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

When Generation Settings Change the Gap between Image Quality and Prompt Alignment

Blake Sanders[1] · Brandon Walker[2] · Ryan King[3]

[1] Netherlands · [2] Belgium · [3] Austria

Original publication: October 2, 2026 | Article 578

Abstract

A setting can improve how an image looks without improving how closely it follows its prompt. We paired the two released rating dimensions for 287 prompts observed under ten AGIQA-3K settings. Relative to the standard Stable Diffusion 1.5 setting, the alignment-minus-quality gap was 0.441 units larger for low-step Midjourney and 0.306 units smaller for standard Midjourney. Prompt-clustered intervals excluded zero for both contrasts. Rankings of setting means agreed only partly across dimensions, with Spearman correlation 0.770. These are differences on the released scales, rather than evidence that the underlying perceptual constructs share a natural unit. The result favors reporting both dimensions when comparing generation settings.

Research draft: public rating metadata reanalysis; scientific interpretation and authorship await human review.

When Generation Settings Change the Gap between Image Quality and Prompt Alignment

Research draft: public rating metadata reanalysis; scientific interpretation and authorship await human review.

Abstract

A setting can improve how an image looks without improving how closely it follows its prompt. We paired the two released rating dimensions for 287 prompts observed under ten AGIQA-3K settings. Relative to the standard Stable Diffusion 1.5 setting, the alignment-minus-quality gap was 0.441 units larger for low-step Midjourney and 0.306 units smaller for standard Midjourney. Prompt-clustered intervals excluded zero for both contrasts. Rankings of setting means agreed only partly across dimensions, with Spearman correlation 0.770. These are differences on the released scales, rather than evidence that the underlying perceptual constructs share a natural unit. The result favors reporting both dimensions when comparing generation settings.

1. Two judgments attached to the same image

A text-to-image system faces at least two evaluation questions. Is the output visually convincing, and does it depict the requested content? A visually polished picture can omit an object or relation. An awkward picture can nevertheless follow a prompt closely. Comparing generation settings with one scalar therefore risks hiding a tradeoff. Our question concerns the released human assessments themselves, before fitting another predictor to them: do setting contrasts have the same direction on the quality and alignment scales?

AGIQA-3K provides a useful paired structure because its metadata contain both perceptual-quality and text-image-alignment scores (Li et al., 2023). The distinction is consistent with the explicit separation of semantic and distortion representations proposed by Meng, 2026. We use that work as motivation for examining the evaluation target. We do not implement its encoders, fusion mechanism, or training procedure, and the regressions below cannot establish the performance of that architecture.

The research problem also differs from constructing a universal visual embedding. CLIP learns image-language correspondence from paired data (Radford et al., 2021), whereas language-based image assessment adapts such representations to quality questions (Wang et al., 2022). Neither role implies that correspondence and perceptual acceptability should be interchangeable labels. A benchmark can support both tasks while retaining different scales, different response distributions, and different sources of observer disagreement. Those distinctions are central to our outcome definition.

Question-answering evaluation in Hu et al., 2023 and the compositional tests of Ghosh et al., 2023 illustrate ways to make correspondence more specific. Huang et al., 2023a likewise treats compositional generation as a collection of evaluative challenges. We do not calculate those metrics here. They identify a reason to resist treating a favorable global quality mean as evidence that all requested attributes, objects, counts, and spatial relationships were successfully rendered.

2. Estimand and expectations

Our first proposition is descriptive: the mean alignment-minus-quality gap differs across released generation settings after matching prompts. A zero setting coefficient would mean that the gap matches the selected baseline on the available scales. It would not mean that either setting has high quality. A positive coefficient could arise from relatively stronger alignment, relatively weaker quality, or both. We therefore retain setting means for each dimension alongside the regression contrasts.

The second proposition concerns the accompanying dispersion fields. Their average alignment-minus-quality difference may be nonzero within complete prompt blocks. This is a separate outcome, not a measure of regression precision. A larger released dispersion value does not automatically indicate a less reliable construct, because rating transformations and response distributions may differ. Measurement-error considerations in Carroll et al., 2006 and the agreement distinction discussed by Bland and Altman, 1986 motivate keeping score differences, disagreement, and estimation uncertainty separate.

Human-preference datasets such as Wu et al., 2023 and Kirstain et al., 2023 further show why evaluation depends on the judgment requested. A preference can incorporate attractiveness, relevance, or other considerations, whereas our two supplied dimensions retain their published labels. NIMA considers image assessment through rating information (Talebi and Milanfar, 2017), and MUSIQ represents visual quality at multiple scales (Ke et al., 2021). These are relevant alternatives for future predictive comparisons, but none was rerun for this metadata analysis.

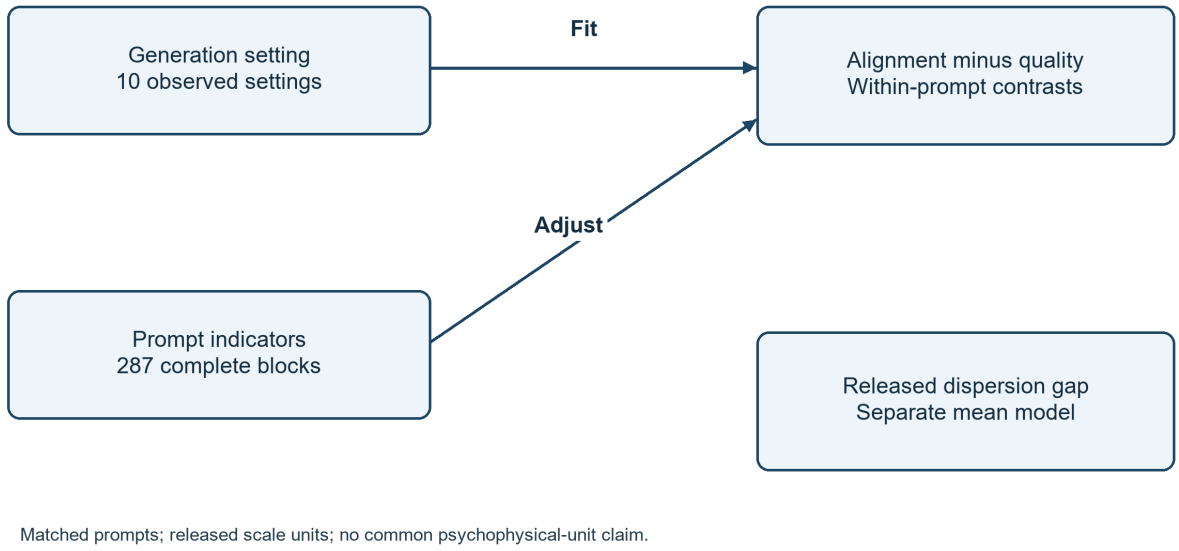

Figure 1. Generation-setting indicators explain the alignment-minus-quality gap within matched prompt blocks; a separate mean model describes the difference in released rating dispersions. Arrows denote fitted comparisons rather than causal effects.

3. Constructing a balanced prompt panel

The acquired CSV contains 2,982 images and 300 distinct prompts. Filenames identify ten settings after removing the final image-number suffix. We checked uniqueness of each prompt-setting pair. A prompt enters the main analysis only if all ten settings are observed. This leaves 287 prompts, 2,870 images, and 112 excluded images belonging to incomplete blocks. Complete blocks ensure that every setting contrast uses the same prompt collection instead of changing the prompt mixture across systems.

This restriction is a design decision, not an assertion that the excluded images are unimportant. Missing blocks could be associated with content difficulty or generation availability. The missing-data framework of Little and Rubin, 2002 cautions against interpreting complete-case convenience as evidence of random omission. Our estimand is explicitly the balanced subset. We do not impute absent scores or claim that its results describe all possible prompts. Source documentation and the processing record preserve the released and retained counts.

For each retained image, the primary outcome is the released alignment MOS minus the released quality MOS. Both fields are analyzed in their supplied scaled units. We do not relabel them as original five-point observer responses. The full-file distributions have similar overall normalization, but that feature alone cannot prove psychophysical comparability. Concordance and reliability concepts in Lin, 1989 and Koo and Li, 2016 help frame this limitation: related rankings, agreement of measurements, and validity of a common construct are different claims.

The secondary outcome subtracts the quality dispersion field from the alignment dispersion field. We use neither field as an inverse-variance weight. The release does not provide the observer-level design needed to reconstruct standard errors for every image under a verified common transformation. Consequently, the regression intervals quantify variation across retained prompt blocks under the stated covariance model. They do not reproduce the original rating experiment's uncertainty or account for all shared-rater dependence.

4. A within-prompt comparison

The main ordinary least squares model includes generation-setting indicators and prompt indicators. The reference is sd1.5_normal. Prompt intercepts remove additive differences in the gap across prompts, leaving setting comparisons within the balanced panel. This resembles the general repeated-measure problem treated by Laird and Ware, 1982, but we fit fixed prompt intercepts rather than a random-effects distribution. No prompt characteristic or image-content predictor is selected after examining statistical significance.

Covariance is clustered by prompt, with a finite-sample correction and t-based intervals across 287 clusters. Cameron and Miller, 2015 explains why repeated observations belonging to one sampling unit should not receive an independence assumption merely because they occupy separate rows. The heteroskedasticity argument of White, 1980 is relevant but does not alone address within-prompt dependence. Prompt clustering still leaves any dependence across prompts from common observers or generation pipelines outside the model.

The dispersion-gap model contains only an intercept and uses the same prompt clustering. Its coefficient is therefore the mean difference across retained images. Because the panel is balanced, each prompt and setting receives equal total representation. An intercept-only fit has zero explained variance by construction. We report that fact instead of presenting its R-squared as a substantive model-quality statistic. The meaning of each coefficient takes priority over comparing fit measures across unlike outcomes.

Table 1. Selected setting-gap contrasts relative to sd1.5_normal and the separate mean dispersion difference. Estimates and uncertainty use the released score units.

Model	Term	Estimate (SE)	95% CI	N
M1	AttnGAN_normal gap	-0.2509 (0.0438)	[-0.3372, -0.1646]	2870
M1	DALLE2_normal gap	0.2416 (0.0495)	[0.1442, 0.3389]	2870
M1	midjourney_lowstep gap	0.4414 (0.0461)	[0.3507, 0.5320]	2870
M1	midjourney_normal gap	-0.3061 (0.0433)	[-0.3914, -0.2209]	2870
M1: fit	R-squared = 0.3694	-	-	2870
M2	Mean contrast	0.1438 (0.0071)	[0.1297, 0.1578]	2870
M2: fit	R-squared = 0.0000	-	-	2870

Model labels: M1 = Alignment-minus-quality setting contrasts; M2 = Mean alignment-minus-quality rating dispersion.

The full setting model has R-squared 0.3694 with 2,870 rows. Table selection emphasizes two Midjourney settings and DALLE2 to keep the main comparison readable; the archived result file retains all setting and prompt terms. These entries were not selected as an exhaustive ranking of statistically significant generators. Influential-observation reasoning from Cook, 1977 remains useful for future prompt-level diagnostics, especially if an investigator wants to generalize beyond this particular prompt collection.

5. What the paired scores show

The standard sd1.5 setting has mean quality 2.8276 and alignment 2.8185, giving a gap of -0.0092. Low-step Midjourney has means 2.3745 and 2.8067, respectively. Its larger positive gap therefore does not indicate higher absolute quality than the baseline. The estimated relative gap is 0.4414, with standard error 0.0461 and interval 0.3507 to 0.5320. The practical interpretation requires both component means, not merely the sign of the contrast.

Standard Midjourney shows the opposite pattern: mean quality 3.6295 exceeds mean alignment 3.3142. Its gap relative to the reference is -0.3061, with interval -0.3914 to -0.2209. DALLE2 has a positive relative gap of 0.2416, with interval 0.1442 to 0.3389. These comparisons establish that the two dimensions do not differ by one common additive offset across all settings in this retained sample.

Across the ten setting means, the Spearman correlation between quality and alignment rankings is 0.7697. That association is substantial, yet it leaves meaningful reordering. It is a descriptive statistic based on ten setting summaries, not an independent image-level predictive validation. Structural similarity (Wang et al., 2004) and learned perceptual distance (Zhang et al., 2018) address other aspects of visual comparison; their existence does not allow us to infer correspondence performance from these human-score rankings.

The released alignment dispersion exceeds the quality dispersion by 0.1438 on average, with prompt-clustered interval 0.1297 to 0.1578. This result describes supplied fields only. It cannot determine whether alignment judgments are intrinsically harder, whether observers used different internal thresholds, or whether normalization altered dispersion. Calibration research such as Guo et al., 2017 concerns relationships between predictions and outcomes; a dispersion difference alone is not a calibration diagnosis of any deployed model.

6. Reading the result without overextending it

The estimates support a two-column reporting practice: show perceptual quality and prompt alignment separately, then specify any intended aggregation. A composite score requires a decision about relative importance, and that decision may differ between advertising, educational illustration, and faithful scene description. Our analysis does not estimate those application-specific weights. It demonstrates why silently fixing equal weights can obscure the source of a setting's apparent advantage.

Neither prompt matching nor a narrow confidence interval identifies a causal generation effect. Seed selection, image curation, software versions, and rating procedures may vary with settings. The causal-diagram perspective in Greenland et al., 1999 makes clear that controlling one shared prompt identifier does not eliminate all common causes. We therefore describe setting associations in released metadata. Assertions about the mechanism behind a semantic-distortion tradeoff would require additional experimental control and image-level evidence.

Multiple setting coefficients also make isolated threshold language inappropriate. Benjamini and Hochberg, 1995 and Benjamini and Yekutieli, 2001 provide frameworks for handling families of tests, including dependent tests, but this draft reports unadjusted descriptive intervals rather than a confirmatory discovery list. The interpretive guidance in Wasserstein and Lazar, 2016 argues against treating a p-value boundary as a scientific conclusion. Concerns about selective findings in Ioannidis, 2005 reinforce the need to retain the full coefficient record, including small and uncertain contrasts.

7. Data release and a focused conclusion

An additional distinction concerns the unit at which a setting is compared. Each coefficient summarizes 287 within-prompt differences. It does not describe the probability that one randomly selected image will be preferred to another, because the release supplies only one observed image for each retained prompt-setting combination. Repeated generations could reveal substantial variation around those cells. Treating that unobserved variation as zero would overstate the precision of any operational claim about what a user will receive on the next generation.

The gap outcome also depends on the relative scaling of its two components. Multiplying only the alignment field by a positive constant would preserve that dimension's ranking while changing the numerical gap and potentially its setting contrasts. Consequently, the gap is an audit of the published evaluation convention. The separate component means and rank summary are retained so that readers can recognize which findings rely on the subtraction convention and which concern ordering within each dimension.

A team wishing to adopt a single score could first specify a utility function in terms of actual task failures. For example, omission of a requested object may invalidate an instructional image even when its surface appearance is appealing. In another setting, modest correspondence errors may matter less than visible artifacts. The current metadata contain neither these application costs nor independent task-success labels. We therefore cannot infer a preferred aggregation from the observed setting averages alone, and we do not recommend a universal weighting rule.

Prompt clustering is appropriate for the repeated-setting design but is not a substitute for sampling prompts from an intended use population. The 287 complete prompts may omit rare linguistic structures, specialized terminology, or culturally specific descriptions. A future study should define that population before collecting generations and retain enough repeated images to separate prompt variability from generation variability. The existing panel provides a useful diagnostic starting point, while its intervals remain conditional on the narrower data construction described here.

The analysis package contains the original rating CSV, the balanced analysis table, the model outputs, and the selection checks. Documentation follows the dataset-accountability logic of Gebru et al., 2021 and the findability and reuse principles of Wilkinson et al., 2016. Public availability does not erase the source's attribution and reuse conditions. No raw observer identities are needed for our analysis, and no claim of new human-subject data collection accompanies the manuscript.

This study shares the broad topic of image evaluation with other drafts in the batch, but it uses a paired two-dimension question and a complete-prompt panel. It is not independent confirmation from a second rating experiment. A useful next investigation would hold prompts, seeds, observer procedures, and image selection fixed while varying a generation parameter. For the present release, the defensible finding is narrower: generator-setting comparisons depend on which human-rating dimension is retained.

References

1. 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.
2. 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.
3. 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.
4. 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.
5. Benjamini, Yoav and Daniel Yekutieli. "The control of the false discovery rate in multiple testing under dependency." *The Annals of Statistics*, vol. 29, no. 4, 2001, pp. 1165-1188. https://doi.org/10.1214/aos/1013699998.
6. 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.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
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. Talebi, Hossein and Peyman Milanfar. "NIMA: Neural Image Assessment" *arXiv preprint arXiv:1709.05424*, 2017. https://arxiv.org/abs/1709.05424.
15. 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.

16. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.

17. Ioannidis, John P. A.. (2005). Why Most Published Research Findings Are False. PLoS Medicine. https://doi.org/10.1371/journal.pmed.0020124.

18. 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.

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

20. 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.

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

22. Laird, Nan M. and James H. Ware. “Random-Effects Models for Longitudinal Data.” *Biometrics*, vol. 38, no. 4, 1982, pp. 963. https://doi.org/10.2307/2529876.

23. 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.

24. 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.

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

26. 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.

27. Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002. https://doi.org/10.1002/9781119013563.

28. Koo, Terry K., and Mae Y. Li. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." *Journal of Chiropractic Medicine*, vol. 15, no. 2, 2016, pp. 155-163. https://doi.org/10.1016/j.jcm.2016.02.012.

29. 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.

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