Operational Evidence and Outcomes
Volume 1 · Number 1 · 2026

What a Balanced Quality-Alignment Gate Gives Up: Equal-Capacity Selection of Generated Images

Oliver Turner[1] · Spencer Douglas[2] · Jason Reed[3]

[1] Luxembourg · [2] Canada · [3] Germany

Original publication: November 8, 2026 | Article 634

Abstract

Selecting images by quality alone and selecting them by a balanced quality-alignment score serve different objectives. On 2,982 AGIQA-3K records, two within-generator rules each retain 596 images. Replacing quality ranking with the minimum of the two percentile ranks raises mean alignment by 0.300 native units and lowers mean quality by 0.122 across six generators. Only 381 images appear in both selections. The quality reduction follows from departing from a quality-maximizing list; the observed alignment gain quantifies what that departure buys in this finite, fully rated collection.

What a Balanced Quality-Alignment Gate Gives Up: Equal-Capacity Selection of Generated Images

Abstract

Selecting images by quality alone and selecting them by a balanced quality-alignment score serve different objectives. On 2,982 AGIQA-3K records, two within-generator rules each retain 596 images. Replacing quality ranking with the minimum of the two percentile ranks raises mean alignment by 0.300 native units and lowers mean quality by 0.122 across six generators. Only 381 images appear in both selections. The quality reduction follows from departing from a quality-maximizing list; the observed alignment gain quantifies what that departure buys in this finite, fully rated collection.

1. A retention rule encodes a priority

A generated image can be visually attractive while departing from the text that requested it. A workflow retaining only a limited number of candidates must decide how to handle that conflict. Ranking by a quality score gives one clear answer: use the available places for the largest values of that score. A balanced rule gives another: avoid an image whose weaker dimension is too low relative to its peers. This article asks what changes in the selected images when those priorities are compared at exactly the same capacity.

The separation of semantic information and distortions in Meng, 2026 motivates keeping quality and alignment visible as distinct evaluation dimensions. We do not implement that architecture or claim to recover its latent streams. Our data already contain two released human aggregate ratings per image. They permit a full-information retrospective selection exercise: the two policies use the actual quality and alignment summaries, and their selected outcomes can be calculated exactly. This differs from learning an automatic gate that must infer those scores from an image.

The source is the generated-image database in Li et al., 2023, with 2,982 records, six generator prefixes, and 300 exact prompt strings. The present question concerns a decision imposed on those existing records, rather than whether the two ratings are correlated. Correlation does not specify a retention policy, a capacity, or a tradeoff between the dimensions. By making the capacity and score construction explicit, we can identify which numerical sacrifice and gain accompany a particular balancing rule without presenting it as the uniquely correct preference for every user.

2. Different evaluation traditions imply different gates

Human preference-related evaluation in Wu et al., 2023 and Xu et al., 2023 illustrates that an image's desirability can depend on what people value in the task. Content-oriented frameworks such as Ghosh et al., 2023, Huang et al., 2023a, and Hu et al., 2023 instead inspect aspects of correspondence to a requested description. Those objectives can overlap with perceived quality without being identical to it. We do not use the present quality rating as an independent truth label for alignment, nor infer that a higher alignment average would necessarily improve an unmeasured pairwise human preference.

Representation methods also do not choose the institutional objective. Radford et al., 2021 establishes a broad image-text representation setting, and Wang et al., 2022 explores image-quality assessment using such representations. Talebi and Milanfar, 2017 and Ke et al., 2021 address assessment from images through different modeling choices. A workflow could use these or other predictors to estimate scores, but the final retention rule would still require a choice about their relative importance. Our comparison

isolates that choice by supplying both released scores directly, avoiding an extra layer of predictor error that was not measured in this study.

One simple response is a weighted sum, but its meaning depends on scale and weights. Following the measurement distinction in Stevens, 1946, we instead convert each score to its within-generator percentile rank and take the smaller percentile as the joint score. This avoids adding raw units from different rating fields. It introduces a different assumption: performance is judged by the weaker relative position, and relative positions within a generator are treated as commensurable. That assumption is transparent but is not validated by the fact that it produces a balanced-looking shortlist.

The decision literature represented by Elkan, 2001 makes costs and preferences central to a classification policy. The selective-classification framework in Geifman and El-Yaniv, 2017 also treats coverage as part of a decision. We have no elicited utility function, review cost, or cost of a semantic failure. Equal retained capacity is therefore a controlled comparison device rather than an economically optimal review allocation. The policy may be appropriate for a user who dislikes weak performance on either dimension, but the experiment cannot establish that such a preference is universal.

3. Two rules, identical source records and capacity

Within each source generator, we retain the floor of twenty percent of its image count. The capacity is calculated before either ranking is applied and is the same under both policies. Summed over the six generators, each retains 596 images. Generator-wise allocation prevents the pooled list from changing its generator composition merely because one system has a higher score distribution. It does not correct unequal prompt composition, different rating precision, or every other source difference between generator categories.

The quality-only rule sorts by quality MOS descending and then filename ascending. The joint rule calculates average-tie percentile ranks separately for quality and alignment within the generator, assigns each image the smaller of those ranks, and sorts by that joint score descending. Remaining ties are resolved by quality descending and then filename ascending. These tie rules are explicit because a percentile-based joint score can create tied values even when the underlying two raw scores differ. No tie is broken using the eventual group-level loss or a preferred example.

Both policies receive the complete released ratings of every candidate image. There is no training split, unseen-image holdout, or estimated missing value. This makes the numerical decision comparison exact for the finite collection, but also makes it more informed than an automatic production gate. If a real gate had to estimate alignment, score errors could alter membership and the resulting tradeoff. We do not borrow the current score availability as evidence that such a gate would be inexpensive or accurate.

The quality-only list maximizes the sum, and therefore the mean, of quality among all lists of the same size within the generator. Departing from it cannot increase that mean unless ties yield an equal result. A quality decrease under the joint rule is consequently expected by construction, and its sign is not an empirical discovery about a failure of balancing. The alignment change is different: it is measured from the selected records and need not be positive for every conceivable source distribution or alternative joint rule.

4. Measuring the tradeoff without combining its units

For each generator, we calculate the selected mean quality and mean alignment under both rules, their paired differences, and the number of shared selected filenames. The primary summaries average differences equally over the six generator categories. That weighting is deliberate and differs from pooling all selected images when capacities vary slightly. It answers what the mean generator-level change is among these source categories. No claim is made that the categories constitute a random sample of every image-generation system that will exist in deployment.

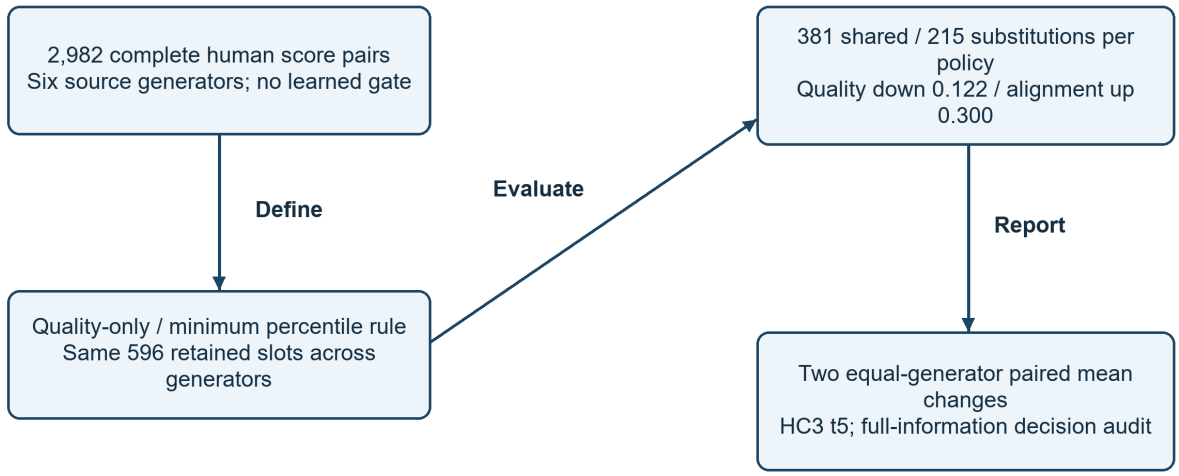

Figure 1. *Complete released ratings feed two equal-capacity rules. Their different priorities change membership, after which quality and alignment consequences are evaluated separately in the source rating units.*

Two intercept-only OLS models summarize the quality and alignment changes, respectively. Covariance is HC3, with a Student reference using five degrees of freedom. White, 1980 supplies the general heteroskedasticity-aware perspective. Cameron and Miller, 2015 makes the dependence limitation clear: treating six category summaries as working rows does not remove shared prompts, common observer structure, or related generators. These intervals express a small-cohort descriptive summary and should not be presented as a population guarantee for a deployment policy chosen from a large menu of candidate rules.

The paired-measurement perspectives in Bland and Altman, 1986 and Lin, 1989 also discourage collapsing the two outcomes into a single association statistic. A quality loss and an alignment gain are measured in their own source units, so subtracting one from the other would require an explicit exchange rate. The agreement framework in Fleiss, 1971 concerns another object again, individual-rater agreement, which cannot be recovered from these two aggregate scores. We leave the tradeoff as two reported quantities rather than inventing an overall benefit index.

Table 1. *Joint-rule minus quality-only selected means, estimated separately by intercept-only OLS on six generator summaries. N is six, covariance is HC3, and the reference distribution has five degrees of freedom. There are no controls or fixed effects. Quality and alignment coefficients remain in their respective native rating units.*

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	-0.122347 (0.009728)	[-0.147354, -0.097340]	6
M1: fit	R-squared = 0.0000	-	-	6
M2	Mean outcome/contrast	0.300358 (0.042147)	[0.192016, 0.408699]	6
M2: fit	R-squared = 0.0000	-	-	6

Model labels: M1 = Joint minus quality-only mean quality among selected images; M2 = Joint minus quality-only mean alignment among selected images.

5. What the joint rule changes in this collection

The equal-generator average selected quality is 3.002498 under quality-only ranking and 2.880152 under the joint rule. The paired change is negative 0.122347, with standard error 0.009728 and a 95% interval from negative 0.147354 to negative 0.097340. All six generators experience a quality reduction. The magnitude describes how far this particular joint policy moves away from the quality-maximizing reference; the common negative sign follows from the fixed-capacity comparison rather than establishing a broadly harmful effect of multiobjective selection.

Selected alignment rises from 2.607107 to 2.907464 in the equal-generator average. The change is positive 0.300358, with standard error 0.042147 and an interval from 0.192016 to 0.408699. Every generator has a positive alignment change in this source collection. That pattern is consistent with the joint score penalizing weak alignment ranks, but it is not a formal optimality result. A different loss, capacity, or source composition could yield another balance of quality and alignment consequences.

The lists share 381 of their 596 selected images. Each therefore contains 215 images absent from the other, with the overall capacity unchanged. This reveals a meaningful membership change that would be obscured by describing the two original score fields as highly associated. It is also different from comparing a quality-only list with an alignment-only list: the joint rule uses both fields and can exclude an image that scores highly on one but relatively poorly on the other. The saved membership record allows those actual substitutions to be inspected.

The threshold perspective in Fawcett, 2006 helps distinguish score ranking from a specific operating decision. Our twenty-percent capacity is an illustrative retention constraint, not a validated acceptability threshold. Gneiting and Raftery, 2007 likewise emphasizes that evaluation rewards a specified objective. We do not claim that either rule is a proper probabilistic score, that both dimensions have equal social importance, or that the percentile minimum represents a measured user utility. The selected score means describe the consequences of an openly chosen priority, leaving the priority itself open to substantive judgment.

6. What the tradeoff does not certify

Perceptual comparison and generative-distribution evaluation concern other levels of evidence. Ding et al., 2020 and Zhang et al., 2018 compare visual representations, while the evaluations associated with Heusel et al., 2017 and Bińkowski et al., 2018 operate on distributions rather than the current per-image retention objective. Their values could not be inserted into this table as interchangeable quality or alignment units. The present result also says nothing about diversity among selected images, representation of sensitive subjects, or the coverage of a particular compositional skill. Those outcomes were not measured merely by retaining a balanced pair of score ranks.

The benchmark-comparison perspective in Demšar, 2006 cautions against treating many correlated image records as independent replications of an overall method advantage. We report the finite membership counts directly and use only six generator-level rows for the displayed working intervals. The sampling distinction in Neyman, 1934 further clarifies that selecting the highest-scored fifth is a decision sample, not a representative probability sample of images. The output list is intentionally enriched for the chosen objectives, so its average cannot stand in for the quality of the full generator output distribution.

The reporting principles associated with Wasserstein and Lazar, 2016 support making the mechanical part of the result explicit. A small p-value for quality loss would add little to the fact that quality-only ranking maximizes that same quality score at fixed capacity. The informative questions are its observed size, the alignment gain, and the actual membership substitutions. Those quantities remain conditional on complete human ratings and one policy. We do not search capacities or joint scoring rules until a favorable tradeoff appears and then label that choice confirmatory.

Documentation follows Gebru et al., 2021 and Wilkinson et al., 2016 by retaining source filenames, generator labels, score ranks, capacities, selection flags, and paired generator outcomes. This makes it possible to compare a different policy openly rather than silently changing how the existing results were produced. The experiment is a decision-consequence audit on already available aggregate judgments. Deployment would require separate evidence about score acquisition, prediction error, utility, costs, and whether the retained images serve the task that motivated the gate in the first place.

Percentile balancing also discards information about cardinal distances. Two images separated by a small raw alignment difference can occupy noticeably different relative positions in a dense score distribution, while a larger raw gap may cross fewer ranks in a sparse region. This is a consequence of choosing relative standing as the input to the joint gate. It is neither a defect discovered by the experiment nor proof that raw-score weighting would be preferable. A workflow should choose that representation because it matches an articulated decision purpose, rather than because percentiles look automatically comparable.

Capacity is another substantive parameter. A very permissive gate could retain nearly all images and make the two lists converge, whereas a strict capacity could emphasize a few extreme score combinations. We report one declared fifth-of-source capacity instead of presenting an unexecuted curve or suggesting that the observed tradeoff is constant across budgets. The stored per-image ranks permit future capacity comparisons, but their interpretation would require acknowledging the additional policy search and any resulting choice made after looking at the outcomes.

7. Conclusion

At equal capacity, the balanced percentile rule substitutes 215 images and trades a lower selected quality mean for a higher selected alignment mean in the released source. That tradeoff is measurable without declaring either dimension the single correct objective. The quality-loss sign is built into departing from a quality-maximizing list; the alignment gain supplies the empirical counterpart. A defensible retention policy should expose both consequences and the preference that makes them acceptable.

References

1. 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.
2. Geifman, Yonatan, and Ran El-Yaniv. "Selective Classification for Deep Neural Networks." *Advances in Neural Information Processing Systems*, vol. 30, 2017. https://arxiv.org/abs/1705.08500.
3. Lin, Lawrence I-Kuei. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics. https://doi.org/10.2307/2532051.
4. 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.
5. 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.
6. Talebi, Hossein and Peyman Milanfar. "NIMA: Neural Image Assessment" *arXiv preprint arXiv:1709.05424*, 2017. https://arxiv.org/abs/1709.05424.
7. Bińkowski, Mikołaj, et al. "Demystifying MMD GANs." *arXiv preprint arXiv:1801.01401*, 2018. https://arxiv.org/abs/1801.01401.
8. 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.
9. Heusel, Martin, et al. "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium" *arXiv preprint arXiv:1706.08500*, 2017. https://arxiv.org/abs/1706.08500.
10. Fawcett, Tom. (2006). An introduction to ROC analysis. Pattern Recognition Letters. https://doi.org/10.1016/j.patrec.2005.10.010.
11. Ding, Keyan, et al. "Image Quality Assessment: Unifying Structure and Texture Similarity." *arXiv preprint arXiv:2004.07728*, 2020. https://arxiv.org/abs/2004.07728.
12. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.
13. 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.
14. Hu, Yushi, et al. "TIFA: Accurate and Interpretable Text-to-Image Faithfulness Evaluation with Question Answering." *arXiv preprint arXiv:2303.11897*, 2023. https://arxiv.org/abs/2303.11897.
15. 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.
16. 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.
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. Demšar, Janez. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research. https://www.jmlr.org/papers/v7/demsar06a.html.

19. 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.

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

21. 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.

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. 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.

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

25. 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.

26. 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.

27. 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.

28. Radford, Alec, et al. "Learning Transferable Visual Models From Natural Language Supervision" *arXiv preprint arXiv:2103.00020*, 2021. https://arxiv.org/abs/2103.00020.

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

30. Elkan, Charles. “The Foundations of Cost-Sensitive Learning.” *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, 2001, pp. 973-978. https://cseweb.ucsd.edu/~elkan/rescale.pdf.