**Journal of Data Efficiency**
Volume 1 · Number 1 · 2026

# More Aggregate Ratings for a Metadata-Only Image-Score Predictor

Gregory Griffin[1] · Ian Lawson[2] · Evan Anderson[3]

[1] Switzerland · [2] Ireland · [3] New Zealand

Original publication: October 24, 2026 | Article 627

## Abstract

Retaining 4,096 rather than 512 training prompt groups reduced squared prediction error for a fixed metadata-only predictor of released image scores. The budgets supplied 4,244 and 526 image-level MOS records respectively. On 3,985 held images, mean squared error fell from 0.215438 to 0.205243, a paired difference of -0.010195 with a prompt-clustered interval of -0.013901 to -0.006489. Absolute error also declined modestly. The target combines perceptual quality and alignment judgments, while the predictors contain generator and configuration fields. This experiment measures the value of retaining aggregate labels within known generators; it does not estimate new annotation cost or validate a visual quality-assessment model.

*Nested prompt-group budgets in AIGIQA-20K, evaluated on identical held prompt groups without using image pixels.*

# More Aggregate Ratings for a Metadata-Only Image-Score Predictor

Nested prompt-group budgets in AIGIQA-20K, evaluated on identical held prompt groups without using image pixels.


## Abstract

Retaining 4,096 rather than 512 training prompt groups reduced squared prediction error for a fixed metadata-only predictor of released image scores. The budgets supplied 4,244 and 526 image-level MOS records respectively. On 3,985 held images, mean squared error fell from 0.215438 to 0.205243, a paired difference of -0.010195 with a prompt-clustered interval of -0.013901 to -0.006489. Absolute error also declined modestly. The target combines perceptual quality and alignment judgments, while the predictors contain generator and configuration fields. This experiment measures the value of retaining aggregate labels within known generators; it does not estimate new annotation cost or validate a visual quality-assessment model.


## 1. A label budget can improve a baseline without measuring visual understanding

A generated-image assessment project can spend effort obtaining more aggregate ratings or developing a richer image representation. Those choices need different evidence. A model that knows the generator and its settings may predict part of a released score even without viewing an image. Additional labels can improve that metadata baseline, but the resulting gain should not be described as evidence that it understands image semantics, detects distortion, or judges a particular generated scene correctly.

We ask whether increasing a nested budget of prompt-grouped MOS records improves a fixed ordinary least-squares predictor on the same held prompt groups. The proposition favors negative larger-minus-smaller prediction-error differences under squared and absolute loss. Both budgets use the same features and contain all source generator categories. This is a study of aggregate label retention for a simple baseline, with no search over architectures or attempt to identify the minimum sufficient annotation budget.

The representation question in Meng, 2026 motivates distinguishing the target score from the information a predictor actually receives. Our executed baseline uses configuration metadata and prompt length, not visual-language alignment features. We do not run the cited assessment method or infer its data requirements. The comparison instead quantifies a more limited source of predictability that a richer assessment experiment may need to recognize when interpreting its own results.

## 2. A composite rating target and a diverse generator table

The source is the public AIGIQA-20K resource associated with Li et al., 2024. The captured table contains twenty thousand image records, fifteen generator families, and 19,259 exact prompt strings. Available columns provide image name, prompt, generator name, iteration setting, guidance setting, width, height, and released MOS. The source study's subjective assessment considers perceptual quality and text-image alignment together. We use its aggregate scalar rather than treating it as a separate distortion-only or alignment-only label.

Observed MOS values range from 0.122290 to 4.981356 in the preserved table. We retain that released numerical scale and do not reconstruct its rater aggregation from missing individual responses. An image-level MOS record is an aggregate label, not one newly collected human response. Our simulated

budgets therefore cannot be converted into respondent counts, working hours, or monetary costs. No new annotation session was conducted for either budget.

The generated-image setting in Li et al., 2023 and broader quality databases represented by Hosu et al., 2019, Fang et al., 2020, and Gu et al., 2020 demonstrate that assessment resources can differ in content, distortion mechanisms, and label construction. Neither those records nor their image pixels enter this analysis. Controlled-distortion material in Ponomarenko et al., 2015 and rating distributions in Murray et al., 2012 likewise answer different questions from this metadata table. We do not pool incompatible MOS scales or claim that the retained-label result transfers automatically between those resources.

## 3. Split exact prompts before choosing a budget

Each exact prompt receives the SHA256 digest of a fixed prefix, `budget20261107:`, followed by its text. Sorting prompts by this digest defines a reproducible ordering independent of MOS values. The first eighty percent of prompt groups form a training pool and the remainder a holdout. The small budget takes the first 512 groups in that pool; the large budget takes its first 4,096. Thus all small-budget prompt groups also occur in the large budget, while no held prompt group appears in either fit.

The budgets yield 526 and 4,244 image records because some prompts occur for multiple images. The held set contains 3,985 images belonging to 3,852 exact prompt groups. Counting groups and images separately matters: retaining one prompt group does not always retain just one MOS. Both fitted samples and the holdout contain all fifteen generator categories. Consequently the evaluation concerns new exact prompts within known generators, not transfer to an unseen generator family.

Grouping protects against exact repeated prompt strings crossing the local boundary. It does not detect paraphrases, nearly identical prompts, shared source concepts, or duplicate visual content. We make no semantic independence claim from a hash rule. Validation concerns in Roberts et al., 2017 motivate recording this limited separation, and model-selection concerns in Varma and Simon, 2006 limit using held errors to redesign the budgets or feature list and then calling the same data an untouched test.

The hash ordering is one fixed pseudo-random arrangement of the available source prompts, not repeated sampling from all possible annotation populations. A different ordering could change which settings and content enter the smaller budget. The uncertainty below conditions on these two realized fits. It does not include repeated budget draws, an independent generator sample, or the uncertainty of recruiting another group of human assessors.

## 4. Twenty regression coefficients and no pixels

Both predictors regress released MOS on generator category, raw guidance setting, log of one plus iteration setting, log image area, log of one plus prompt word count, and log width-to-height ratio. There are twenty fitted coefficients including the intercept and generator indicators. Both design matrices have full column rank. The feature definition and ordinary least-squares method are identical across budgets; only the retained training records change. No regularization parameter or feature subset is selected using held MOS.

Some source settings are fixed or zero-coded for particular generators. Those values remain in the predictive table alongside generator indicators. They are not interpreted as comparable physical doses across all systems. A coefficient in this specification can absorb source coding and generation-family differences without identifying what would happen if an operator changed one setting. The purpose is prediction from the recorded fields, rather than a causal experiment in guidance or iteration count.

Ridge ideas in Hoerl and Kennard, 1970 identify a possible regularized alternative, but no ridge fit is included. Visual assessment approaches represented by Talebi and Milanfar, 2017, Ke et al., 2021, Wang et al., 2022, Yang et al., 2022, and Mittal et al., 2013 use other representations or assumptions. None supplies predictions to our comparison. A metadata-only result cannot inherit the visual or semantic capabilities implied by those methods' inputs. Their relevance is to why a baseline's information boundary must remain explicit when evaluating a more complex assessor.

Predictions remain unbounded ordinary least-squares outputs. We apply no post hoc clipping based on test scores or an observed MOS range. All error calculations use those actual predictions. The held response never enters preprocessing, feature selection, or model fitting. Prompt word count measures length only: the model receives no prompt embedding, token identity, object list, or semantic relation extracted from the text.

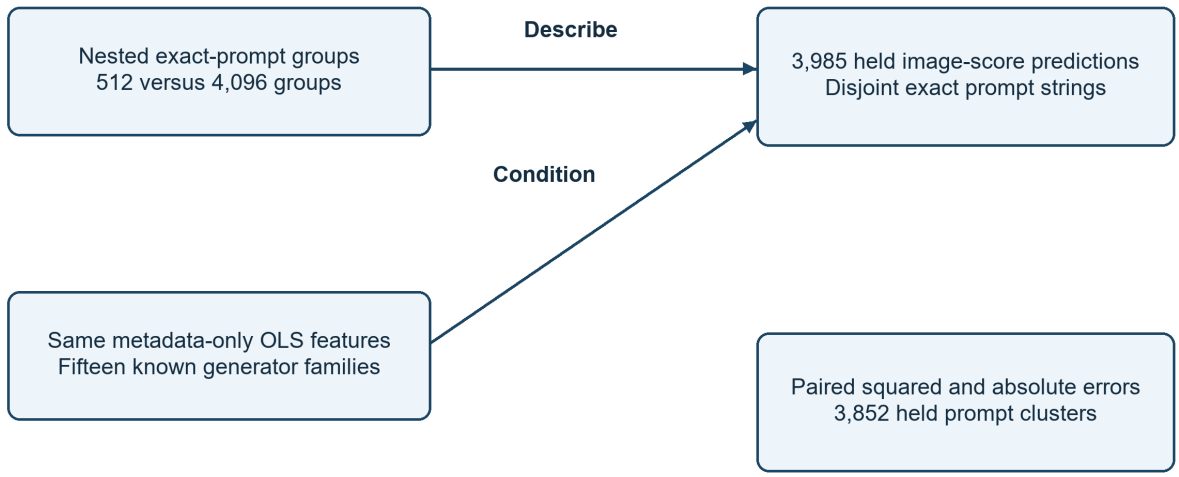



*Figure 1. Nested groups supply two aggregate-rating fits with identical metadata features; predictions are compared on the same disjoint exact-prompt holdout.*

## 5. Prediction endpoints and inferential units

For every held image we calculate squared and absolute differences between predicted and released MOS. We subtract the small-budget error from the large-budget error at that identical image. Separate intercept-only ordinary least-squares models summarize each paired difference. Their units are squared MOS units and MOS units respectively. A negative contrast favors the larger retained-label budget on that endpoint. We do not combine the two loss scales into a single generic quality improvement.

Error-measure considerations in Hyndman and Koehler, 2006 help explain why the absolute and squared outcomes are reported separately. Scoring principles in Gneiting and Raftery, 2007 also distinguish predictive loss from a latent judgment of whether an image is acceptable. Calibration questions in Guo et al., 2017 would require additional evidence about the relationship between scores and desired decisions. We estimate neither a calibrated probability of acceptable generation nor a threshold for selecting images for use.

Covariance clusters by the 3,852 held exact prompts, using a finite-sample correction and a t reference with 3,851 degrees of freedom. The grouping approach follows the dependence considerations in Cameron and Miller, 2015. The heteroskedasticity-consistent framework in White, 1980 alone would not

account for repeated exact prompts. This adjustment does not account for all semantic similarity or shared original raters. The target MOS values are treated as the available labels, not independently remeasured human outcomes.

*Table 1. Larger-minus-smaller retained-rating budget error on 3,985 held images.*

| Model | Term | Estimate (SE) | 95% CI | N |
|---|---|---|---|---|
| M1 | Mean outcome/contrast | -0.0102 (0.0019) | [-0.0139, -0.0065] | 3985 |
| M1: fit | R-squared = 0.0000 | - | - | 3985 |
| M2 | Mean outcome/contrast | -0.0077 (0.0013) | [-0.0102, -0.0052] | 3985 |
| M2: fit | R-squared = 0.0000 | - | - | 3985 |

Model labels: M1 = Larger minus smaller budget squared error; M2 = Larger minus smaller budget absolute error.

These mean-contrast regressions have zero R-squared because they contain no explanatory covariates. Their fit statistic is not the predictive R-squared of the metadata model. We report the latter separately on the shared holdout. Keeping those quantities apart prevents a table about uncertainty in average error from being mistaken for a model of how image content determines quality.

## 6. Modest held improvement from more retained ratings

Mean squared error is 0.215437567 for the small-budget fit and 0.205242736 for the large-budget fit. The paired difference is -0.010194831, with clustered standard error 0.001890187 and interval -0.013900694 to -0.006488968. The reduction is about 4.7 percent of the smaller fit's squared error on this held set. Its denominator is the specified baseline error, not the underlying image quality or the human assessment effort.

Mean absolute error falls from 0.343630793 to 0.335901623. The contrast is -0.007729171, with standard error 0.001266867 and interval -0.010212965 to -0.005245377. Both point estimates and intervals favor retaining more aggregate labels under this realized split. They do not imply that every held image improves, that the larger budget is optimal, or that an additional label has the same value throughout a broader learning curve.

Held predictive R-squared increases from 0.534155 to 0.556200. These values use the same held response variation and support the statement that the larger fit predicts this local target somewhat better. Training R-squared is 0.599188 in the smaller sample and 0.561956 in the larger. Because those training samples differ, that downward change is not evidence that the additional data worsened held prediction. Training and held summaries answer different questions and should not be conflated.

The fifteen generator categories remain visible to the predictor at evaluation. Part of the explained variation can therefore come from generator-associated score differences rather than content-specific judgments. We have not decomposed held R-squared into semantic, distortion, model-family, and configuration components. Calling the total explained fraction evidence of visual understanding would assign the metadata baseline capabilities it was never given an opportunity to demonstrate.

## 7. Aggregate MOS remains a measurement choice

The perceptual motivations in Wang et al., 2004 and Zhang et al., 2018 make clear that multiple numerical endpoints can respond differently to an image. Neither metric is computed in this metadata-only experiment. In particular, the released composite MOS does not provide separate labels with which to test whether the predictor recognizes semantic mismatch independently of visual distortion. A single aggregate regression cannot recover that distinction simply because one of its features describes image dimensions.

Measurement-error reasoning in Carroll et al., 2006 also directs attention to label construction. The source aggregates subjective responses, while the individual rater records are not part of the analyzed table. Shared rater tendencies and image-specific judgment uncertainty may remain. Missing-data principles in Little and Rubin, 2002 further distinguish our deliberately unused training labels from naturally absent or selectively collected assessments. This experiment does not identify the mechanism of a real annotation project's missing ratings.

Influence considerations in Cook, 1977 would motivate inspecting images that contribute strongly to the error difference; none is removed from the result. Multiplicity guidance in Benjamini and Hochberg, 1995 limits broader claims from a search over numerous budgets or error measures. We retain the two declared group counts and two stated endpoints. The interpretive emphasis of Wasserstein and Lazar, 2016 favors reporting their small gain and concrete limits, rather than treating an interval below zero as proof that more annotation is always the preferred investment.

## 8. What a larger data-requirement study would add

A stronger study could repeat the grouped budget sampling and evaluate a planned range of sample sizes. It could hold complete generator families out, collect independent scores for semantic and technical quality, or run defined visual assessors on actual images. Those operations would address different uncertainty and generalization dimensions. They have not been performed by adding more records from the same source to this fixed ordinary least-squares fit.

The current comparison also cannot price the retained labels. The larger budget includes more aggregate MOS records, but the table supplies no estimate of the effort needed to reproduce each aggregate. Whether to collect more ratings rather than richer annotations would depend on actual costs and the intended assessment objective. The observed gain is therefore an empirical baseline quantity, not a recommendation about how a new annotation project should allocate its resources.

## 9. Keep label count, target definition, and representation together

The completed experiment finds a modest local prediction benefit from more retained ratings when the predictor already knows generator identity and configuration. It also exposes why an assessment model's data requirement cannot be summarized by label count alone. Group structure, generator coverage, score composition, and the information provided to the model all determine what such an experiment measures. None can be supplied after the fact by interpreting a favorable held-error contrast more broadly.

Documentation under Gebru et al., 2021 records the exact hash rule, retained group and image counts, feature formula, source coding, and unbounded predictions. Preservation following Wilkinson et al., 2016 retains the held predictions, paired errors, and executable budget construction alongside the source table. These records make the result reproducible. Establishing an independent visual-quality contribution, a general sufficient sample size, or an annotation-cost recommendation requires additional experiments and scientific review beyond this metadata baseline.

## References


1. Talebi, Hossein and Peyman Milanfar. "NIMA: Neural Image Assessment" *arXiv preprint arXiv:1709.05424*, 2017. https://arxiv.org/abs/1709.05424.
2. Mittal, A., et al. "Making a “Completely Blind” Image Quality Analyzer" *IEEE Signal Processing Letters*, 2013. https://doi.org/10.1109/lsp.2012.2227726.
3. 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.
4. 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.
5. 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.
6. 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.
7. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.
8. 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.
9. 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.
10. 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.
11. 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.
12. Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002. https://doi.org/10.1002/9781119013563.
13. 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.
14. 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.
15. 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.
16. 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.
17. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.
18. Ke, Junjie, et al. "MUSIQ: Multi-scale Image Quality Transformer" *arXiv preprint arXiv:2108.05997*, 2021. https://arxiv.org/abs/2108.05997.
19. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.
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. 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.

22. Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. “On Calibration of Modern Neural Networks.” *Proceedings of the 34th International Conference on Machine Learning*, PMLR, vol. 70, 2017, pp. 1321-1330. https://proceedings.mlr.press/v70/guo17a.html.

23. 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.

24. Hyndman, Rob J. and Anne B. Koehler. “Another look at measures of forecast accuracy.” *International Journal of Forecasting*, vol. 22, no. 4, 2006, pp. 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001.

25. 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.

26. 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.

27. 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.

28. 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.

29. 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.

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.