Journal of Audit and Accountability
Volume 1 · Number 1 · 2026

One Guidance Column, Several Operating Regimes: A Configuration Audit of AIGIQA-20K

Jack Reed[1] · Connor Franklin[2] · Matthew Fisher[3]

[1] Montenegro · [2] Slovakia · [3] Germany

Original publication: November 7, 2026 | Article 573

Abstract

In AIGIQA-20K, the same guidance field includes varying positive settings, fixed settings, and zeros attached to several generation regimes. A regression pooling all 20,000 images associated higher guidance with higher subjective ratings, with a coefficient of 0.00656. Adding generator indicators changed the estimate to -0.00184, with an interval spanning zero. A specification restricted to 14,000 images from generators with positive, varying settings was also inconclusive after adjustment. These comparisons expose the role of model composition and configuration interpretation. They do not establish that changing guidance causes an improvement or deterioration in generated images.

Empirical research draft using released ratings and configuration metadata. No new images were generated and no image-quality predictor was run.

One Guidance Column, Several Operating Regimes: A Configuration Audit of AIGIQA-20K

Empirical research draft using released ratings and configuration metadata. No new images were generated and no image-quality predictor was run.

Abstract

In AIGIQA-20K, the same guidance field includes varying positive settings, fixed settings, and zeros attached to several generation regimes. A regression pooling all 20,000 images associated higher guidance with higher subjective ratings, with a coefficient of 0.00656. Adding generator indicators changed the estimate to -0.00184, with an interval spanning zero. A specification restricted to 14,000 images from generators with positive, varying settings was also inconclusive after adjustment. These comparisons expose the role of model composition and configuration interpretation. They do not establish that changing guidance causes an improvement or deterioration in generated images.

1. Why a parameter column needs its generator

Large public image-quality datasets make it tempting to relate a generation setting directly to a subjective rating. A single column labeled classifier-free guidance appears to offer a common numerical scale across generators. Yet its interpretation depends on the operating regime that produced each row. A zero in an API-based generator's metadata does not establish that the service used the same internal procedure as a diffusion implementation explicitly configured without conventional guidance. Fixed settings also provide no within-generator variation from which to estimate a guidance relationship.

This paper asks how the observed association between the released guidance field and mean subjective score changes when generator composition is made explicit. The data come from AIGIQA-20K, whose design combines outputs from 15 generators and records configurations alongside subjective evaluations (Li et al., 2024). We analyze its public table of 20,000 images rather than the pixels themselves. The contribution is a configuration audit: it identifies the support for a numerical comparison, fits transparent associations, and keeps unsupported claims about parameter interventions outside the findings.

The distinction between semantics and perceptual degradation gives this audit a substantive motivation. Decoupling Semantics from Distortions proposes a multi-scale, two-stream vision-language approach to AI-generated image assessment (Meng, 2026). Its representational question differs from our metadata question. Nevertheless, both require specifying what a quality outcome measures. A dataset's single subjective score can combine several judgments, while a generation field can combine several operational meanings. Our analysis evaluates neither the proposed two-stream architecture nor its reported benchmark advantage.

2. Quality, alignment, and configuration are different measurements

Image-quality research has long distinguished ways of representing a visible discrepancy. Structural similarity compares image structure, learned perceptual metrics use feature representations, and a no-reference approach such as NIQE draws on statistical regularities without a pristine comparison image (Wang et al., 2004; Zhang et al., 2018; Mittal et al., 2013). These methods do not supply the outcome in our regressions. They illustrate why the phrase image quality cannot identify a measurement procedure by itself. Here the outcome is the released mean opinion score, preserving the dataset's subjective evaluation design.

Benchmark construction also changes the domain of a quality measure. KADID-10k supplies artificially distorted images, whereas KonIQ-10k addresses a natural-image quality setting (Lin et al., 2019; Hosu et al., 2019). AI-generated images may contain both aesthetic defects and failures to follow a prompt. AGIQA-3K provides another public resource for investigating generated-image assessment (Li et al., 2023). We do not merge its scores with AIGIQA-20K, since different assessment protocols and score scales cannot be assumed interchangeable merely because both datasets concern generated images.

Representation and supervision introduce further distinctions. NIMA studies neural image assessment, and MUSIQ uses a multi-scale architecture for quality assessment (Talebi and Milanfar, 2017; Ke et al., 2021). CLIP learns transferable image-language representations, while preference resources such as Pick-a-Pic and Human Preference Score investigate judgments about generated images (Radford et al., 2021; Kirstain et al., 2023; Wu et al., 2023). These sources motivate care in naming the observed judgment. Our table contains no separate respondent-level rating stream, pairwise preference outcome, or measured text-image correspondence score, so none is silently inferred from the aggregate MOS column.

The same restraint applies to distributional evaluation. Fréchet-based generative comparisons concern collections of image representations rather than the subjective quality of each individual output (Heusel et al., 2017). Assigning a model's distributional score to every one of its images would create a different dataset and a different level of analysis. We retain one released rating per image. Generator indicators are adjustment terms for this image-level association, and are not presented as a league table of generators whose performance has been experimentally equalized.

3. Operating regimes and the research question

Diffusion research provides a clear example of why configuration semantics matter. Denoising diffusion and denoising implicit formulations involve different sampling ideas, while latent diffusion moves substantial image modeling into a compressed representation (Ho et al., 2020; Song et al., 2020; Rombach et al., 2021). A recorded iteration count consequently does not automatically measure equal computational effort across architectures. Guidance likewise interacts with implementation details and the model's training procedure. Our analysis uses the released numerical values while avoiding a conversion into standardized computational cost or a universal intervention dose.

Differences among named generators reinforce this point. SDXL introduces conditioning and architecture choices within a latent diffusion setting, and hierarchical text-conditional generation through CLIP latents uses another design (Podell et al., 2023; Ramesh et al., 2022). We cite these works to establish model context, not to reconstruct undocumented settings for individual benchmark images. Where a generator has a fixed or zero-coded field, the public table tells us how the field was recorded. It does not reveal every internal sampling decision or authorize replacing the value with a guessed positive setting.

We frame three exploratory comparisons. P1 is the pooled image-level association between guidance and MOS across the complete table. P2 adds generator-specific intercepts, shifting the guidance comparison toward variation within generators. P3 restricts the sample to generators whose recorded guidance is positive and varies, then adjusts for iterations, image area, and prompt length. These are distinct estimands and are labeled accordingly. A change between their coefficients can reveal sensitivity to model composition or specification, but cannot identify the unique causal contribution of zero coding.

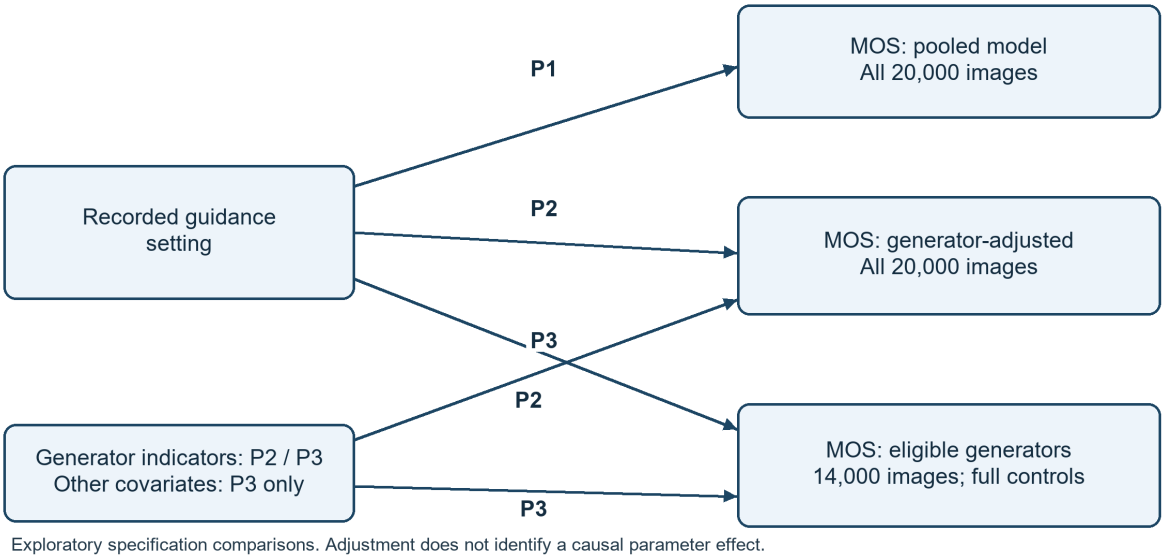

Figure 1. Research model for three exploratory specification comparisons. Guidance is related to the released MOS under a pooled model, a model with generator indicators, and an eligible-generator model with additional configuration covariates. Eligibility requires positive and varying guidance within a generator. Arrows describe regression terms and do not imply randomized parameter assignment.

4. Public table and eligibility decisions

We downloaded the public AIGIQA-20K metadata table and retained its original file and hash. It has 20,000 rows with image name, prompt, generator name, iterations, guidance, width, height, and MOS. Required fields used in this analysis contain no missing values. There are 15 named generator categories and 19,259 distinct prompt strings. The observed ratings range from approximately 0.1223 to 4.9814. These are aggregate subjective outcomes from the release, with no access in this analysis to individual raters' repeated responses or their covariance structure.

A generator-by-field profile precedes modeling. Five thousand five hundred rows have guidance recorded as zero. Several generators have only one observed guidance setting; DeepFloyd-IF, for example, has a fixed positive value in this table. Seven generators have guidance that is both strictly positive and varying: Dreamlike Photoreal v2, Pixart XL 2, Playground v2, Segmind Stable Diffusion 1B, Stable Diffusion XL, Stable Diffusion v1.4, and Stable Diffusion v1.5. Their combined 14,000 images form the third specification's eligible sample.

Eligibility is defined from the field profile rather than from the resulting rating association. We do not relabel every zero as a missing measurement. Some zeros can represent a distinct operating mode, whereas other implementations may not expose a comparable setting. Missing-data theory makes this distinction between an observed code and a substantive missingness assumption important (Little and Rubin, 2002). The audit preserves all original values in its complete-table models and uses the restricted sample to show what can be learned where positive within-generator variation is directly observed.

For the adjusted eligible-sample model, image area is width multiplied by height and entered on a natural-log scale. Prompt length is the count of word-pattern matches, transformed as log(1 + words). Iteration count remains on its recorded scale. These operational definitions are stored with the derived analysis table. Dataset documentation makes such construction choices available for inspection and reuse (Gebru et al., 2021). Prompt length is a textual summary, not a direct measurement of semantic complexity or the difficulty of satisfying a requested composition.

5. Estimating associations with different support

All three specifications use ordinary least squares. The first regresses MOS only on guidance. The second includes the same numerical predictor and generator indicators. The third adds iterations, logged area, and logged prompt length within the eligible group, again retaining generator indicators. A common guidance slope is assumed inside each specification. Generator-specific interactions could relax that assumption, but would answer a more detailed question and are not fitted opportunistically to find a favorable subgroup coefficient.

HC3 standard errors and 95 percent intervals accompany each estimate, following the general motivation for heteroskedasticity-consistent inference (White, 1980). In this dataset, a model with generator indicators does not make the included generator collection a random sample from a defined population. Nor does image-level HC3 account for every possible relationship among repeated prompts or common generation workflows. Cluster-robust inference requires a defensible grouping and sampling interpretation (Cameron and Miller, 2015). We therefore describe the intervals as conditional model summaries, with residual dependence and generator selection retained as limitations.

The subjective score is treated as a continuous released outcome. It combines perceptual quality and alignment considerations in the source evaluation rather than separating them into independent response variables. Measurement-error analysis provides a useful distinction between an observed proxy and the latent construct a researcher hopes to explain (Carroll et al., 2006). A guidance coefficient on MOS cannot establish whether any association operates through cleaner texture, stronger prompt compliance, aesthetic style, or a mixture of these mechanisms. That information is absent from the analyzed table.

Table 1. Executed configuration-association models. Parentheses contain HC3 standard errors and brackets contain 95 percent intervals. P1 uses all images without generator adjustment; P2 adds generator indicators. P3 uses 14,000 eligible images and includes generator indicators, iterations, log image area, and log-transformed prompt length. Fit rows report each specification's R-squared.

Model	Term	Estimate (SE)	95% CI	N
Pooled zero-coded audit	Recorded guidance	0.0066 (0.0013)	[0.0040, 0.0091]	20000
All-model intercept adjustment	Recorded guidance	-0.0018 (0.0016)	[-0.0049, 0.0012]	20000
Positive varying guidance settings	Recorded guidance	-0.0025 (0.0014)	[-0.0053, 0.0003]	14000
Pooled zero-coded audit: fit	R-squared = 0.0015	-	-	20000
All-model intercept adjustment: fit	R-squared = 0.4964	-	-	20000
Positive varying guidance settings: fit	R-squared = 0.4965	-	-	14000

6. A positive pooled coefficient does not survive the same interpretation

The complete-table pooled guidance coefficient is 0.00656, with interval [0.00401, 0.00911]. The associated R-squared is only 0.00153. This positive association is statistically distinguishable from zero under the fitted covariance assumptions, but guidance alone explains very little of the observed rating variation. Because the table mixes generator categories with different settings and score distributions, the coefficient combines between-generator differences and within-generator comparisons. It should not be translated into advice that increasing a particular system's guidance will raise its expected quality.

After generator indicators enter, the coefficient is -0.00184 with interval [-0.00490, 0.00123]. Its direction differs from the pooled point estimate, and the interval includes zero. Model R-squared rises to 0.4964, reflecting the much larger amount of variation represented by generator-specific mean differences. This fit improvement does not establish generator identity as a causal explanation of quality. It shows that the observed collection has substantial group structure that the one-predictor pooled calculation leaves inside its guidance comparison.

The eligible-sample model estimates a guidance coefficient of -0.00250, with interval [-0.00531, 0.00031] and R-squared 0.4965. Again the interval spans zero. The result concerns positive, varying configurations within seven generators after the stated adjustments. It is not directly the same population contrast as the complete-table model. In particular, removing fixed-setting generators and adding configuration covariates are simultaneous specification choices in P3. Their separate effects cannot be identified merely by comparing its final coefficient with P1.

These differences limit a simple dose-response account. The positive pooled association is compatible with no clear common conditional relationship in the adjusted specifications. It does not follow that guidance has no effect in every model, that a negative setting change is preferable, or that zero-coded entries alone explain the sign change. The evidence establishes sensitivity to the modeling unit and adjustment set. Any stronger explanation would require a design capable of separating generator selection, parameter assignment, and the components of subjective judgment.

7. Consequences for benchmarking practice

A useful metadata audit can be performed before training a new quality predictor. It should publish generator-level ranges, identify constant columns, distinguish exposed settings from internal procedures, and name the eligible comparison population. The current study does this without substituting a fitted configuration equation for visual evaluation. Future model comparisons should also preserve appropriate validation boundaries: data with shared generation conditions may require structured splitting (Roberts et al., 2017), while selecting features or settings on evaluation outcomes can lead to optimistic performance estimates (Varma and Simon, 2006).

Measurement disagreement deserves its own design. Agreement between two scores is not established by a shared average relationship with guidance. Method-comparison approaches explicitly consider paired measurements and their discrepancies (Bland and Altman, 1986). Reliability assessment similarly requires specifying which sources of variation and repeated judgments are being evaluated (Koo and Li, 2016). Our table contains one aggregate MOS per image and no independently computed quality score. Consequently, the regressions provide neither an agreement coefficient nor an estimate of inter-rater reliability, even though those would be relevant additions to a future assessment study.

Threshold-based applications create another layer. If a deployment system labels images acceptable or unacceptable, it must define that boundary and evaluate predictions against an appropriate outcome. Precision-recall considerations become relevant when acceptable or unacceptable cases are unevenly represented (Saito and Rehmsmeier, 2015). We do not choose such a cutoff from the present ratings. Retaining the continuous outcome avoids turning an exploratory coefficient into an unsupported moderation or publication rule. The dataset's observed MOS range also provides no universal acceptability threshold for other rating protocols.

Finally, an expanded analysis might compare many model-specific slopes or several quality dimensions. Such a family should be declared and reported in full, with an explicit approach to multiplicity where inferential screening is undertaken (Benjamini and Hochberg, 1995). Our three specification comparisons

are exploratory and all remain visible, including the adjusted intervals that cross zero. We make no claim of preregistration, causal identification, or out-of-sample predictive validation. These boundaries allow the configuration audit to inform later experiments without overstating what the downloaded metadata alone can establish.

8. Conclusion

A shared guidance column does not guarantee a shared operating regime. In AIGIQA-20K, its positive pooled association with subjective ratings changes direction and becomes inconclusive after generator adjustment; an eligible-generator specification is also compatible with zero. The result supports documenting configuration support and generator composition before interpreting a benchmark-wide slope. An experiment that varies a specified parameter within a defined generator, while evaluating perceptual and semantic outcomes separately, would be needed to move from this audit to claims about how settings change image quality.

References

1. 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.
2. Song, Jiaming; Meng, Chenlin; Ermon, Stefano. (2020). Denoising Diffusion Implicit Models. arXiv:2010.02502. https://arxiv.org/abs/2010.02502.
3. Ke, Junjie, et al. "MUSIQ: Multi-scale Image Quality Transformer" *arXiv preprint arXiv:2108.05997*, 2021. https://arxiv.org/abs/2108.05997.
4. 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.
5. 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.
6. Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002. https://doi.org/10.1002/9781119013563.
7. 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.
8. Ramesh, Aditya; Dhariwal, Prafulla; Nichol, Alex; Chu, Casey; Chen, Mark. (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125. https://arxiv.org/abs/2204.06125.
9. 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.
10. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.
11. Podell, Dustin; English, Zion; Lacey, Kyle; Blattmann, Andreas; Dockhorn, Tim; Müller, Jonas; Penna, Joe; Rombach, Robin. (2023). SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv:2307.01952. https://arxiv.org/abs/2307.01952.
12. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.
13. 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.
14. Saito, Takaya and Marc Rehmsmeier. “The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets.” *PLOS ONE*, vol. 10, no. 3, 2015, pp. e0118432. https://doi.org/10.1371/journal.pone.0118432.
15. 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.

16. Mittal, A., et al. "Making a “Completely Blind” Image Quality Analyzer" *IEEE Signal Processing Letters*, 2013. https://doi.org/10.1109/lsp.2012.2227726.

17. 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.

18. 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.

19. 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.

20. Rombach, Robin; Blattmann, Andreas; Lorenz, Dominik; Esser, Patrick; Ommer, Björn. (2021). High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752. https://arxiv.org/abs/2112.10752.

21. 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.

22. Ho, Jonathan; Jain, Ajay; Abbeel, Pieter. (2020). Denoising Diffusion Probabilistic Models. arXiv:2006.11239. https://arxiv.org/abs/2006.11239.

23. 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.

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

25. 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.

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

27. 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.

28. 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.

29. 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.

30. 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.