Generalization across Prompts in a Small Public Quality-Score Table

Arthur Brooks (UA), Oliver Wheeler (KP), Elijah Bailey (TT)

Abstract

A simple score-based regression predicts released quality means across held-out prompts with a root mean squared error of 0.493, compared with 0.998 for a training-mean baseline. The five-fold exercise groups all observations of each prompt together. Crucially, the predictor requires the already observed human alignment mean, alongside condition, style, and adjective count. It is therefore a secondary association model rather than a deployable image-only assessor. The result illustrates both the value of prompt-aware validation and the importance of declaring what information prediction requires.

A predictive result needs an information boundary

Predicting one published human outcome from another can be useful for understanding a benchmark, but it is a different task from assessing an unseen image without human judgments. The distinction should be stated before interpreting a favorable error rate. This study examines a small linear model of quality means in AGIQA-3K, using alignment means and recorded metadata. Its question is whether the fitted relationship transfers to held-out prompt identifiers within the same released table and generation-condition population.

Decoupling Semantics from Distortions proposes an image-quality architecture using multiscale visual-language evidence (Meng, 2026). The present regression does not reproduce that architecture or compete with it on the same input task. It has access to human alignment information that an image-only predictor would ordinarily need to estimate. The comparison is therefore conceptual and evaluative: it clarifies why semantic information can be associated with quality while emphasizing that a model's practical claim depends on what is available at prediction time.

Grouped validation and model inputs

The input contains 2,982 rows associated with 300 prompt identifiers and ten generation conditions. A fixed random partition, using seed 541, assigns the prompt identifiers to five folds after shuffling their sorted list. All rows of a prompt remain together. For each held-out fold, ordinary least squares is fitted on the remaining folds. Predictors are the released alignment mean, condition indicators, style indicators, and the count of nonmissing entries in the two adjective fields. The quality mean is the target.

Predictions are collected only for each fold's held-out rows. The baseline predicts the training portion's mean quality for every held-out row. No hyperparameter search, image-feature extraction, or newly trained visual model is part of the exercise. The grouping prevents identical prompt identifiers from appearing on both sides of a fold boundary. It does not guarantee semantic disjointness among differently worded prompts or independence from every source of benchmark construction. Those stronger boundaries have not been measured here.

Pooled out-of-fold root mean squared error is 0.493, and mean absolute error is 0.375. The training-mean baseline has root mean squared error 0.998. The pooled predictive coefficient of determination is 0.756 when errors are compared with variation around the overall observed target mean. Correlation between held-out predictions and targets is 0.869. These quantities summarize one fixed prompt partition. They are not confidence intervals and do not establish stability across every possible partition or a new external dataset.

What transfers and what remains untested

The exercise supports a bounded conclusion: within this released population, alignment scores and available metadata contain information associated with quality on prompts not used to fit that fold's coefficients. It does not establish transfer to new generation conditions, different rating scales, or images without human alignment judgments. The condition categories represented in training remain part of the prediction setting. A new-generator test would require another split and potentially a model capable of handling genuinely unseen condition levels.

The human alignment input is particularly consequential. Calling the regression an automated quality system would conceal a substantial part of its information budget. A practical pipeline could attempt to replace that input with an automated alignment estimate, but then both the new measurement error and the complete pipeline would need evaluation. The present results cannot be carried over unchanged to that hypothetical system. Similarly, no comparison with the focal image-quality architecture follows without obtaining its predictions on a suitably matched assessment task.

The grouped design is still useful as a reproducible baseline for score relationships. It improves on a row split that would distribute observations of the same prompt across fitting and assessment. Future work could reserve external prompts, test repeated partitions, and obtain independent visual predictions under declared input constraints. Those additions should answer specific unresolved questions rather than merely repeat evaluation until a preferred result appears. The saved fold assignments and held-out predictions make the current exercise auditable under its stated boundary.

A prediction setting with human alignment already available

A comparison of training mean baseline and metadata and human-alignment regression can clarify out-of-fold error on previously withheld prompt identifiers. An analyst estimates the likely quality mean for a new prompt using its already collected human alignment ratings. This hypothetical example would be recorded at the level of prompt held out as a group. A better score-only model may still be impractical for image-only screening because the human predictor is costly. Report the available inputs and reserve new generator families for a separate transport assessment.

Make the prediction task explicit at the level of inputs available for a new row. If a human alignment mean has already been collected, the model can use it to predict a human quality mean. An image-only scorer has a different information budget. A test that holds out entire prompts probes one kind of transfer. Holding out generation conditions or a dataset would probe others. Document fitting, baseline and validation on the same eligible inputs before interpreting any reduction in error.

An initial table can record the eligibility and observation status of prompt held out as a group, before any performance statistic is calculated. For the prompt held out as a group, mark ambiguous images and distinguish a low score from an unavailable judgment.

Record the prompt, model settings and rating instructions separately for prompt held out as a group. If there are many observations of prompt held out as a group, decide which ones share an underlying history and keep that grouping in any uncertainty analysis. When examining metadata and human-alignment regression, compare sensitivity to localized changes with sensitivity to a modification affecting the whole frame.

A sensitivity display can show where the conclusion about out-of-fold error on previously withheld prompt identifiers changes as assumptions about the eligible observations change. The analysis of training mean baseline should state whether pair order, prompt display or image scaling could change the human

judgment. Set aside examples of prompt held out as a group for independent assessment, rather than testing transfer by revisiting the cases that motivated the protocol.

Measurement and error

Another study uses natural image statistics for blind quality assessment (Mittal et al., 2013). Opinion-score distributions retain information about rater agreement that disappears when every image is represented by a single mean value. Unusual image statistics can signal degradation. Include deliberate styles that depart from those statistics without looking defective. Related work learns to predict distributions of human image assessment scores (Talebi and Milanfar, 2017).

A methodological precedent collects subjective judgments at image and patch scales (Ying et al., 2019). In an evaluation, one study provides human quality judgments for controlled image distortions (Lin et al., 2019). Controlled degradations allow comparisons at known distortion levels and provide a useful contrast with uncontrolled artifacts in model-generated images. Paired image-level and patch-level assessments allow investigation of whether a single defective region drives the rating of an otherwise plausible composition.

An existing method uses autoregressive modeling for text-conditioned image synthesis (Yu et al., 2022). Another study discusses a kernel-based distributional distance for comparing generated samples (Bińkowski et al., 2018). During validation, comparing feature distributions can answer a collection-level question, but the sample composition and case-level errors still need to be described. An autoregressive generator can broaden an evaluation split beyond diffusion outputs and test whether the scoring rule transfers across generation regimes.

Separating uncertainty in the measured inputs from variability in the underlying property is important when the same noisy proxy also defines the strata. Collection-level feature statistics depend on sample count and composition, which can mask a local failure affecting only a few cases. In an evaluation, one study investigates statistical consequences of imperfectly measured variables (Carroll et al., 2006). An existing method compares feature distributions of generated and reference image collections (Heusel et al., 2017).

An existing method uses multiscale image representations for quality prediction (Ke et al., 2021). During validation, one approach examines uncertainty estimates under dataset shift (Ovadia et al., 2019). A multiscale score can help distinguish sensitivity to resampling or crop size from a stable image-level judgment. Confidence calibrated on the original collection may change under a new capture or generation regime, so calibration belongs in transfer testing.

Another study learns visual representations through self-supervision (Caron et al., 2021). An existing method collects human quality and alignment judgments for generated images (Li et al., 2023). During validation, its score fields make it possible to compare visual ratings with prompt alignment and examine whether disagreement is concentrated in particular prompts. Using a representation learned without explicit class labels gives a useful comparison when a prompt-derived score may reward the wrong visual feature.

A methodological precedent uses attention-based image representations for no-reference quality assessment (Yang et al., 2022). In practical terms, a language-image quality representation should be checked against pairs with controlled content so that prompt relevance does not dominate the distortion judgment. A no-reference model provides a comparison on images without clean originals, especially when known distortion manipulations can be added as controls. Related work uses contrastive language-image features for assessing visual qualities (Wang et al., 2022).

A methodological precedent tests compositional text-to-image correspondence across multiple attributes and relationships (Huang et al., 2023a). In practical terms, an existing method learns a reward model from human preferences for text-to-image generations (Xu et al., 2023). Attribute and relation scores can distinguish drawing the correct object from assigning it the wrong color or location. Human preference supervision provides a ranking benchmark, with disagreements worth examining where raters might favor stylistic polish over the specified content.

A methodological precedent models image generation through CLIP latents (Ramesh et al., 2022). In a controlled test, the separation of language conditioning and successive image resolutions provides a useful distinction when analyzing failures in composition versus small visual detail. Related work uses a language encoder and cascaded diffusion for text-to-image generation (Saharia et al., 2022). A generator using aligned visual and language latents motivates comparing semantic scores with ratings of technical artifacts.

A whole-image embedding reflects scene description and background alongside subject appearance. A local check can distinguish these contributions. In-the-wild ratings capture naturally occurring quality variation that may be absent from controlled distortion benchmarks. One study collects quality judgments for images with in-the-wild distortions (Hosu et al., 2019). In practical terms, an existing method learns aligned visual and language representations through contrastive training (Radford et al., 2021).

Question-based scoring can distinguish missing objects from incorrect relations when the questions and answers are archived alongside the generated examples. Pairwise preferences support ranking checks, provided ties and same-prompt pairs are handled explicitly before summarizing disagreement between methods. For a prospective comparison, a relevant contribution evaluates prompt faithfulness using questions about generated images (Hu et al., 2023). Related work records comparative human preferences for generated images (Kirstain et al., 2023).

One approach investigates a human-preference score for generated images (Wu et al., 2023). One published study uses object-level checks to assess compositional image generation (Ghosh et al., 2023). During validation, object-level checks make individual omissions and counting errors observable, although detection failures can still contaminate the final comparison if they are excluded. A comparative preference score reflects aesthetics, composition and prompt adherence. Compare it with a measure of the particular failure under study.

Analysis and interpretation

A methodological precedent uses diffusion in a learned latent space for conditional image synthesis (Rombach et al., 2021). An existing method shows bias from reusing validation for selection and error estimation (Varma and Simon, 2006). For implementation, reusing a validation split for repeated configuration selection can make the final error estimate optimistic; retain a separate assessment for the chosen rule. Latent compression may influence reconstruction details independently of prompt compliance, making downstream checks on both properties informative for comparisons across generators.

Linking evidence to the evaluation

An existing method adapts image synthesis to a specific subject using example images (Ruiz et al., 2022). One published study connects validation splits to temporal, spatial and hierarchical dependence (Roberts et al., 2017). For implementation, one approach reviews matching as a design step for comparing observational units (Stuart, 2010). Keep reference and evaluation views separate so that recognition after adaptation can be distinguished from successful reproduction of a training portrait. Matching can improve comparability on recorded attributes; an overlap display should identify which observations are lost before the outcome contrast is computed. Validation should respect shared identities, scenes or capture times when units near one another would otherwise leak into nominally held-out data.

Conclusion

The score-based regression improves on a training-mean baseline across held-out prompt identifiers in one fixed five-fold exercise. Its favorable error depends on access to released human alignment means and existing metadata. The result is therefore a bounded predictive association, not an image-only quality benchmark. It shows why grouping and input disclosure belong together: a held-out result is informative only when readers know both what was withheld and what information the predictor still received.

References

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.

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.

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.

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.

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.

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.

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.

Ramesh, Aditya, et al. "Hierarchical Text-Conditional Image Generation with CLIP Latents." *arXiv preprint arXiv:2204.06125*, 2022. https://arxiv.org/abs/2204.06125.

Saharia, Chitwan, et al. "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding." *arXiv preprint arXiv:2205.11487*, 2022. https://arxiv.org/abs/2205.11487.

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.

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

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.

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.

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.

Ruiz, Nataniel, et al.. "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation" *arXiv preprint arXiv:2208.12242*, 2022. https://arxiv.org/abs/2208.12242.

Rombach, Robin, et al. "High-Resolution Image Synthesis with Latent Diffusion Models." *arXiv preprint arXiv:2112.10752*, 2021. https://arxiv.org/abs/2112.10752.

Meng, Zijie. "Decoupling Semantics from Distortions: Multi-Scale Two-Stream Vision-Language Alignment for AI-Generated Image Quality Assessment." *arXiv preprint arXiv:2606.16799* (2026).

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.

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

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

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.

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

Yu, Jiahui, et al. "Scaling Autoregressive Models for Content-Rich Text-to-Image Generation." *arXiv preprint arXiv:2206.10789*, 2022. https://arxiv.org/abs/2206.10789.

Bińkowski, Mikołaj, et al. "Demystifying MMD GANs." *arXiv preprint arXiv:1801.01401*, 2018. https://arxiv.org/abs/1801.01401.

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.

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

Stuart, Elizabeth A. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science*, vol. 25, no. 1, 2010, pp. 1-21. https://doi.org/10.1214/09-sts313.

Ovadia, Yaniv, et al. “Can You Trust Your Model’s Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift.” *Advances in Neural Information Processing Systems*, vol. 32, 2019. https://arxiv.org/abs/1906.02530.

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

Caron, Mathilde, et al.. "Emerging Properties in Self-Supervised Vision Transformers" *arXiv preprint arXiv:2104.14294*, 2021. https://arxiv.org/abs/2104.14294.