *IEEE Signal Processing Letters*, 2013. https://doi.org/10.1109/lsp.2012.2227726.

25. Talebi, Hossein and Peyman Milanfar. "NIMA: Neural Image Assessment" *arXiv preprint arXiv:1709.05424*, 2017. https://arxiv.org/abs/1709.05424.

26. Hu, Yushi, et al. "TIFA: Accurate and Interpretable Text-to-Image Faithfulness Evaluation with Question Answering." *arXiv preprint arXiv:2303.11897*, 2023. https://arxiv.org/abs/2303.11897.

27. Li, Chunyi, et al. "AGIQA-3K: An Open Database for AI-Generated Image Quality Assessment" *arXiv preprint arXiv:2306.04717*, 2023. https://arxiv.org/abs/2306.04717.

28. 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.

29. Lin Zhang, ; Lei Zhang, ; Xuanqin Mou, ; Zhang, D.. (2011). FSIM: A Feature Similarity Index for Image Quality Assessment. IEEE Transactions on Image Processing. https://doi.org/10.1109/tip.2011.2109730.

30. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.