Journal of Transfer and Generalization
Volume 1 · Number 1 · 2026

Human Attribute Ratings Travel Unevenly into Night Scenes: A Frozen Mapping of Smartphone Photo Quality

Scott Parker[1] · Noah Wilson[2] · Jason Reed[3]

[1] Norway · [2] Belgium · [3] Austria

Original publication: November 1, 2026 | Article 629

Abstract

Five subjective attribute ratings predict overall smartphone-photo MOS closely, even when night-associated images are excluded from fitting. A regression trained on 8,124 nonnight SPAQ photographs achieves MSE 24.89 on 969 night-associated photographs, compared with 439.67 for the training-mean baseline. Its error is nevertheless 6.12 squared rating units higher than on held nonnight photos. The result concerns transfer among human rating summaries, with attributes supplied at prediction time. It does not establish a blind image-quality model or a causal effect of night photography.

Human Attribute Ratings Travel Unevenly into Night Scenes: A Frozen Mapping of Smartphone Photo Quality

Abstract

Five subjective attribute ratings predict overall smartphone-photo MOS closely, even when night-associated images are excluded from fitting. A regression trained on 8,124 nonnight SPAQ photographs achieves MSE 24.89 on 969 night-associated photographs, compared with 439.67 for the training-mean baseline. Its error is nevertheless 6.12 squared rating units higher than on held nonnight photos. The result concerns transfer among human rating summaries, with attributes supplied at prediction time. It does not establish a blind image-quality model or a causal effect of night photography.

1. A strong association can still have a domain boundary

Overall photo quality and judgments of sharpness, brightness, colorfulness, contrast, and noisiness are closely related by construction of the evaluation task. Their association does not imply that a mapping learned from one set of scenes behaves identically in another. Night-associated photography is a useful observational boundary: features of illumination and subject matter can change together, and the meaning of a favorable attribute rating may depend on the visual context. The research question is whether a relationship fitted without any night-associated photographs remains useful for predicting their overall quality ratings.

This question requires a careful description of what is transferred. We predict an aggregate human judgment using five other aggregate human judgments of the same photograph. The attributes are available to the prediction rule. No pixel model estimates them. The two-stream semantic and distortion design proposed by Meng, 2026 supplies a reason to keep distinct dimensions visible, but it is not executed here. Our experiment evaluates the portability of a simple relationship among released scores, rather than the robustness of that architecture or a newly trained visual evaluator.

The data come from the smartphone photography resource in Fang et al., 2020. Night association is defined by any positive released Night scene share, including images with mixed scene annotations. That definition retains ambiguous mixtures rather than selecting only photographs whose unique dominant category is night. It is an operational grouping of existing labels, not a measurement of illuminance, a verified capture timestamp, or a randomized lighting intervention. Its value is that the grouping can be reconstructed exactly and excluded from the training branch before any held loss is computed.

2. Subjective attributes as an information-rich input

Blind quality methods such as Mittal et al., 2012 and Mittal et al., 2013 derive assessment information from visual statistics. Multiscale processing in Ke et al., 2021 provides another image-based approach. Our inputs are substantially different: an observer has already summarized five dimensions of each held image. A large improvement over a constant baseline is therefore plausible and should not be reported as comparable to automatic blind assessment. It reflects the predictive content of human attribute summaries and whatever shared evaluation structure connects them to the overall MOS.

Other databases show why that boundary matters. Hosu et al., 2019 concerns broad image-quality assessment, Ponomarenko et al., 2015 organizes artificially distorted images, and Gu et al., 2020 addresses perceptual restoration. Their rating tasks and image populations differ from smartphone photographs with scene-share labels. We do not pool those resources or infer that an attribute mapping calibrated on one would retain its numerical coefficients on the others. Even within SPAQ, a

night-associated group can differ simultaneously in devices, subjects, exposure strategies, and observer expectations, leaving several plausible explanations for a domain error gap.

The scale perspective in Stevens, 1946 asks what operations the rating representation supports. We use the source numerical scores in their native units and fit a linear relationship, without treating equal rating increments as equal changes in physical image properties. Carroll et al., 2006 further distinguishes measurement uncertainty from regression error conditional on observed covariates. Here both sides are aggregate judgments with possible common-method dependence. A small residual does not prove that the model independently recovers true optical sharpness or that the five attributes exhaust the concept of overall quality.

Agreement is also more demanding than association. Bland and Altman, 1986 and Lin, 1989 provide relevant measurement perspectives because a score mapping can be strongly associated with its outcome while remaining systematically miscalibrated in a subgroup. Our held squared-error comparison responds to that concern by evaluating actual prediction discrepancies. It does not make the mapping a replacement for the original human judgment. The night versus nonnight error contrast is descriptive, and a positive gap can coexist with a large improvement over the much less informative constant baseline.

3. Excluding night-associated photos before fitting

The local SPAQ metadata table contains 11,125 unique image names and complete values for overall MOS, the five selected subjective attribute fields, and scene shares. We use the first Brightness field, which is a subjective score, rather than the separate EXIF-related brightness field. We do not require complete camera settings. A photograph is night-associated when its released Night scene share is greater than zero. This produces 969 night-associated images and 10,156 images with zero night share, all retained for the appropriate branch of the design.

The zero-night-share images are ordered by the SHA256 digest of `scenetransfer20261108:` followed by the image name. The first eighty percent, 8,124 images, form the fitting set. The remaining 2,032 form a held nonnight comparison set. All 969 night-associated images are held from fitting. Thus the final prediction record contains 3,001 images in two observational domains. The split does not use MOS, attribute values, image appearance, or favorable prediction errors. It is a single reproducible menu rather than a repeated-split performance estimate.

Structured validation principles in Roberts et al., 2017 make the training boundary more meaningful than a pooled in-sample scene interaction. The leakage analysis in Kapoor and Narayanan, 2023 motivates preventing held outcomes from choosing the formula or an adjustment for night scenes. We accordingly fit one common model to nonnight data and freeze it. Varma and Simon, 2006 describes the danger of using evaluation outcomes for model selection; no search over scene definitions, transformations, or regression penalties is performed against the held losses in this experiment.

The model is ordinary least squares with an intercept and the five raw subjective attributes. Its six-column design has full rank. No standardization estimated from held records is needed, and predictions are not clipped to held MOS extrema. The constant reference predicts the nonnight training mean, 51.503677, for every held image. Both procedures receive the same training outcomes, while only the attribute model receives the held subjective attribute scores. That information advantage is part of the specified comparison and should remain visible when interpreting its large loss reduction.

4. A research model with two distinct contrasts

Figure 1 separates fitting, held-domain classification, and evaluation. The first contrast asks whether the frozen attribute mapping improves squared error over the training-mean baseline among night-associated images. The second asks whether the attribute model's held MSE differs between the night-associated and nonnight groups. They answer different questions. An informative model can beat a constant in both groups while performing worse in one of them. We therefore do not use the first contrast to dismiss an observed domain gap in the second.

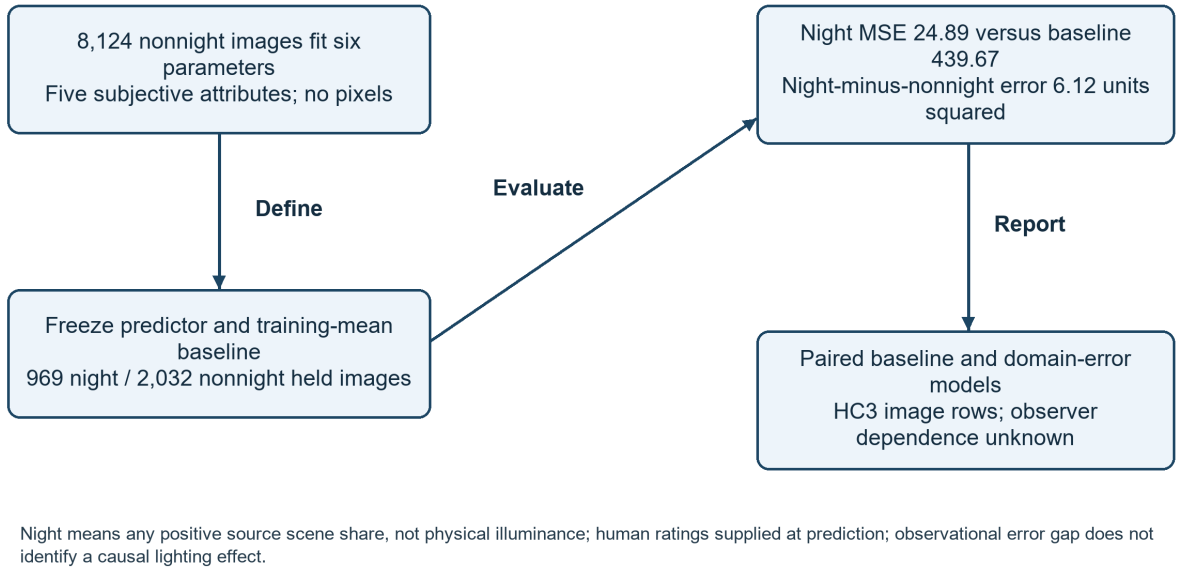


Figure 1. *A six-parameter relationship among subjective scores is fitted only on nonnight images, then applied without adjustment to two held groups. Scene association changes the evaluation subgroup, while the predictor remains frozen.*

The first inferential summary is intercept-only OLS on the paired squared-error difference for 969 night-associated photographs. The second is OLS of attribute-model squared error on an intercept and the night indicator across all 3,001 held photographs. Both use HC3 covariance and a t reference based on the fitted residual degrees of freedom. White, 1980 motivates heteroskedasticity-aware covariance. A stronger dependence model remains unavailable because the released table does not provide the observer, photographer, or phone groupings needed to represent the possible shared sources of rating error.

The error contrast contains no additional controls, because its purpose is to describe the observed domain gap for one frozen predictor. The regression interpretation in Ding, 2020 helps distinguish a conditional coefficient from an unadjusted group difference, and Angrist and Pischke, 2009 emphasizes the requirements for causal identification. Those requirements are not met by this scene indicator. Adding arbitrary covariates after inspecting the error gap would change the estimand without randomizing night association. We therefore report the actual gap with its working uncertainty and describe composition as a possible explanation.

Table 1. *M1 is the paired attribute-model minus constant-baseline squared-error mean on night-associated images. M2 compares attribute-model squared errors between held groups; its night coefficient is the unadjusted night-associated minus nonnight MSE gap. Covariance is HC3. There are no controls in these contrast fits. The underlying predictor has five attribute slopes and an intercept.*

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	-414.779865 (14.206010)	[-442.657990, -386.901740]	969
M1: fit	R-squared = 0.0000	-	-	969
M2	Night-associated minus nonnight MSE	6.121470 (1.468462)	[3.242176, 9.000764]	3001
M2: fit	R-squared = 0.0065	-	-	3001

Model labels: M1 = Attribute model minus training-mean MSE on night-associated photos; M2 = Night-associated minus held nonnight attribute-model MSE.

5. Useful transfer with a remaining error increase

The fitting regression has an in-sample R-squared of 0.954792 on the nonnight training images. This is a description of a relationship among human scores, not an independent visual benchmark achievement. On held nonnight images, attribute-model MSE is 18.767777, against 450.082171 for the constant reference. On night-associated images it is 24.889247, against 439.669112. The attribute information is therefore highly useful for this particular score-prediction task in both observed domains, despite the complete exclusion of night-associated images from fitting.

Among night-associated photos, the paired MSE change is negative 414.779865, with standard error 14.206010 and a 95% interval from negative 442.657990 to negative 386.901740. That reduction should be read relative to the deliberately simple constant reference. It does not compare two equally informed image models. A practical workflow that already has five human attribute ratings has paid an information-acquisition cost that a blind image assessor generally seeks to avoid. We did not measure that cost or claim that the regression saves annotation labor.

The night-associated error gap is positive 6.121470 squared native rating units, with an interval from 3.242176 to 9.000764. A positive gap is compatible with the mapping remaining useful. It indicates that the frozen linear relationship leaves more squared residual variation in the night-associated group under the present source composition. The result does not identify whether changed attribute meanings, nonlinearities, rating precision, scene content, or another unobserved difference accounts for that variation. In particular, it cannot isolate a causal effect of darkness on predictive error.

Squared error is the executed primary loss. Gneiting and Raftery, 2007 motivates specifying the prediction objective, and Willmott and Matsuura, 2005, Chai and Draxler, 2014, and Hodson, 2022 explain why an error summary's interpretation depends on its loss and units. We report squared rating units throughout. No absolute-error or image-ranking comparison is inferred from those results. The training mean, the five-attribute mapping, and the group contrast would need to be recomputed if a different decision objective were adopted; a favorable result under one criterion does not establish dominance under every criterion.

6. What a domain label and a diagnostic cannot establish

The held groups are not disjoint semantic universes. Images with a positive night share can also carry other scene labels, and zero night share means only that this annotation component is absent. The analysis deliberately keeps mixed assignments. This preserves the source annotation structure but weakens any intuitive interpretation of the groups as pure day and pure night. It also means the experiment cannot distinguish an illumination domain from the broader composition of photographs selected into that annotation category.

Diagnostic and interpretability tools would answer additional questions. Cook, 1977 identifies influence in a fitted regression, and Lundberg and Lee, 2017 offers a framework for attributing predictions to inputs. Neither is used to rewrite or validate the source labels here. The label-noise perspective in Natarajan et al., 2013 also does not justify declaring an unexpected rating incorrect. A coefficient on subjective noisiness, for example, remains an association conditional on the other supplied ratings and their source coding. It cannot be interpreted as a physical experiment that adds sensor noise while holding all other image properties fixed.

Alternative representations remain untested. The dimension-reduction approach in Turk and Pentland, 1991 concerns image representations in a different task, and the regularization framework in Hoerl and Kennard, 1970 could stabilize a different regression. Neither establishes how a richer model would transfer across these scene groups. Likewise, the shift concerns in Ovadia et al., 2019 motivate evaluating changed populations but do not provide a confidence guarantee for our single local split. New photographers, devices, or observers could change the mapping even if their scene annotation were the same.

The retained split, formula, coefficients, and per-image losses follow the documentation aims in Gebru et al., 2021 and the reuse principles in Wilkinson et al., 2016. They make the unusually informative inputs and the night-exclusion rule explicit. These records distinguish this held-domain study from an in-sample analysis of scene-specific attribute slopes. A future prospective assessment could estimate attributes automatically or collect independent observers, but that would require a new evidence chain rather than a reinterpretation of the present human-score mapping.

One design choice deserves particular attention: the nonnight holdout and the night-associated holdout are evaluated by the same fitted function, but their outcome distributions need not match. The error-gap coefficient therefore contains every way their joint attribute and MOS distributions differ. It cannot distinguish lack of predictor support from a changed conditional relationship. Checking support would be a useful additional analysis, yet its result is not silently assumed. The present comparison establishes a concrete held discrepancy for the observed groups while leaving the decomposition of that discrepancy open.

The single hash split also means that the displayed standard errors do not include uncertainty in the six fitted coefficients. Retraining on another nonnight subset might change predictions and the resulting gap. The image-level loss regressions condition on the stored function, which makes their interpretation narrower than a full assessment of the learning procedure. This distinction matters most when applying the method to future collections: repeat training, source grouping, and independent observer evidence would be needed to estimate the variability of that broader procedure.

7. Conclusion

A five-attribute relationship learned without night-associated photos remains informative for their overall quality ratings, yet its held squared error increases relative to nonnight photos. The combination supports evaluating usefulness and domain stability separately. It also shows why a strong score association should retain its information boundary: the predictor uses human attribute ratings and measures transfer among aggregate judgments, with no claim to automate visual quality assessment or identify the cause of the observed night-associated error gap.

References

1. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.
2. 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.
3. Chai, T.; Draxler, R. R.. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature. Geoscientific Model Development. https://doi.org/10.5194/gmd-7-1247-2014.
4. Natarajan, Nagarajan; Dhillon, Inderjit S.; Ravikumar, Pradeep; Tewari, Ambuj. (2013). Learning with Noisy Labels. Advances in Neural Information Processing Systems 26. https://papers.nips.cc/paper_files/paper/2013/hash/3871bd64012152bfb53fdf04b401193f-Abstract.html.
5. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.
6. 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.
7. 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.
8. Kapoor, Sayash; Narayanan, Arvind. (2023). Leakage and the reproducibility crisis in machine-learning-based science. Patterns. https://doi.org/10.1016/j.patter.2023.100804.
9. 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.
10. 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.
11. Mittal, A., et al. "Making a “Completely Blind” Image Quality Analyzer" *IEEE Signal Processing Letters*, 2013. https://doi.org/10.1109/lsp.2012.2227726.
12. Hodson, Timothy O.. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. Geoscientific Model Development. https://doi.org/10.5194/gmd-15-5481-2022.
13. 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.
14. Lin, Lawrence I-Kuei. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics. https://doi.org/10.2307/2532051.
15. Stevens, S. S.. (1946). On the Theory of Scales of Measurement. Science. https://doi.org/10.1126/science.103.2684.677.
16. Willmott, CJ; Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research. https://doi.org/10.3354/cr030079.

17. 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.

18. Turk, M.A.; Pentland, A.P.. (1991). Face recognition using eigenfaces. Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/cvpr.1991.139758.

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

20. Hoerl, Arthur E.; Kennard, Robert W.. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics. https://doi.org/10.1080/00401706.1970.10488634.

21. 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.

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

23. Lundberg, Scott M., and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems*, vol. 30, 2017. https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

24. Angrist, Joshua D.; Pischke, Jörn-Steffen. (2009). Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press. https://press.princeton.edu/books/paperback/9780691120355/mostly-harmless-econometrics.

25. 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.

26. 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.

27. 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.

28. Ovadia, Yaniv, et al. "Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift." *Advances in Neural Information Processing Systems*, vol. 32, 2019. https://arxiv.org/abs/1906.02530.

29. Ding, Peng. (2020). The Frisch--Waugh--Lovell Theorem for Standard Errors. arXiv:2009.06621. https://arxiv.org/abs/2009.06621.

30. 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.