Journal of Measurement and Evidence
Volume 1 · Number 1 · 2026

A High Correlation Can Conceal Different Shortlists: Quality and Text Alignment in AGIQA-3K

Brandon Wells[1] · Lucas King[2] · Ryan King[3]

[1] Malta · [2] Canada · [3] Germany

Original publication: October 2, 2026 | Article 584

Abstract

Would choosing generated images by visual quality produce the same shortlist as choosing them by text alignment? Across 2,982 AGIQA-3K records, the standardized score association was 0.814 before adjustment and 0.505 after generator and exact-prompt fixed effects. Yet equal-capacity top-quarter lists within each generator shared only 367 of 746 selected images. Aggregate correlation therefore offered limited assurance that the two rating fields would support the same selection. The analysis uses released human-score summaries, without evaluating an automated image-quality model.

Exploratory research draft; interpretation, source verification by an accountable human, and publication suitability remain under review.

A High Correlation Can Conceal Different Shortlists: Quality and Text Alignment in AGIQA-3K

> Exploratory research draft; interpretation, source verification by an accountable human, and publication suitability remain under review.

Abstract

Would choosing generated images by visual quality produce the same shortlist as choosing them by text alignment? Across 2,982 AGIQA-3K records, the standardized score association was 0.814 before adjustment and 0.505 after generator and exact-prompt fixed effects. Yet equal-capacity top-quarter lists within each generator shared only 367 of 746 selected images. Aggregate correlation therefore offered limited assurance that the two rating fields would support the same selection. The analysis uses released human-score summaries, without evaluating an automated image-quality model.

1. A substitution question behind image evaluation

An image can look polished while failing to depict an important part of its prompt. Conversely, a recognizable depiction of the requested content can contain visible artifacts. Evaluation systems that report separate quality and alignment scores acknowledge this distinction, but users may still simplify a comparison by retaining only one score. A high correlation between the two fields can make that simplification appear harmless. The relevant empirical question is whether the apparent agreement persists after composition is considered and whether it supports the same image-level choices.

We ask whether released aggregate quality MOS can substitute for text-alignment MOS in a fixed public collection. The question concerns the relationship between two observed rating fields, not whether quality causes alignment or whether either field completely measures its intended construct. Li et al., 2023 provides the paired scores and generation metadata for the analysis. The collection contains several generator families and repeated exact prompt strings, allowing a distinction between pooled association and association remaining after these two sources of composition are represented explicitly.

The motivating architectural literature also separates these concerns. Meng, 2026 proposes multi-scale two-stream vision-language alignment for generated-image quality assessment, explicitly framing semantics and distortions as separable considerations. That framing motivates our measurement question. We do not implement the proposed model, reproduce its experiments, or attribute any observed score relationship to its architecture. The empirical contribution is a transparent audit of whether two released outcomes support substitution under a stated numerical model and a stated shortlist rule.

Our analytical proposition has two parts. Generator and prompt composition may explain some of the pooled relationship, so adjustment may change its magnitude. Even a positive adjusted relationship need not imply equivalent rankings at a fixed selection capacity. The distinction between measurement scales emphasized by Stevens, 1946 and the difference between correlation and agreement examined by Bland and Altman, 1986 make these separate questions necessary. They also discourage calling the two scores interchangeable merely because one is statistically informative about the other.

2. Related evaluation objects

Image evaluation already uses several kinds of evidence. Structural similarity Wang et al., 2004 compares image structure, learned perceptual similarity Zhang et al., 2018 uses deep representations, and structure-texture similarity Ding et al., 2020 emphasizes another perceptual construction. These are methodologically relevant alternatives, but no score from those methods appears in our regression. A relationship between released human quality and alignment means cannot establish that any one perceptual metric is a sufficient replacement for either. It describes the recorded human summaries at their own observation level.

Text-image representation models add a different route to evaluation. CLIP Radford et al., 2021 and the OpenCLIP training work Cherti et al., 2022 motivate using shared visual-language representations, while CLIP-based image-quality assessment Wang et al., 2022 applies such representations to an evaluative task. A shared embedding space does not by itself specify which quality dimensions should be aggregated or how much a missing prompt attribute should matter. Our analysis has no downloaded pixels or embeddings, so it cannot compare these systems or inspect which visual features account for a particular score disagreement.

Several benchmarks make the alignment object more explicit. TIFA uses question answering to assess faithfulness Hu et al., 2023, GenEval evaluates object-focused alignment Ghosh et al., 2023, and compositional benchmarking expands the combinations of requested attributes and relations Huang et al., 2023a. These approaches help identify why a generic high-quality label may omit important semantic requirements. The current alignment MOS remains the source's aggregate rating field. It is not reconstructed from object counts, questions, or compositional annotations, and we do not assert equivalence among those evaluation procedures.

Human preference provides another nearby but distinct object. The Human Preference Score work Wu et al., 2023, Pick-a-Pic Kirstain et al., 2023, and ImageReward Xu et al., 2023 concern preference-oriented data or evaluation. A preference can combine visual appeal, instruction adherence, novelty, and other considerations in ways that two separate MOS fields do not identify. Therefore, the shortlist exercise below should be read as a comparison of two recorded ordering rules. It is not an inferred majority preference and does not establish a deployment policy on behalf of the original observers.

3. Released observations and the model diagram

The complete metadata file has 2,982 unique image names, six generator prefixes, and 300 distinct exact prompt strings. Generator counts are uneven: 1,200 records have the sd1.5 prefix, 592 midjourney, 300 each AttnGAN, glide, and xl2.2, and 290 DALLE2. Prompt multiplicities range from seven to ten records. Missing adjective and style fields are not used. Both MOS fields are observed for every included record, so the analysis retains the full file without imputing either outcome.

Generator identity is read from the prefix before the first underscore in the documented image name. Prompt identity means exact equality of the released string. It does not merge paraphrases, resolve semantic equivalence, or establish that all records with the same string share every generation setting. Each MOS field is standardized once using its complete-file mean and population standard deviation. Quality has mean 2.5004172 and standard deviation 0.9977410; alignment has mean approximately 2.5000000 and standard deviation 0.9988256. That fixed scaling is maintained across all specifications.

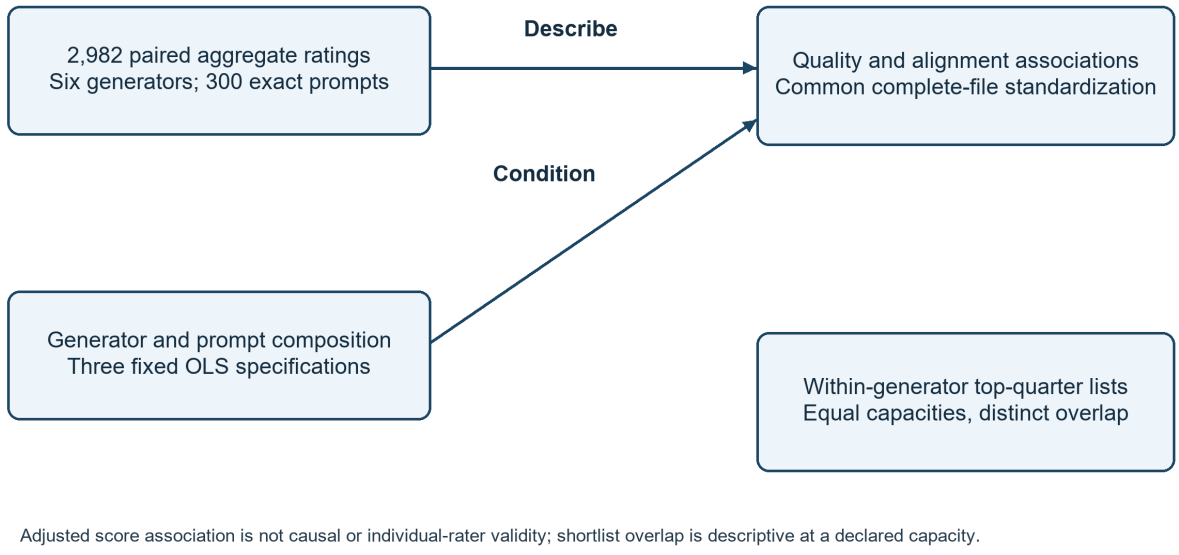

Figure 1. Composition-adjusted score association and equal-capacity selection overlap.

Figure 1 contains two linked descriptive exercises. The regression branch relates standardized quality to standardized alignment, adding generator and exact-prompt indicators in stages. The selection branch ranks images separately by each raw MOS within each generator, applying the same capacity to both rankings. These branches share observations but answer different questions. Neither contains an arrow asserting that an intervention on visual quality would change alignment. The causal-graph distinctions discussed by Greenland et al., 1999 help prevent a statistical adjustment diagram from acquiring a causal interpretation that this design cannot support.

The observation is an image-level aggregate score pair, not an individual judgment. Measurement-error analysis Carroll et al., 2006 reminds us that aggregation does not remove uncertainty about the measurement process. The available per-image standard-deviation fields do not provide observer identities, cross-dimension response covariance, or the raw paired judgments needed to reconstruct individual-level agreement. Consequently, we do not estimate inter-rater reliability, factor loadings, or the probability that an individual observer would prefer one image over another.

4. Three regressions and two ordering rules

Model M1 regresses standardized alignment on standardized quality with an intercept. Because both variables have the same complete-file standardization, its slope equals the pooled Pearson correlation. M2 adds five generator indicators. M3 additionally includes 299 exact-prompt indicators. All three use the same 2,982 records. The script checks that each numerical design matrix has full column rank, so an omitted reference category handles the categorical intercept relationships rather than an unidentified coefficient set.

Covariance estimates cluster on the 300 exact prompt strings, with a finite-sample correction and t reference using 299 degrees of freedom. This choice recognizes repeated content requests within the source file. Cluster-robust inference follows the methodological discussion of Cameron and Miller, 2015, building on robust covariance reasoning White, 1980. It does not account for unknown observer dependence or make the six generators a random sample of all generation systems. The intervals remain conditional on the observed metadata definitions and the specified linear association model.

The three slopes are not directly interchangeable with three correlations. Only M1 has the simple pooled-correlation interpretation. Once indicators are added, the slope relates residual variation in the two globally standardized scores after the represented composition is accounted for. A higher R-squared in M3

can partly reflect the many prompt indicators; it is not held-out predictive validation. No regression coefficient is used to generate the shortlists, which depend directly on each recorded raw MOS field.

Within a generator with N images, each shortlist contains the ceiling of one quarter of N. Images are ordered by descending score, with the unique filename breaking score ties. The quality and alignment capacities are therefore exactly equal within every generator. We record the shared images, the images selected only by each rule, and the Jaccard intersection-to-union ratio. Combining counts across generators weights them by their selected capacities. It does not give the six generator families equal weight or estimate an unobserved future selection frequency.

5. Association weakens after composition adjustment

The pooled standardized slope is 0.8141071, with a 95% prompt-clustered interval from 0.7916869 to 0.8365274. Generator adjustment reduces the estimate to 0.5369860, with interval 0.5003168 to 0.5736552. Including exact-prompt indicators yields 0.5050635, with interval 0.4714153 to 0.5387118. The association remains positive under all three specifications. Its reduction shows that the pooled value includes a substantial compositional component under the declared indicator model, without identifying the generator-level mechanisms responsible.

Table 1. Quality MOS associations with standardized alignment MOS

Model	Term	Estimate (SE)	95% CI	N
M1	Quality MOS, per global SD	0.8141 (0.0114)	[0.7917, 0.8365]	2982
M1: fit	R-squared = 0.6628	-	-	2982
M2	Quality MOS, per global SD	0.5370 (0.0186)	[0.5003, 0.5737]	2982
M2: fit	R-squared = 0.7400	-	-	2982
M3	Quality MOS, per global SD	0.5051 (0.0171)	[0.4714, 0.5387]	2982
M3: fit	R-squared = 0.8257	-	-	2982

Model labels: M1 = Pooled standardized score association; M2 = Generator-adjusted standardized score association; M3 = Generator and exact-prompt adjusted association.

Table 1 presents OLS coefficients and standard errors clustered on 300 exact prompts, using t299 intervals and finite-sample correction. M1 has an intercept and quality only; M2 adds generator fixed effects; M3 adds generator and prompt fixed effects. All columns use 2,982 images and the same complete-file standardization. R-squared values are 0.6628, 0.7400, and 0.8257. Categorical coefficients are included in estimation but omitted from the displayed table, which focuses on the common quality slope.

The association is not close to a deterministic relationship even when its direction is precise. An estimated conditional slope of about one half says that residual variation in quality accompanies residual variation in alignment under the fitted model. It does not specify how often the highest-scoring images coincide. Nor does it validate a universal conversion between the two scales. The interpretation of scoring rules discussed by Gneiting and Raftery, 2007 depends on defining the predictive or evaluative object; changing that object changes what an apparently successful score relationship can justify.

6. Equal capacity still produces different selections

Across the six generators, each ordering rule selects 746 images. Only 367 appear in both selections, leaving 379 images unique to the quality list and 379 unique to the alignment list. The combined union contains 1,125 images, yielding Jaccard overlap 0.3262222. Thus slightly fewer than half of the images selected by either individual rule are shared with the other rule. This is a property of the fixed recorded lists, not a confidence interval for future annotation or a randomized decision experiment.

The mismatch is present in every generator family. AttnGAN has 30 shared images among two lists of 75, DALLE2 has 33 among lists of 73, and glide has 34 among lists of 75. Midjourney has 79 shared among lists of 148, sd1.5 has 159 among lists of 300, and xl2.2 has 32 among lists of 75. Their Jaccard values range from 0.2500 to 0.3641. We do not interpret these differences as rankings of generator quality, since they describe agreement between dimensions within a generator's own records.

This selection result supports a concrete answer to the substitution question. Replacing alignment-based ordering with quality-based ordering would change many selected images at the declared capacity, despite the positive association. It does not say which replacements are desirable. That judgment requires a stated task objective and loss for selecting a visually strong but weakly aligned image versus the reverse. A selection rule can be reproducible without being substantively appropriate for a particular publisher, user, or safety-sensitive application.

7. Scope, documentation, and implications

A reader might ask whether the reduced adjusted slope proves that generator identity is the principal source of agreement. The sequential models do not support that attribution. Generator indicators absorb all average differences associated with each prefix, including observed and unobserved settings. Prompt indicators absorb mean differences among exact strings. Their contribution depends on the order and specification of adjustment, and the remaining slope can still reflect jointly changing image properties. The useful finding is the sensitivity of the pooled relationship to a clearly described composition adjustment, rather than a decomposition into independently identified psychological mechanisms.

The two numerical exercises also differ in their sensitivity to transformations. A positive affine rescaling of either MOS leaves its within-generator ordering unchanged, so the declared shortlist overlap would be identical. The standardized regression likewise removes a common shift and scale, but it still assumes a linear relationship after its indicators. A nonlinear monotonic transformation could preserve the top lists while changing that fitted slope. This distinction explains why reporting both exercises adds information: one evaluates a linear association under fixed scaling, while the other evaluates a concrete ordering consequence without requiring a linear conversion between the two fields.

Alternative quality models, including NIMA Talebi and Milanfar, 2017, MUSIQ Ke et al., 2021, and multi-dimension attention-based assessment Yang et al., 2022, could be evaluated against the two fields in a separate study. Our results do not choose among them. Likewise, KonIQ-10k Hosu et al., 2019, PIPAL Gu et al., 2020, and KADID-10k Lin et al., 2019 have their own image constructions and evaluation aims. They are relevant reminders that a score name can span different evidence settings; they are not additional samples in the regressions reported here.

The strongest limitations concern interpretation rather than numerical precision. Released aggregate scores cannot establish that quality and alignment are psychologically independent latent traits. Fixed effects do not remove all omitted settings or recover individual raters' preferences. Exact prompt matching leaves semantic near-duplicates unresolved. The quarter-capacity rule is deliberately declared, not

estimated as an optimal threshold. Small p-values for association cannot settle these issues, consistent with the contextual interpretation urged by Wasserstein and Lazar, 2016.

The source file, its checksum, the feature construction, shortlist membership counts, and regression outputs are retained locally. Documentation distinguishes image-level summaries from unavailable observer records, as encouraged by Gebru et al., 2021, while stable source identifiers and structured outputs support the reuse principles in Wilkinson et al., 2016. The numerical answer is modest but operationally relevant: quality and alignment move together in this collection, yet they do not nominate the same top-quarter images. Evaluation reports should preserve that distinction when a later decision depends on which dimension was actually measured.

References

1. Talebi, Hossein and Peyman Milanfar. "NIMA: Neural Image Assessment" *arXiv preprint arXiv:1709.05424*, 2017. https://arxiv.org/abs/1709.05424.
2. 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.
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. 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.
5. 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.
6. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.
7. 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.
8. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.
9. Cherti, Mehdi, et al. "Reproducible scaling laws for contrastive language-image learning." *arXiv preprint arXiv:2212.07143*, 2022. https://arxiv.org/abs/2212.07143.
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. 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.
12. Kirstain, Yuval, et al. "Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation" *arXiv preprint arXiv:2305.01569*, 2023. https://arxiv.org/abs/2305.01569.
13. 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.
14. 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.
15. Radford, Alec, et al. "Learning Transferable Visual Models From Natural Language Supervision" *arXiv preprint arXiv:2103.00020*, 2021. https://arxiv.org/abs/2103.00020.
16. Greenland, Sander, Judea Pearl, and James M. Robins. “Causal Diagrams for Epidemiologic Research.” *Epidemiology*, vol. 10, no. 1, 1999, pp. 37-48. https://doi.org/10.1097/00001648-199901000-00008.

17. 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.

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

19. 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.

20. 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.

21. 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.

22. 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.

23. 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.

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

25. 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.

26. 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.

27. 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.

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

29. 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.

30. 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.