Resource Allocation and Optimization
Volume 1 · Number 1 · 2026

Equal Image Budgets, Different Generator Coverage: An Allocation Experiment with Released Quality Ratings

Alexander Stewart[1] · Aaron Knight[2] · Paul Lewis[3]

[1] Poland · [2] Greece · [3] Croatia

Original publication: October 31, 2026 | Article 641

Abstract

Keeping one hundred rated images from each generator changes representation, but does it improve prediction at a fixed total budget? Two 1,500-image menus from AIGIQA-20K were used to fit the same metadata regression, with exact prompt groups held out. Generator-balanced selection and uniform image selection produced almost identical equal-generator mean squared errors, 0.21765 and 0.21757. Seven generators favored balancing and eight favored the uniform menu. Balanced counts alone therefore supplied no clear predictive gain in this fixed comparison, although they changed which generators received the retained ratings.

Equal Image Budgets, Different Generator Coverage: An Allocation Experiment with Released Quality Ratings

Abstract

Keeping one hundred rated images from each generator changes representation, but does it improve prediction at a fixed total budget? Two 1,500-image menus from AIGIQA-20K were used to fit the same metadata regression, with exact prompt groups held out. Generator-balanced selection and uniform image selection produced almost identical equal-generator mean squared errors, 0.21765 and 0.21757. Seven generators favored balancing and eight favored the uniform menu. Balanced counts alone therefore supplied no clear predictive gain in this fixed comparison, although they changed which generators received the retained ratings.

1. Allocation is a separate decision from budget size

An evaluation team retaining a limited number of rated images must decide both how many to retain and where those images should come from. Those are distinct resource decisions. Increasing a total budget can reduce sampling variability while leaving a generator underrepresented; balancing generator counts can alter representation without adding a single rated image. This article isolates the latter question: at a fixed number of retained source images, does generator-balanced selection improve prediction when each generator is given equal weight during evaluation?

The distinction matters for generated-image quality assessment because source collections often contain unequal numbers of images from different systems. A model trained on such a collection may receive considerably more information about one generator than another. Meng, 2026 addresses quality assessment through a separation of semantic and distortion information. We examine an upstream allocation choice rather than that architecture. The prediction model uses only released configuration metadata, and no target neural model, image pixels, semantic embedding, or new rating exercise enters the experiment.

The outcome is the released composite MOS in AIGIQA-20K, whose database is described by Li et al., 2024. It should not be relabeled as a pure distortion score. The related Li et al., 2023 resource separately highlights quality and alignment assessment, illustrating why an aggregate label's construction affects what a predictor can claim to learn. Our resource question is consequently narrow: how two subsets of already rated images support prediction of the source's composite outcome. Neither subset represents a measured financial annotation cost or a fresh sample of human judgments.

2. Representation, target weights, and a nonautomatic hypothesis

Balanced allocation has an intuitive appeal: every generator receives a minimum amount of information. But equal training counts do not automatically minimize equal-generator prediction loss. Generators may differ in residual variation, predictor support, and how much their coefficients benefit from the common slopes. The sampling perspective in Neyman, 1934 makes allocation an explicit design choice instead of a synonym for representativeness. The proposition tested here is therefore two-sided. Balance might help poorly represented categories, yet it could also remove useful observations from categories with more varied settings.

The role of evaluation weights is equally important. Solon et al., 2015 discusses the interpretation of weighted and unweighted analyses, while Kitagawa, 1955 separates composition from within-group quantities in comparisons of rates. Here the primary evaluation averages mean squared error across fifteen

source generator categories with equal category weight. A secondary descriptive summary weights all held images equally. These targets need not agree because held generator counts are unequal. We do not choose the preferred weighting after inspecting which menu has the lower loss.

This allocation question differs from visual model design. Talebi and Milanfar, 2017 learns image-assessment outputs, Ke et al., 2021 uses a multiscale architecture, and Ying et al., 2019 addresses picture-quality assessment from visual inputs. Our deliberately simple regression makes the retained-data decision easier to inspect, but its simplicity also narrows generalization. A generator-balanced menu could interact differently with a high-capacity visual learner, particularly if visual feature distributions vary across systems in ways that configuration labels and prompt length fail to represent.

Nor are preference and prompt fidelity interchangeable with the present outcome. Wu et al., 2023 and Xu et al., 2023 concern human preference-related evaluation, while Ghosh et al., 2023 and Huang et al., 2023a organize tests of content correspondence and compositional abilities. Those perspectives motivate preserving the source label's meaning. A menu that predicts composite MOS well has not necessarily learned which objects appear in an image, whether a spatial relationship matches a prompt, or which image a particular user would prefer. Equal generator coverage does not resolve those construct differences.

3. A common training pool and genuinely held prompt groups

The starting table contains 20,000 image records from fifteen generator categories and 19,259 distinct exact prompt strings. Prompts are ordered by the SHA256 digest of a fixed prefix, `budget20261107:` followed by the exact prompt. The first eighty percent of distinct groups form the training pool and the rest form the holdout. This produces 16,015 pool images and 3,985 held images representing 3,852 prompt groups. Every source generator occurs in both sets. Exact repeated strings remain together, while semantic near-duplicates are not identified or merged.

This grouping responds to the validation concerns in Roberts et al., 2017: records sharing a prompt are not equivalent to unrelated new inputs. The leakage account in Kapoor and Narayanan, 2023 further motivates keeping held outcomes out of both retention menus and fitting. The held set is frozen before either model is trained. Nothing about a held rating, a held loss, or a favorable generator result is used to change the allocation. It remains possible for semantically similar prompts with different strings to occur across sets, so the split should not be described as a complete semantic-content separation.

For uniform selection, pool images are sorted by a second fixed SHA256 rule using the prefix `allocation20261108:` and the image name, and the first 1,500 are retained. For generator-balanced selection, that same ordering is applied separately within each category, retaining one hundred images from every generator. The resulting menus overlap on 1,165 image records. This overlap is intentional: the comparison changes allocation while holding the source pool and deterministic ordering principle constant. It is not a comparison between independent newly acquired rating samples.

Uniform-menu counts range from 28 for DALLE 3 to 166 for Pixart XL 2. These unequal counts are an observed consequence of the declared image-level selection rule, not a claim that either generator merits that allocation. The balanced menu replaces the count distribution with exactly one hundred per category. Both menus contain 1,500 existing MOS labels. Differences in the expense of obtaining those labels, the number of individual raters behind them, or the practical cost of identifying a generator were not observed and are not inferred from the retained counts.

4. The fixed prediction model and research diagram

Each menu fits the same ordinary least-squares formula. It contains an intercept, fourteen generator-category indicators, raw guidance configuration, log one plus iteration count, log image area, log one plus prompt word count, and log aspect ratio. The design has twenty parameters and full numerical rank in both fits. Source-coded zeros remain as recorded configuration values alongside the category indicators. We do not treat those zeros as common physical doses across all generators or interpret the fitted slopes as causal effects of changing a generation setting.

The modeling choice is intentionally modest. The dimensional-representation ideas of Pearson, 1901 and Tipping and Bishop, 1999 suggest alternatives for reducing richer inputs, but no principal components or latent image representation are fitted here. Hoerl and Kennard, 1970 offers a regularized alternative when correlated predictors threaten stability; we do not tune a ridge penalty in this experiment. Keeping one ordinary regression specification limits the degrees of freedom through which an allocation comparison could become an implicit model-selection exercise. Full rank establishes estimability, not suitability for every generator's response surface.

The error-estimation concern in Varma and Simon, 2006 explains why we do not optimize a menu or model specification against the held ratings. Predictions are generated once from each fitted model and remain unbounded, with no clipping based on holdout extrema. The same held image then has a paired squared-error difference, balanced minus uniform. Those differences are averaged within generator, giving fifteen category-level contrasts. An intercept-only OLS model estimates their equal-generator average. Figure 1 shows how the menu choice enters the analysis through fitting rather than through a changed evaluation population.

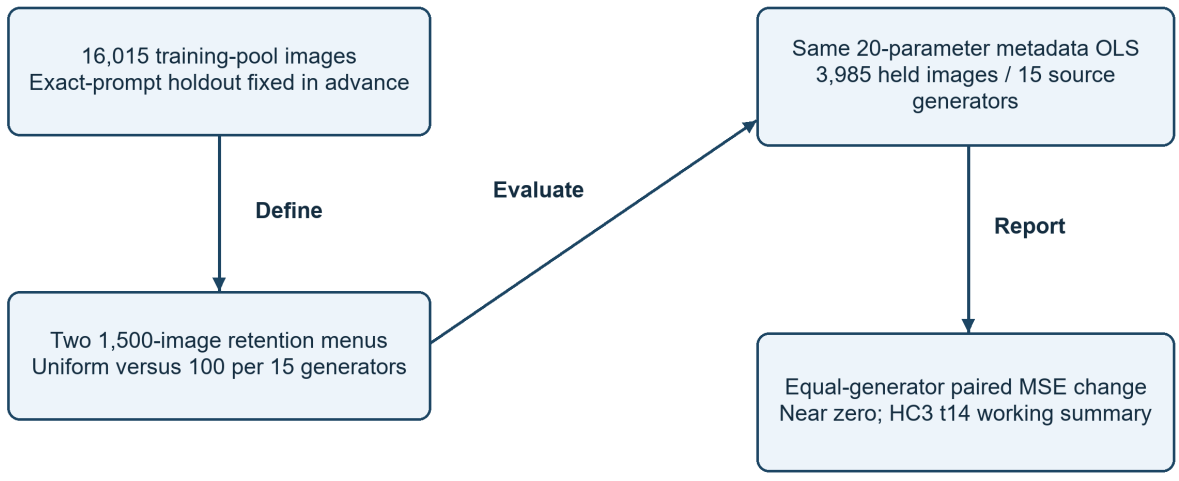


Figure 1. *Equal retained-image budgets produce different generator coverage. The two frozen metadata regressions face the same prompt-grouped holdout, and the primary outcome weights the fifteen held generator categories equally.*

We use HC3 covariance and a t reference with fourteen degrees of freedom for the fifteen generator-level rows. White, 1980 supplies the broader heteroskedasticity-robust perspective, but the small category count warrants restraint. The dependence discussion in Cameron and Miller, 2015 identifies limitations that this working summary does not eliminate: all categories share fitted common slopes, and exact

prompts can cross generator categories. The interval therefore conditions on one pair of fitted menus and does not incorporate retraining variation, menu randomness, or all cross-category dependence.

Table 1. *Balanced minus uniform held mean squared error. Each row in the intercept-only OLS analysis is one generator's paired mean loss difference. N is fifteen source generators, covariance is HC3, and the reference distribution has fourteen degrees of freedom. No controls or fixed effects enter this contrast model; the underlying prediction regressions each use the same twenty parameters.*

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	0.000075 (0.001300)	[-0.002713, 0.002864]	15
M1: fit	R-squared = 0.0000	-	-	15

Model labels: M1 = Balanced minus uniform held MSE, equal generator.

5. A count advantage without a clear loss advantage

Equal-generator mean squared error is 0.217571 for uniform retention and 0.217646 for balanced retention. The estimated difference is positive 0.000075, with a standard error of 0.001300 and a 95% interval from negative 0.002713 to positive 0.002864. The point estimate slightly favors the uniform menu, but the interval includes improvements and deteriorations of considerably greater magnitude than the estimate. The result supplies no clear average prediction advantage for either allocation in this fixed comparison; it does not demonstrate equivalence.

Seven generators have lower loss under the balanced fit and eight have higher loss. This heterogeneity matters because identical category counts do not imply identical predictive demands. The shared model may gain from more observations in one part of the configuration space while losing fit elsewhere. The observed direction count is descriptive and is not a sign test over independent generator draws. All fifteen categories are source-defined and share data-generation and model-fitting structures, so they do not constitute a random sample of every future image-generation system.

The image-weighted summary also slightly favors uniform retention: mean squared error is 0.208642 against 0.209943 for balancing. That summary has a different target from the table and is not substituted for it. Its lower numerical level reflects the distribution of held images across categories, rather than necessarily better prediction for every generator. Reporting both summaries makes the evaluation weights visible and prevents the apparently benign phrase overall error from concealing a change in which source categories receive the most influence.

The loss itself is a choice. Gneiting and Raftery, 2007 provides a general basis for specifying what a predictive score rewards, while the error-measure discussions in Willmott and Matsuura, 2005, Chai and Draxler, 2014, and Hodson, 2022 show why absolute and squared-error criteria should not be casually interchanged. We use squared error throughout the executed primary comparison. We do not add an absolute-error result that was not computed, or infer that the ranking of the menus would be identical under a different loss, a pairwise preference objective, or a content-correctness score.

6. Boundaries of the allocation result

This experiment studies source-category balance rather than population representativeness. The generators are known, their metadata labels are available, and the holdout includes every category. It does not estimate performance on an unseen generator. A generator-balanced menu also uses information that a fully anonymous image pool might not provide. Its one-hundred-per-category constraint is feasible in the observed source but could be infeasible or expensive for a rare generator in practice. No conclusion about the optimal allocation follows from comparing just these two fixed rules.

The source corpus is likewise distinct from other visual-quality resources. Hosu et al., 2019 concerns an ecologically motivated image-quality collection, Fang et al., 2020 addresses smartphone photographs, and Murray et al., 2012 supplies aesthetic vote distributions. Their category meanings and rating protocols differ from the generated-image source used here. A balanced allocation over devices, distortion types, photographers, or semantic categories would be a new experiment. The present findings should not be generalized to such designs simply because each can be described informally as balanced sampling.

The retained records and processing rules are documented in the spirit of Gebru et al., 2021 and the reuse aims of Wilkinson et al., 2016. They include the two image menus, generator counts, exact prompt grouping, fitted formula, per-image held losses, and generator summaries. Those records make it possible to separate this allocation study from a larger-budget experiment on the same database. Reusing a frozen holdout also creates an accumulation risk across successive analyses: the present question was specified before its result, but the collection is not an untouched external validation resource for unlimited future decisions.

The equal-generator target also does not require the training set to have equal counts as a matter of mathematics. Different fitting weights, category-specific models, or explicitly optimized allocations could target that loss in other ways. We did not search those alternatives. The comparison thus evaluates a familiar coverage heuristic against one uniform menu, with a common inexpensive regression. Treating the small observed gap as proof that representation never matters would extend the evidence as far as treating balance itself as a predictive guarantee.

7. Conclusion

Generator-balanced retention substantially changed source representation while leaving the total retained-image budget unchanged. In the executed metadata-only comparison, that change yielded virtually no average equal-generator MSE difference, with uncertainty spanning benefits and harms. The finding separates a legitimate representation objective from a predictive guarantee. Teams choosing a retention policy should state whether they seek coverage, a particular weighted loss, or measured annotation efficiency, because equal category counts alone establish only the first of those outcomes.

References

1. 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.
2. 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.
3. 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.
4. Ke, Junjie, et al. "MUSIQ: Multi-scale Image Quality Transformer" *arXiv preprint arXiv:2108.05997*, 2021. https://arxiv.org/abs/2108.05997.
5. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.

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. 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.

9. 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.

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. Ghosh, Dhruba, et al. "GenEval: An Object-Focused Framework for Evaluating Text-to-Image Alignment." *arXiv preprint arXiv:2310.11513*, 2023. https://arxiv.org/abs/2310.11513.

12. 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.

13. 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.

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

15. 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.

16. Ying, Zhenqiang, et al. "From Patches to Pictures (PaQ-2-PiQ): Mapping the Perceptual Space of Picture Quality." *arXiv preprint arXiv:1912.10088*, 2019. https://arxiv.org/abs/1912.10088.

17. 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.

18. Huang, Kaiyi, et al. "T2I-CompBench++: An Enhanced and Comprehensive Benchmark for Compositional Text-to-image Generation." *arXiv preprint arXiv:2307.06350*, 2023. https://arxiv.org/abs/2307.06350.

19. Solon, Gary; Haider, Steven J.; Wooldridge, Jeffrey M.. (2015). What Are We Weighting For?. Journal of Human Resources. https://doi.org/10.3368/jhr.50.2.301.

20. 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.

21. 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.

22. Xu, Jiazheng, et al. "ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation." *arXiv preprint arXiv:2304.05977*, 2023. https://arxiv.org/abs/2304.05977.

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

24. 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.

25. Neyman, Jerzy. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. Journal of the Royal Statistical Society,97(4),558-606. https://doi.org/10.1111/j.2397-2335.1934.tb04184.x.

26. 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.

27. 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.

28. 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.

29. Kitagawa, Evelyn M.. (1955). Components of a Difference Between Two Rates*. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.1955.10501299.

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.