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

Knowing Quality Helps Predict Alignment, but Does Not Replace It: Human-Score Transfer across Generators

Lucas Reed[1] · Jeffrey Hayes[2] · Ryan King[3]

[1] Lithuania · [2] Iceland · [3] Luxembourg

Original publication: November 1, 2026 | Article 693

Abstract

Can human quality ratings stand in for text-alignment ratings on a different generator? Six leave-one-generator fits on AGIQA-3K map quality MOS to alignment MOS while retaining shared prompt content in training. Every held generator improves over a constant reference, but residual errors and signed biases remain. Equal-generator MSE falls from 1.545 to 0.431; its mean difference has a wide interval, whereas the MAE difference is negative under a small-cohort working summary. The experiment transfers a relationship between human scores, with no automatic image assessment or unseen-prompt claim.

Knowing Quality Helps Predict Alignment, but Does Not Replace It: Human-Score Transfer across Generators

Abstract

Can human quality ratings stand in for text-alignment ratings on a different generator? Six leave-one-generator fits on AGIQA-3K map quality MOS to alignment MOS while retaining shared prompt content in training. Every held generator improves over a constant reference, but residual errors and signed biases remain. Equal-generator MSE falls from 1.545 to 0.431; its mean difference has a wide interval, whereas the MAE difference is negative under a small-cohort working summary. The experiment transfers a relationship between human scores, with no automatic image assessment or unseen-prompt claim.

1. Predictive information is weaker than interchangeability

Two rating dimensions can contain useful information about one another without being interchangeable. A visually poor generated image may also fail to express its requested content, but a polished image can still depict the wrong object or relationship. The research question is whether a simple mapping from quality MOS to alignment MOS, learned on other generators, predicts alignment better than a constant for a held generator. A favorable answer would establish information transfer among aggregate scores, not prove that the second rating is redundant.

The distinction is relevant to Meng, 2026, which proposes a semantic and distortion separation for generated-image assessment. Our predictor does not implement those streams or observe an image. It is supplied the held image's existing human quality MOS and tries to predict its human alignment MOS. This unusual information boundary is intentional: it isolates the transportability of a score relationship from the much harder task of estimating either score automatically. A successful mapping would still require collecting quality judgments before it could be used.

The analysis uses the public score table in Li et al., 2023, containing 2,982 images from six source generator prefixes and 300 exact prompts. The generator groups have unequal sizes. Rather than treating every held image as an independent test of a broad substitution claim, we calculate a separate held result for each generator and summarize those six results. Prompt strings represented in the held generator remain in other-generator training. The design concerns generator transfer for familiar prompt content, with no assertion that text content itself is unseen.

2. Why a mapping can move between groups and still be biased

The measurement framework in Carroll et al., 2006 distinguishes imperfect observed scores from latent constructs. Both fields here are aggregate human judgments, with potential shared observer and task structure. Neither is an error-free reading of true visual quality or semantic correctness. Stevens, 1946 adds a numerical-scale consideration: we use the released scores in their original coding and fit a linear mapping. That assumption makes the calculation transparent, but it does not imply that a one-unit difference means the same thing across fields, generators, or hypothetical rating populations.

The agreement perspectives in Bland and Altman, 1986 and Lin, 1989 explain why a useful slope is insufficient for substitution. A mapping can reduce average prediction loss while retaining a generator-specific offset, or while making individual errors too large for the intended decision. We therefore save signed held errors as well as absolute and squared losses. No correlation coefficient alone is treated as proof that alignment ratings can be eliminated. Independent-rater agreement, as studied in

Fleiss, 1971, cannot be reconstructed because the released aggregate table does not provide individual rater links.

Domain shift supplies another distinction. Ovadia et al., 2019 motivates examining how uncertainty behaves when evaluation conditions change, while Roberts et al., 2017 emphasizes validation under structured data. Our held unit is a source generator category, not a new observer population or a new prompt universe. The folds deliberately retain overlapping prompt content and share many training images. This design is appropriate for the narrow generator question but cannot support a claim that the method generalizes to entirely novel requests, independent future generators, or the joint absence of both familiar content and familiar systems.

Image-based assessors illustrate how different the present inputs are. Talebi and Milanfar, 2017, Ke et al., 2021, and Yang et al., 2022 learn assessment from visual representations, and Wang et al., 2022 uses a vision-language representation setting. Ding et al., 2020 also concerns a perceptual comparison distinct from a human-score mapping. We execute none of those methods. Their role in the argument is to prevent a low error from being mislabeled as a blind quality or alignment result. The known quality MOS supplies information that those systems would ordinarily have to infer from the visual input.

3. Six explicitly separated fitting exercises

For each generator prefix, all of its records form the held group and every other generator's records form training. A two-parameter OLS model predicts alignment MOS from an intercept and quality MOS. A constant baseline uses the mean alignment MOS of that same training set. Both rules are frozen before generating predictions for the held group. No generator-specific correction is fitted from held outcomes, and predictions are not clipped to the held group's observed score range.

The held sizes are 300 for AttnGAN, 290 for DALLE2, 300 for glide, 592 for midjourney, 1,200 for sd1.5, and 300 for xl2.2. Corresponding training sizes therefore range from 1,782 to 2,692 images. All exact prompt strings present in each held group are also represented among other-generator training records: 290 to 300 distinct strings depending on the group. This overlap is reported directly. Removing every such training prompt would answer a different and, for this highly shared prompt set, much more restricted question.

The leakage discussion in Kapoor and Narayanan, 2023 motivates declaring what information crosses the evaluation boundary. Shared prompt content is allowed here, whereas held generator outcomes are not. The error-estimation account in Varma and Simon, 2006 also discourages changing the regression specification after viewing held performance. We therefore use the same two-parameter formula in all six folds. No polynomial degree, robust loss, penalty, generator weighting, or prompt effect is selected by the held results. The experiment is an exploratory transfer audit with a fixed specification, not a tuned contest among score mappings.

Dimensional and regularized approaches could produce different mappings. Pearson, 1901 and Tipping and Bishop, 1999 concern latent representation strategies, and Hoerl and Kennard, 1970 provides a regularization framework for regression. None is needed to estimate the present two-column full-rank design, and none is executed as a competing method. Their relevance is to distinguish this intentionally simple baseline from a general upper bound on how well the human score relationship could be modeled. A remaining error does not prove that quality contains no further predictive information beyond a straight line.

4. Research model and loss comparisons

For each held image, we save the source quality and alignment MOS, the fitted prediction, and the training-mean prediction. Within generator, we calculate mean squared error, mean absolute error, and mean signed error for the fitted mapping, together with the corresponding constant-baseline losses. The paired loss differences are mapping minus constant, so negative values favor using the quality score. Generator-level differences are then averaged equally across the six categories rather than weighted by their unequal image counts.

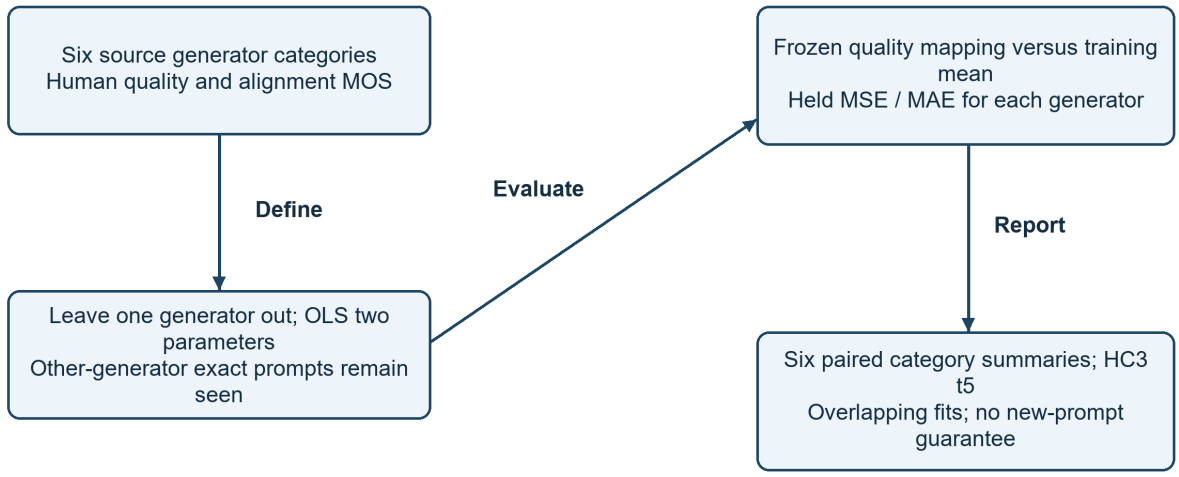


Figure 1. *Other-generator human scores fit a two-parameter mapping. Known held quality MOS predicts held alignment MOS, while shared prompt content remains represented in training. Six generator summaries measure the loss change relative to a training mean.*

Two intercept-only OLS models summarize the generator-level MSE and MAE differences. They use HC3 covariance and a Student reference with five degrees of freedom. The heteroskedasticity-aware perspective in White, 1980 supports allowing unequal spread among category results, but the clustering concerns in Cameron and Miller, 2015 reveal what this does not solve: folds overlap in training and prompts, and source generators may share ancestry or observers. The displayed intervals are small-cohort working descriptions, not independent-fold significance tests or guaranteed coverage for a new-generator population.

Loss choice is substantive. Gneiting and Raftery, 2007 motivates specifying what prediction performance rewards. Willmott and Matsuura, 2005, Chai and Draxler, 2014, and Hodson, 2022 clarify the different weighting of large deviations by absolute and squared error. We calculate both criteria rather than infer one from the other, and we keep their units separate. A generator with a large baseline squared error can have disproportionate influence on the MSE difference even under equal generator weighting. The two summaries consequently need not have the same inferential precision or the same practical interpretation.

Table 1. *Quality-score mapping minus constant-baseline loss, averaged equally over six held generators. Each model is intercept-only OLS with HC3 covariance and a t reference with five degrees of freedom. There are no controls. Underlying fold predictions come from alignment MOS regressed on quality MOS using only other-generator outcomes.*

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	-1.113969 (0.569288)	[-2.577372, 0.349433]	6
M1: fit	R-squared = 0.0000	-	-	6
M2	Mean outcome/contrast	-0.521942 (0.187671)	[-1.004364, -0.039519]	6
M2: fit	R-squared = 0.0000	-	-	6

Model labels: M1 = Quality-score mapping minus constant alignment MSE; M2 = Quality-score mapping minus constant alignment MAE.

5. Consistent finite-cohort gains with remaining calibration errors

The equal-generator mean MSE is 0.430940 for the quality mapping and 1.544910 for the constant baseline. The average difference is negative 1.113969, with standard error 0.569288 and a 95% interval from negative 2.577372 to positive 0.349433. All six observed generators have negative MSE differences. That finite-cohort pattern and the broad working interval are compatible: effects vary considerably across the categories, and only six category summaries support the displayed uncertainty estimate.

The corresponding MAEs are 0.525462 and 1.047404. Their difference averages negative 0.521942, with standard error 0.187671 and an interval from negative 1.004364 to negative 0.039519. Under the stated working inference, this interval remains below zero. It should not be promoted into a stronger claim than the design supports. The folds are dependent, the source categories are not random draws, and a known human quality rating is supplied to every prediction. The result establishes useful information in this finite mapping exercise, with those conditions attached.

The held signed errors show that improvement does not remove generator-specific calibration problems. The mapping overpredicts AttnGAN alignment by an average 0.809947 units and underpredicts DALLE2 by 0.290602. Other signed errors are smaller but also nonzero. These quantities are retained as descriptive diagnostics rather than fitted away using held labels. An institution attempting to replace alignment judgments with these predictions would need to decide whether such biases and individual residuals are tolerable for its actual decision, which the experiment does not specify.

Training slopes vary across folds, from approximately 0.694849 to 0.854954. That variation reflects which source generator is excluded and the resulting training composition; it is not a causal estimate of a generator's effect on semantic quality. The intercepts change as well. A single pooled relationship cannot be assumed to preserve subgroup calibration merely because its overall association is strong. Reporting the fold coefficients and held losses makes this composition sensitivity visible without searching for a post hoc correction that makes every generator look well calibrated.

6. What cannot be transferred from this score experiment

The result is specific to the source rating task. Fang et al., 2020 addresses smartphone photography, Murray et al., 2012 aesthetic rating distributions, and Lin et al., 2019 artificial image distortions. Those resources illustrate why a relationship among one database's aggregate fields need not retain its coefficients or error levels in another. We do not standardize unrelated outcome scales and combine them to create additional transfer tests. Nor do we infer that the known quality score contains the same semantic information for a natural photograph, a controlled distortion, and a generated image paired with text.

The mapping also cannot identify which visual attributes explain the relationship. A high quality MOS could co-occur with alignment because a generator performs well on both, because some prompts are easier, or because shared observer impressions influence both ratings. Other mechanisms are possible. The two-parameter model does not separate them, and generator holdout alone does not randomize the predictor. Better prediction than a constant therefore remains an informational result rather than a causal statement that improving visual quality will improve text correspondence.

The prompt overlap is particularly important for interpreting a future use. A workflow evaluating familiar prompt categories on a new generator may find the current question relevant. A workflow facing entirely new instructions would need additional evidence, because the other-generator training already covers every held exact prompt string here. Even that statement does not imply that all visual or stylistic variations are familiar: different outputs can realize the same text differently. Exact-string overlap defines the available boundary without pretending to resolve semantic or perceptual similarity completely.

Documentation follows Gebru et al., 2021 and Wilkinson et al., 2016 by preserving generator prefixes, prompt-overlap counts, training sizes, fold coefficients, held predictions, and both loss summaries. This record distinguishes the study from an in-sample association analysis and from metadata-only transfer with prompt purging. Reuse of the same public collection still warrants caution about accumulated exploratory decisions. Human review is needed to decide whether this score-transport question has an independent contribution and whether its dependence limits suit a proposed publication or operational claim.

There is also a weighting distinction within the design. Each fold's OLS fit gives training images equal weight, so a larger generator contributes more observations to the fitted relationship. The final comparison instead gives the six held generators equal weight. We do not claim that this combination minimizes an equal-generator risk, and the constant reference follows the same image-weighted training convention. The choice keeps the baseline simple and explicit. An alternative training-weight scheme would need to be specified and evaluated as a new comparison, rather than assumed to reproduce the current fold errors.

7. Conclusion

Human quality MOS improves prediction of alignment MOS for every held generator in the observed collection, yet residual errors and generator-specific biases remain. The MSE and MAE summaries also have different uncertainty under a six-category working analysis. These results support treating quality as informative about alignment while retaining the distinction between the two constructs. They do not establish automatic assessment, unseen-prompt transfer, or a justification for replacing alignment judgments without evaluating the intended downstream decision.

References

1. 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.

2. 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.

3. 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.

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

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

6. 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.

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. 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.

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

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

11. 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.

12. Fleiss, Joseph L.. (1971). Measuring nominal scale agreement among many raters.. Psychological Bulletin. https://doi.org/10.1037/h0031619.

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. 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.

15. 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.

16. Pearson, Karl. (1901). LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. https://doi.org/10.1080/14786440109462720.

17. 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.

18. 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.

19. 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.

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

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

23. 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.

24. 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.

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

26. 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.

27. Tipping, Michael E.; Bishop, Christopher M.. (1999). Probabilistic Principal Component Analysis. Journal of the Royal Statistical Society Series B: Statistical Methodology. https://doi.org/10.1111/1467-9868.00196.

28. 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.

29. 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.

30. 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.