



OPEN Beyond distortions: a benchmark for subjective evaluation of image rendering quality

Vsevolod Plokhotnyuk^{1,2,3}, Artyom Panshin^{1,3}, Nikola Banić⁴, Simone Bianco⁵✉, Michael Freeman⁶ & Egor Ershov^{1,2,3}

Traditional Image Quality Assessment (IQA) has primarily aimed to quantify perceptual quality in terms of technical degradations such as noise, blur, or compression artifacts. However, in image rendering, the key factor influencing perceived quality is not the presence of degradations but the manner in which color processing algorithms are applied, as they directly shape the overall aesthetic appearance of the image. To date, the quantitative evaluation of how rendering methods affect image quality has been insufficiently addressed. In this work, we introduce Image Rendering Quality Assessment (IRQA) as a new problem setting within IQA and present REPID, a benchmark designed for its study. REPID contains 30,000 edited images and preference annotations collected from 13,648 voters, resulting in an over 2.5 million unique votes. Based on REPID, we investigate content-dependent render preferences and the influence of rendering parameters, and further explore applications such as aesthetic preference prediction (including personalization), render ranking, and benchmarking of aesthetic evaluation methods. We further perform an extensive experimental comparison of traditional IQA metrics, handcrafted features, deep learning approaches, and foundation-model embeddings. On the REPID benchmark, IRQA-specific models achieve up to 40% better precision than conventional distortion-oriented IQA methods.

Keywords Image quality assessment, Rendering, Crowd-sourcing, Subjective metrics

Image Quality Assessment refers to the task of automatically evaluating the quality of images in a way that aligns with human perception. Accurate IQA is critical for the development and optimization of computer vision and image processing algorithms, as it provides an objective signal that guides models toward producing results that are visually pleasing or perceptually faithful². Whether in image compression, enhancement, restoration, or generation, IQA metrics are expected to help ensure that algorithmic outputs meet human expectations of quality, as well as to measure the performance of the algorithms themselves.

Traditional IQA research has focused primarily on assessing the degradation of perceived image quality introduced by the presence of distortions such as noise, blur, or compression artifacts. Over time, however, the field recognized that perceived image quality extends beyond simple degradations: viewers judge an image not only by the absence of artifacts, but also by how convincingly it conveys realism and visual appeal. This shift in perspective motivated the creation of modern subjective IQA datasets, such as UHD-IQA³, which capture human judgments on more holistic attributes, including composition, resolution, and color harmony, without requiring a pristine reference. However, even in subjective IQA datasets³, participants are typically asked to rate the overall quality of individual images, not to compare different stylistic renditions of the same scene. As a result, such datasets capture judgments about technical fidelity or perceptual realism rather than aesthetics of rendering, and are not suited for modeling user preferences between alternative rendering interpretations, which could be captured much more easily if datasets such as UHD-IQA contained pairwise evaluation results.

Image Rendering Quality Assessment introduces a qualitatively different challenge. Aesthetics of rendering is not characterized by degradations, but by aesthetic attributes such as color usage, texture, and artistic expression. Importantly, while image degradations are known to be of lesser quality than the original image, multiple renderings can be equally valid and pleasing, making rendering preference inherently subjective and context-dependent. Unlike quality degradations, which tend to have a broadly agreed-upon negative impact, rendering

¹Moscow Independent Research Institute of Artificial Intelligence, Moscow, Russia. ²AXXX, Moscow, Russia.

³Color Reproduction and Synthesis Institute, Moscow, Russia. ⁴Gideon Brothers, Zagreb, Croatia. ⁵University of Milano-Bicocca, Milan, Italy. ⁶Michael Freeman Photography, London, United Kingdom. ✉email: simone.bianco@unimib.it

variation often reflects personal, cultural, or situational preferences. Finally, because there is no objectively correct answer for the most pleasant rendering, it is impossible to guarantee whether users select their responses attentively, making the annotation process vulnerable to careless behavior⁴.

Despite growing interest in modeling visual aesthetics and render, the field lacks robust datasets that reflect the subjective nature of style perception. Existing IQA datasets are not suitable for learning or evaluating models that aim to predict human render judgments, as they are designed for scenarios with clear perceptual rankings, see Fig. 1. This absence of IRQA data hinders progress in developing algorithms that can interpret and predict nuanced aesthetic preferences.

To address this gap, we introduce the Rendering Evaluation of Photographic Image Dataset (REPID), a new benchmark specifically designed to allow learning subjective judgments of image rendering. REPID is constructed by augmenting the MIT-Adobe FiveK dataset⁵ with pairwise human preference scores obtained through a crowdsourced evaluation platform. Each pair of image renderings was evaluated by at least 25 evaluators, with a total of 13,648 unique evaluators participating overall. These annotations enable the development of IRQA models that go beyond degradation-focused assessments and instead capture the more nuanced, human-centered perception of rendering.

The dataset includes both the complex, photographer-specific rendering variations originally present in the MIT-Adobe FiveK dataset, as well as simpler rendering modifications such as subtle adjustments in contrast. This diversity allows REPID to reflect a broader and more realistic range of aesthetic choices. To the best of our knowledge, this is the first dataset of its kind, making it possible to conduct experiments on subjective rendering perception that were previously impractical or infeasible.

The paper is structured as follows: Section 2 gives a preview of the related work, in Section 3 the new Rendering Evaluation of Photographic Image Dataset is described, in Section 4 the analysis of users' preferences is performed, Section 5 describes some new models for IRQA and their quality, in Section 6 the results are presented, and Section 7 concludes the paper.

Related work

The section on related work is organized into two key subsections, aligning with the paper's core contributions. First, existing research on the development of standardized benchmark databases for IQA is comprehensively reviewed. Second, the state-of-the-art methods for no-reference (NR) IQA are critically analyzed. This dual-pronged approach enables a systematic examination of both (1) the foundational datasets essential for reliable evaluation and (2) the most advanced techniques driving progress in NR-IQA.

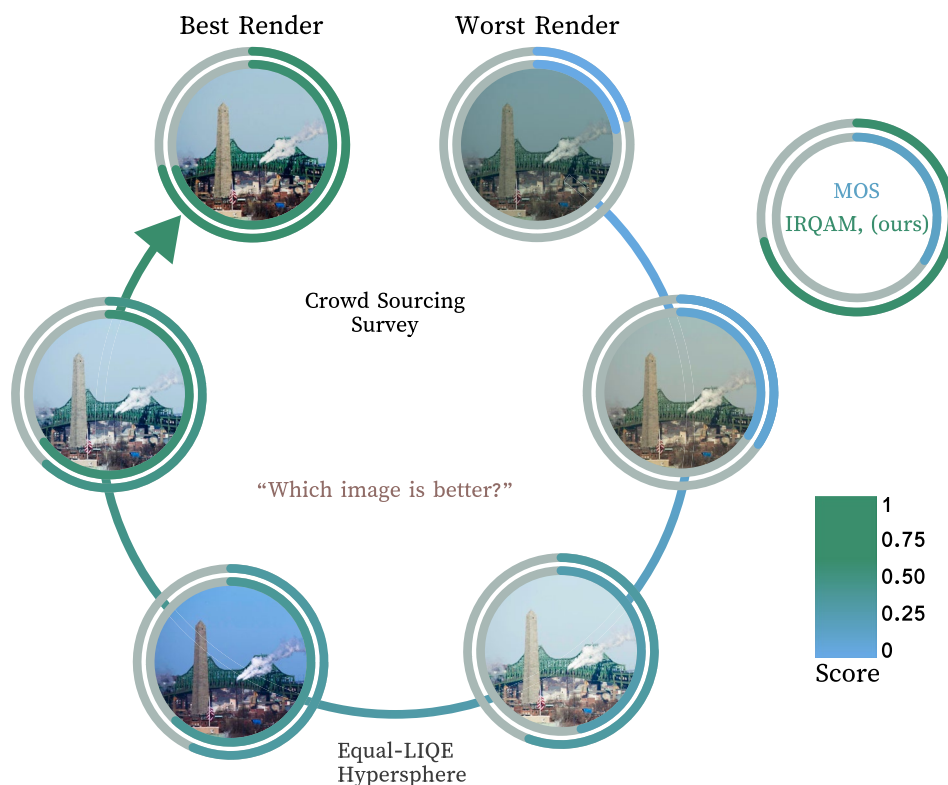


Fig. 1. Limitations of the existing IQA methods: while LIQE¹ assigns nearly identical quality scores with standard deviations from the mean value 3.5% across evaluations, subjective evaluation reveals clear differences. Our work addresses this gap by introducing a new benchmark Rendering Evaluation of Photographic Image Dataset (REPID).

Benchmark IQA datasets

In recent years, numerous IQA datasets have been introduced. Benchmarks such as LIVE⁶, VCL@FER⁷, NITS-IQA⁸, and CSIQ⁹ comprised limited sets of images with diverse distortion types, annotated using controlled laboratory studies. Subsequent efforts such as TID2013¹⁰ and MDID¹¹ expanded dataset sizes but remained insufficient for training data-hungry deep learning models due to their inherent scale limitations. A key bottleneck is the labor-intensive and costly subjective scoring process, which hinders the creation of large-scale datasets. As a result, only a few sizable datasets such as PDAP-HDDS¹² exist. Moreover, synthetically marked-up datasets also exist, e.g., KADIS-700k¹³ with markups obtained by a model trained on KADID-10k¹³, but unfortunately their utility is compromised for certain evaluation tasks.

Recent studies^{14,15} demonstrate that, with proper experimental design, crowdsourced annotations can achieve a reliability comparable to laboratory settings. Moreover, in prior work⁴, the reliability and stability of using crowdsourcing for stylistic and rendering evaluation was investigated, demonstrating its effectiveness as a viable data collection method. Crowdsourcing is becoming increasingly important and relying on it has enabled the creation of large-scale, diverse IQA datasets such as KonIQ-10k¹⁶ and UHD-IQA³, which address the scalability limitations of traditional approaches. Unlike earlier benchmarks constrained by small sample sizes or missing subjective labels, these datasets leverage crowdsourcing to balance scale and annotation quality, making them more suitable for modern data-driven IQA methods.

The aforementioned works primarily address objective IQA, focusing on technical distortions such as noise or compression artifacts. In contrast, a separate line of research targets image aesthetics evaluation, which evaluates subjective preferences and perceptual appeal. Prominent datasets in this domain include AVA¹⁷, AVA-PD¹⁸, and EVA¹⁹. Unlike technical IQA benchmarks, these datasets evaluate aesthetics by comparing entirely different scenes, where users select the most visually pleasing image within a broad theme, e.g., “flying bird”. Finally, recent work PARA²⁰ also investigates image aesthetics. However, its annotation methodology does not involve direct comparisons between multiple renderings of the same scene, which limits its ability to capture user preferences specific to image post-processing, while chosen annotation methodology limits achievable precision²¹. This makes it difficult to assess the contribution of stylization or rendering to overall image quality.

Recently, a growing research trend has emerged to evaluate the quality of AI-generated images, addressing unique challenges such as perceptual realism, semantic consistency, and technical artifacts. Datasets like HPDv2²², AGIQA-3k²³, AIGIQA-20k²⁴, and PKU-I2IQA²⁵ have been developed to assess these aspects. These datasets typically employ a structured methodology: multiple images are generated for each prompt, and then their quality, authenticity, and prompt alignment are rigorously validated using mean opinion scores. This approach provides a standardized framework to assess the rapidly evolving capabilities of generative AI models.

Several tone-mapped image quality datasets have as well been introduced^{26,27}, in which multiple tone mapping operators (TMOs) are applied to high dynamic range (HDR) images to produce low dynamic range (LDR) outputs of varying perceptual quality (we use the term HDR in the traditional sense of a high-range image destined for reduction to SDR — Standard Dynamic Range — rather than the newer display technology that presents actual HDR to the viewer). While TMOs do not introduce degradations in the traditional sense, making these datasets somewhat aligned with the subjective nature of our work, they differ fundamentally in scope and application. Specifically, the input and output images in such datasets belong to different dynamic ranges, making the transformations more technical than stylistic. Moreover, these datasets do not address or address minimally the broader and more common use cases involving stylistic and rendering variation within the same dynamic range, such as color grading, contrast adjustments, or photographic render emulation, which are the core of IRQA and the focus of this work.

The UHD-IQA Benchmark Database³ contains 6073 ultra-high-definition (4K) images collected from Pixabay and curated to exclude synthetic or heavily retouched content. Image quality was assessed in a single-stimulus protocol, where 10 expert raters composed of photographers and graphic artists judged each image twice on a continuous scale from *bad* (1%) to *excellent* (100%), yielding around 20 ratings per image. The resulting mean opinion scores provide ground-truth annotations, with high inter-round consistency measured by the Spearman Rank Order Correlation Coefficient (SRCC), reported at approximately 0.93. Along with perceptual quality ratings, the dataset includes metadata such as image dimensions, popularity indicators, and machine-generated tags, making it one of the largest and most comprehensive publicly available benchmarks for no-reference image quality assessment in the UHD domain. However, the dataset contains only a single rendering per image, limiting its usefulness for evaluating different renderings of the same scene; moreover, the absence of pairwise comparisons further limits its ability to capture fine-grained perceptual differences.

Overall, none of the aforementioned works explicitly addresses IRQA. While existing approaches focus on quantifying technical distortions — such as noise, blur, or compression artifacts — that degrade generic image quality, they largely overlook the perceptual impact of artistic render, see Table 1. The goal of this paper is exactly to fill this niche.

Benchmark NR-IQA methods

For IQA tasks, numerous methodological approaches exist. The field of IQA is dominated by two primary approaches: full-reference and no-reference methods. Full-reference techniques require a complete, high-fidelity reference image for comparison to quantify perceptual degradation. Conversely, no-reference methods operate without any reference image, directly estimating the quality score based solely on the features of the test image. In our work, since we propose a dataset without reference images (i.e., “gold standard” for comparison exists), traditional full-reference metrics like SSIM²⁸, PSNR, or LPIPS²⁹ cannot be directly applied. Consequently, our analysis focuses exclusively on NR-IQA methods.

NR-IQA approaches can be broadly categorized into two paradigms:

Dataset	# Scenes	# Edited Images	IQA Task	# Voters	Method
LIVE ⁶	29	779	Distortion	29	SS
CSIQ ⁹	30	866	Distortion	25	N/A
VCL@FER ⁷	23	552	Distortion	118	SS
TID2013 ¹⁰	25	3,000	Distortion	971	PC
PDAP-HDDS ¹²	250	12,000	Distortion	38	ACR
KonIQ-10k ¹⁶	10,073	0	Distortion, Content	1,459	ACR
KADID-10k ¹³	81	10,125	Distortion	2,209	ACR
UHD-IQA ³	6073	6073	Content	10	SS
REPID	5000	30,000	IRQA	13,648	PC

Table 1. Dataset summary for most relevant IQA datasets compared to our IRQA dataset. In this table in the method column ACR stands for 5-point Absolute Category Rating, SS for Single Stimulus, and PC for pairwise comparison.

- **Statistical** methods, that employ handcrafted heuristics to detect specific artifacts (e.g., noise, blur, compression artifacts) based on measurable image statistics. These methods typically rely on domain knowledge about common distortion types and their observable effects on image properties.
- **Deep learning** (DL) methods, that utilize data-driven feature extraction to automatically learn complex quality-relevant patterns. Unlike statistical approaches, DL models can adapt to image content and capture more sophisticated quality attributes through hierarchical feature learning. This enables them to handle diverse distortion types and content variations more effectively.

One of the most widely used statistical approaches for NR-IQA is BRISQUE³⁰. This method extracts a set of handcrafted features based on natural scene statistics and employs a Support Vector Regressor (SVR)³¹ to predict the final quality score. More recently, PIQUE³² has adopted a similar paradigm, leveraging perceptually relevant features in conjunction with a statistical learning framework to estimate image quality.

BLIINDS-II³³ similarly operates within the natural scene statistic (NSS) paradigm but shifts feature extraction to the frequency domain, constructing a generalized statistical model of local discrete cosine transform (DCT) coefficients. The parameters of this model are used within a Bayesian inference framework to predict perceptual quality scores, requiring only minimal training. Evaluated on the LIVE IQA database, BLIINDS-II achieves correlation with human visual judgements at a level competitive with the full-reference SSIM index.

SSEQ³⁴ departs from both spatial and DCT-domain modelling by extracting local spatial and spectral entropy features, deploying a two-stage SVM-based framework that first classifies the distortion type before estimating quality. The resulting index is capable of assessing image quality across multiple distortion categories and has been shown to outperform several leading NR-IQA methods, including BIQI, DIIVINE, and BLIINDS-II, as well as the full-reference SSIM, while maintaining comparatively low computational complexity.

NIQE³⁵ represents a distinct class of NR-IQA method in that it operates entirely without exposure to distorted images or human opinion scores, relying solely on measurable deviations from the statistical regularities of natural images. Quality is expressed as a distance metric between multivariate Gaussian models fitted to NSS features extracted from the test image and from a corpus of pristine natural images.

With the emergence of large-scale IQA datasets, DL-based methods have gained significant traction. Early works such as OG-IQA³⁶, DeepBIQ³⁷, MUSIQ³⁸, PaQ-2-PiQ³⁹, CONTRIQUE⁴⁰, Re-IQA⁴¹, and ARNIQA⁴² explore various deep learning paradigms, including patch-based learning, contrastive representation learning, and transformer-based architectures, to improve the accuracy and generalizability of IQA models.

A blind NR-IQA method for HDR content was proposed by Wan et al.⁴³, combining multi-scale Retinex-based perception features with deep features extracted from a pretrained VGG16 network, and using SVR for final quality prediction.

Furthermore, a growing number of recent approaches incorporates vision-language models for image quality assessment. For instance, CLIP-IQA⁴⁴ estimates image quality by computing the cosine similarity between visual embeddings and text prompts such as “This photo is good” and “This photo is bad.” Building upon this idea, LIQE¹ improves the framework by fine-tuning both visual and textual encoders and introducing a multi-task optimization strategy, which enhances robustness, particularly due to its use of large-scale training data. Subsequently, QualiCLIP⁴⁵ was designed to be more sensitive to varying levels of image distortion. Most recently, LAR-IQA⁴⁶ demonstrated that incorporating KANs⁴⁷ can further boost model performance.

More recently, RAM-VQA⁴⁸ has extended the vision-language paradigm to the video domain, proposing a Restoration-Assisted Multi-modality framework that leverages video restoration as a proxy task to explicitly model distortion-sensitive features, which are subsequently fused with multi-modal quality representations to improve predictive robustness. A separate line of work addresses the specific challenges posed by omnidirectional video, where the scarcity of annotated training samples presents a fundamental obstacle to learning adequate distortion representations. DS-MAE⁴⁹ tackles this by adapting the masked autoencoder paradigm to the omnidirectional video quality assessment (OVQA) problem. Rather than masking arbitrary patches as in standard MAE, DS-MAE replaces distorted regions with their undistorted counterparts and trains the autoencoder to reconstruct the original distortions, directly imbuing the learned representations with sensitivity to perceptual degradations. A

knowledge replay module is additionally employed to mitigate catastrophic forgetting of content representations during training. The resulting OVQA model, OmniVQA, is evaluated on three public OVQA benchmarks and achieves competitive performance against existing methods.

Complementing these task-specific approaches, a recent survey on LMM evaluation⁵⁰ provides a structured overview of the field, organising evaluation along two main axes: multimodal understanding, divided into general and specialised capabilities, and multimodal generation, organised by output modality across image, video, audio, and 3D content. Notably, low-level visual quality assessment — encompassing IQA and related perceptual tasks — is explicitly situated within the specialised capabilities axis, contextualising the rapid proliferation of LMM-based quality assessment methods and clarifying the broader evaluation landscape within which they operate.

Taking a further step toward unification, X-QA⁵¹ proposes a versatile multimedia quality assessment framework for visual communications that consolidates the evaluation of images, videos, and 3D models within a single large multimodal model, enabling simultaneous cross-modal quality prediction while harnessing shared perceptual knowledge across media types.

Overall, none of the aforementioned works explicitly addresses IRQA. While existing approaches span a broad spectrum — from handcrafted natural scene statistics and entropy-based features to deep learning models and vision-language frameworks — they are uniformly designed to quantify technical distortions such as noise, blur, compression artifacts, and perceptual degradations of generic photographic content. The aesthetic and perceptual dimensions specific to rendered images, including the tonal, chromatic, and stylistic qualities examined in this work, remain largely unaddressed by the existing literature.

Proposed REPID

The MIT-Adobe FiveK dataset⁵ is a widely-used, high-resolution image dataset designed for research in computational photography and image enhancement. It comprises 5000 RAW photographs captured by DSLR cameras under diverse lighting and scene conditions. Each image has been independently rendered by five professional photo renderers, labeled Expert A through E, yielding five stylistically distinct reference renders per image. The dataset includes both the original RAW files and their corresponding JPEG renderings, making it suitable for tasks such as supervised learning of image-to-image translation, automatic photo retouching, tone and exposure correction, white balance adjustment, render-specific enhancement, etc. Its diversity in photographic content, expert renderers, and high-resolution format with up to 5616×3744 pixels makes it a valuable resource for training and evaluating deep learning models in both paired and unpaired settings.

The distribution of labels for these images that is available within the MIT-Adobe FiveK dataset is shown in Fig. 2, together with another distribution obtained by means of querying the BLIP⁵². In order to validate the overall quality of BLIP results we randomly selected 100 images from the dataset and marked them up. The average accuracy for all categories equals 0.96, while F1-score equals 0.86. More detailed analysis of category markup can be seen in the [Supplementary material](#). The variety of the dataset is sufficient to draw general conclusions.

For the selection of renderings, we considered two alternative strategies that fit in research budget: (1) collecting DMOS annotations for all scenes in the MIT-Adobe FiveK dataset under a limited set of rendering variations, or (2) annotating a dense grid of rendering parameters for a restricted subset of scenes. We ultimately adopted the first strategy, focusing on the most probable visually pleasing renderings — namely, the photographer-provided renderings available in the MIT-Adobe FiveK dataset. This choice allows us to leverage expert-curated outputs while maintaining broad scene coverage. In addition, we provide supplementary analysis on a subset of 19 scenes rendered using 36 rendering configurations, see section 8.5. Furthermore, we outline guidelines for future work exploring a wider range of rendering strategies, see Section 8.4.

Each of the MIT-Adobe FiveK images is rendered in six different renders: five photographers, i.e. renderers and one neutral style from Adobe Photoshop *auto* settings, which yields $\binom{6}{2} = 15$ possible unique style pair combinations per image. To perform subjective quality assessment, these pairs were uploaded to the crowdsourcing platform Yandex.Tasks. The Yandex.Tasks interface displayed six image pairs per page. For each pair, the participants, i.e., evaluators were asked: “Which image is preferable?”. Images were presented against a 50% gray background, following widely used recommendations²¹ and standard to minimize bias in color perception by using a neutral background. To further ensure unbiased assessment, no additional controls or information were placed between the image pairs. Each image pair was also spaced adequately to simplify the visual comparison. An example of such an image pair can be seen in Fig. 3.

The evaluators could select one of the following options for each image pair: “The left image is preferable”, “The right image is preferable” and “Both images are equally preferable”. To identify unreliable evaluators, each page included an additional pair of identical images as a control question. If such outlier votes were detected, the corresponding votes were excluded from the pool. Essentially, the whole voting process was similar to the one in⁴.

The dataset consists of 5000 unique scenes captured by professional photographers. For each scene, five rendered versions were produced by photography students, along with one automatically generated version using Adobe Photoshop’s *auto*-mode. For REPID, each pair of scene versions was assessed by 25 participants, who provided one of three possible responses: a preference for one image, a preference for the other, or abstention. The choice of 25 evaluators per pair is grounded in prior work⁴, which showed that for crowdsourced pairwise aesthetic evaluation of image rendering styles, the correct rate — defined as the agreement between results obtained by a given number of participants and those obtained by all participants, simulated via bootstrapping — reaches approximately 0.85 at 25 voters per pair, while increasing to 100 or 400 voters yields only marginal improvements to approximately 0.91 and 0.95, respectively. This diminishing return confirms that 25 evaluators

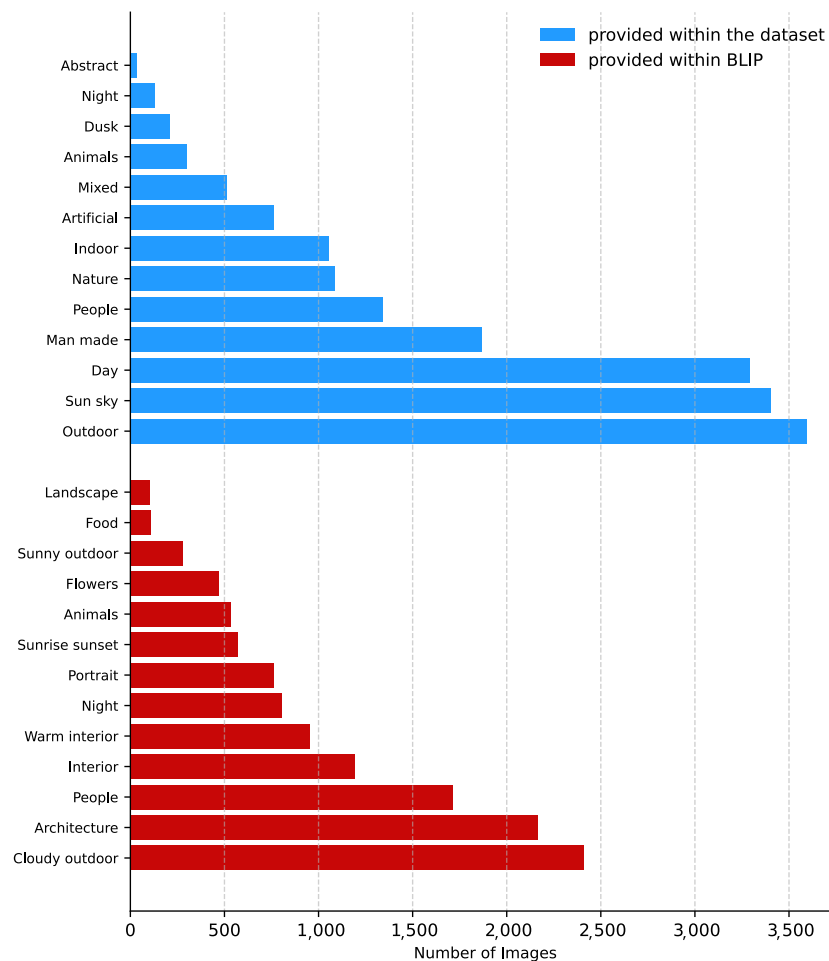


Fig. 2. The distribution of categories for the selected MIT-Adobe FiveK dataset images based on the information provided within the dataset and BLIP; one image can have multiple categories.

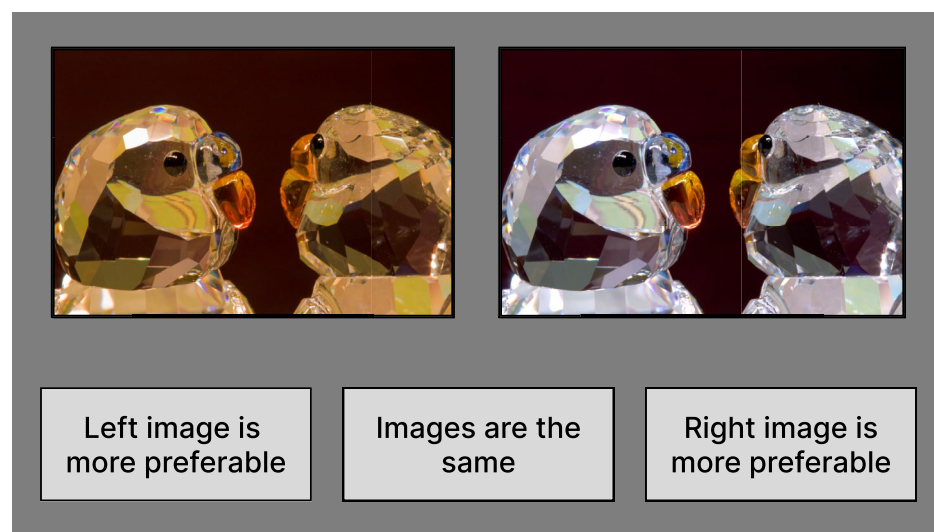


Fig. 3. An example of a pair of aesthetic renderings for the same scene as shown in Yandex.Tasks.

provide a practical trade-off between annotation cost and reliability. An illustration of the dataset is shown in Fig. 4.

Overall, 13,648 people annotated the dataset, completing 2.5 million tasks, including those in the test procedure. The evaluators were recruited through the Yandex.Tasks crowdsourcing platform without imposing constraints on professional background, nationality, or education level. This deliberate design choice ensures that the collected preferences reflect a broad and demographically diverse population rather than a narrow expert cohort, which is appropriate given that rendering preferences are inherently subjective and vary across individuals. The large total number of unique evaluators further mitigates the risk of systematic bias from any single demographic group. In particular, the distribution of votes follows a power law: the most active 20% of users generated 60% of the data, while the top 30% generated 70%. While this has no direct and immediate consequence on the quality of the dataset, it is still interesting to mention. For a detailed view, see Fig. 5a.

The distribution of votes is shown in Fig. 5d. It can be seen that vote “Both images are equally preferable” reminds of the exponential distribution. Introduction of the third option leads to the shift in produced distributions, see Fig. 5b. Due to even numbers being more probable here, the probability of total number of votes that determines quality (“The left image is preferable” and “The right image is preferable”) is more probable to be odd, since subtracting an even number from 25 results in an odd number. Therefore, the distribution becomes jagged and may lead to worse results when used as ground truth to the models. To deal with this issue and smooth the distribution, see Fig. 5c, we processed scores using Equation 1:

$$DMOS = L - R + B \bmod 2, \quad (1)$$

where L , R , B are the numbers of votes for left, right, and both images, respectively. It was further normalized to bring it to the range (0, 1) using Equation 2, which will be interpreted in our experiments as the probability that the left rendering is better than the right.

$$PMOS = \frac{DMOS}{2(L + R + B)} + \frac{1}{2} = \frac{2L + B + B \bmod 2}{2(L + R + B)}. \quad (2)$$

To assess the reliability and potential ceiling of model performance, we estimated the maximum achievable accuracy through bootstrapping. At each iteration, votes were randomly drawn from the original annotation pool to construct a new dataset, subsequently binarized into preferred render labels. The agreement between the resampled dataset and the original annotations was then computed as accuracy and Spearman correlation. Repeating this process 1000 times produced the histogram shown in Fig. 6. The distribution is centered at a mean accuracy of 0.896 with a confidence interval of 0.004, and a mean Spearman correlation of 0.919 with a confidence interval of 0.005. These results indicate that even a model capable of perfectly reproducing the preferences of all 25 evaluators would not be expected to exceed this performance level. This values therefore represent a practical upper bound for predictive performance on the IRQA task.

Possible tasks

Our dataset enables a wide range of research directions related to aesthetic perception and photographic render evaluation. Below, we outline several potential tasks:



Fig. 4. Illustration of the proposed REPID dataset. For each scene in MIT-Adobe FiveK there are 6 different renders: 5 styles from renderers and 1 for Adobe neutral style. For each pair in scene there is a total of 25 votes subdivided between left image, right image, and equal.

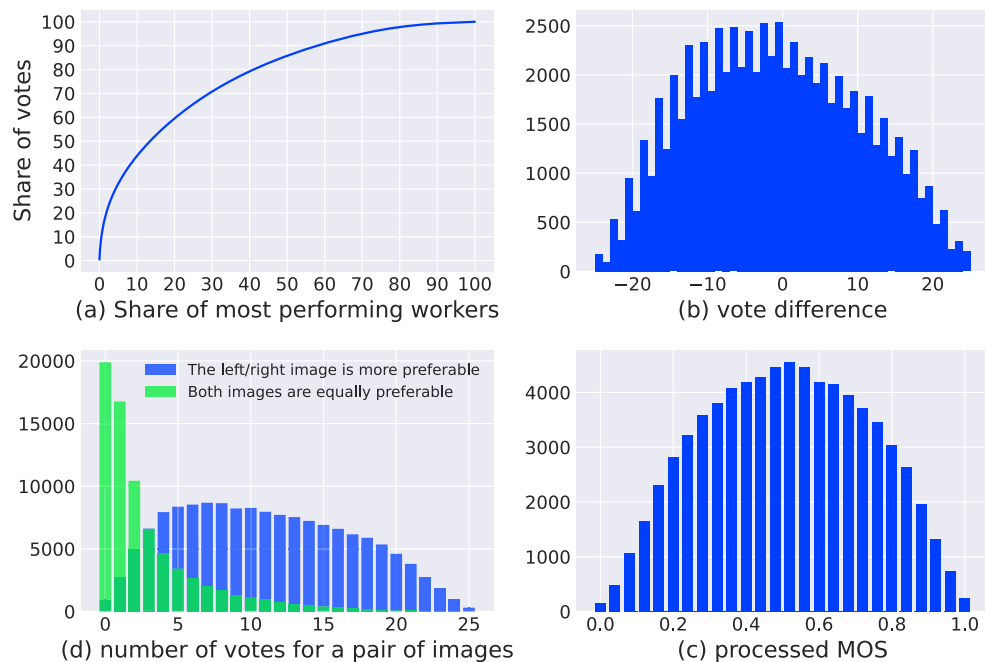


Fig. 5. Analysis of REPID: **a** Distribution of annotation workload. **b** The vote difference distribution. Vote difference is obtained by subtracting votes for left and right image, while votes that claim images look the same are not used. The distribution looks inconsistent due to distribution of votes for “Both images are equally preferable”. **c** PMOS distribution. PMOS is calculated via Equation 2 that smooths the final distribution. **d** The distribution of votes per each image pair in experiment. Votes for “Both images are equally preferable” are not distributed similar to other types. Instead, this category gives bias, making processing more complex.

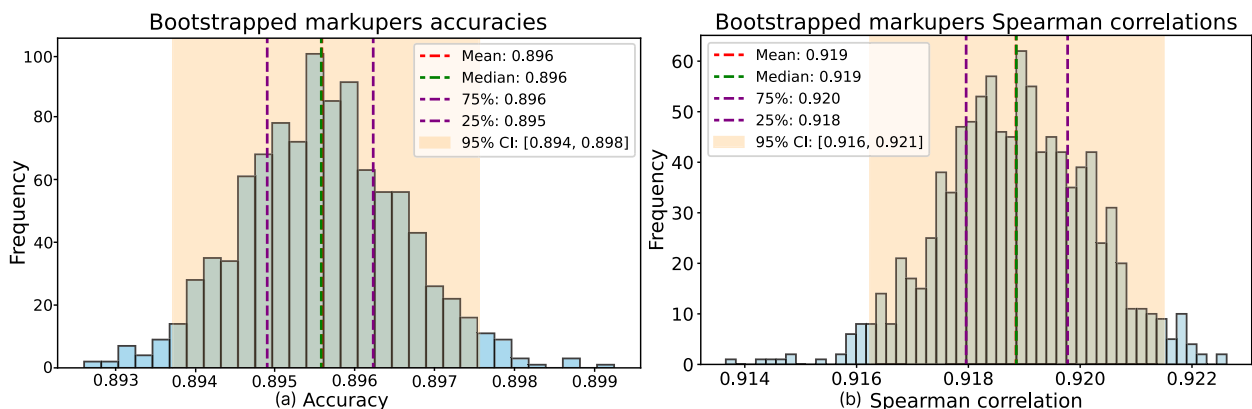


Fig. 6. Histograms of maximum **a** accuracy and **b** Spearman correlation estimation across 1000 resampled datasets. The distribution illustrates the upper bound of performance achievable by a model that perfectly replicates the voting behavior of 25 human evaluators.

1. *Analysis of Human Aesthetic Biases*: Investigate how render preferences vary across different image categories (e.g., landscapes, portraits, food), color distributions, and compositional patterns.
2. *Rendering Preference Prediction*: Train a model to predict which of two retouched versions of an image users would prefer, given only the image content and/or render features.
3. *Style Ranking*: Predict a ranking of user preferences for multiple renders applied to the same original image, instead of just pairwise preference prediction.
4. *Personalized Rendering Preference Prediction*: Leverage evaluator identifiers to train models that predict the preferred render of a specific user. This task allows studying personalization and modeling individual aesthetic profiles.
5. *Style Generation*: Explore generative models capable of producing new renderings that maximize predicted user preference.

6. **Aesthetics Evaluation Benchmarking:** Use the dataset to evaluate and benchmark aesthetics evaluation models, render evaluation methods, and fine-tuned foundation models (e.g., CLIP). This is closely aligned with the current work presented in this paper.

Our dataset consists of 5000 images spanning a diverse set of vernacular subjects. No explicit selection criteria were provided for image collection. Each image was rendered by five knowledgeable photography students, i.e., renderers using Adobe Lightroom in 2011, though no information is available regarding their relative skill levels or aesthetic preferences.

These characteristics naturally lead to two complementary areas of inquiry: (1) understanding viewer preferences and the factors that drive them, and (2) analyzing the original rendering process and its influence on perceived quality.

In the following sections, we focus on four key tasks: *viewer preference analysis*, *rendering preference prediction*, *personalized rendering preference prediction*, and *rendering evaluation benchmarking*.

Analysis of preferences

The obtained data gives us an opportunity to analyze preferences of evaluators for different categories of scenes and different renders, see Fig. 7.

A strong positive correlation was observed between the mean number of parameters changed by the renderer (i.e., Expert in FiveK terminology) and their win probability. Renderer E, who applied the highest mean number of alterations (9.81 ± 1.79), achieved the highest win probability (57.9%).

However, this relationship is significantly moderated by the efficiency and quality of the edits. Notably, renderer D demonstrated a highly efficient strategy, securing a strong win probability (52.1%) with the lowest mean number of changes (6.22 ± 1.52). Furthermore, comparing renderers B and D reveals that a relatively small difference in the number of changes (from 6.58 to 6.22) corresponded to a substantial jump in performance (from 43.7 to 52.1%). This discrepancy underscores that the nature and precision of parameter modifications are more critical to success than the mere volume of edits.

The low and similar standard deviations across all renderers indicate that the application of each strategy was consistent and reproducible.

For our dataset we calculated the importance of settings for different categories to verify that the preferred render depends on content, see Fig. 8. We selected the most promising settings based on the advice of professional photographer Michael Freeman and trained a linear regression for each setting. Note that the Parametric group represents an alternative formulation of brightness and contrast adjustments implemented via Curves. Vibrance can be interpreted as a constrained version of Saturation, designed to preserve perceptual balance. Tint corresponds to the color axis orthogonal to color Temperature. Additionally, Shadows and Highlight Recovery act as local TMOs, selectively modifying brightness and contrast at the extremities of the tonal range. Consequently, the fundamental rendering dimensions can be reduced to Brightness (Exposure), Contrast, Color Temperature, and Saturation. As a measure of importance we used accuracy with PMOS.

For better visibility, we calculated maximum and minimum value among all categories and subtracted minimum and divided by maximum. From the obtained charts we can clearly see that different scene types are described with preferred setting types differently.

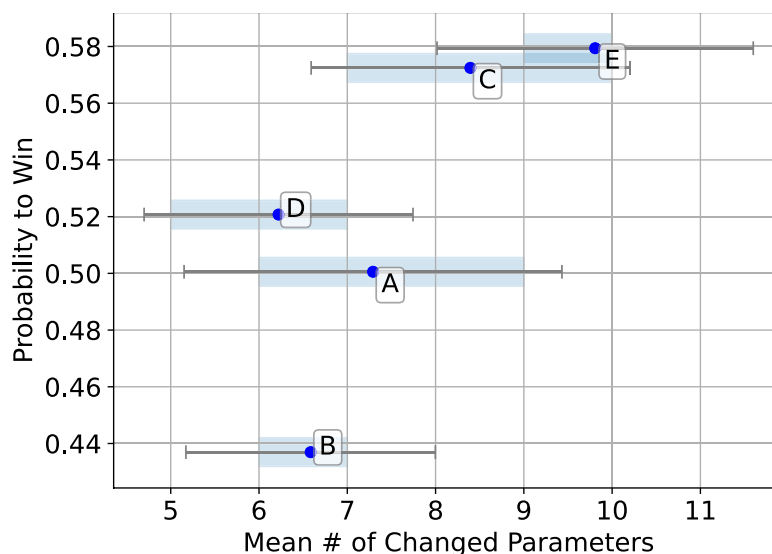


Fig. 7. Analysis of used settings impact of photographers (A,B,C,D,E) to perceived quality. Gray line represent SD, blue dot is the mean and the blue region displays 0.25 to 0.75 quantile. Pearson correlation is 0.77.



Fig. 8. Radar charts of main settings importance in Adobe Lightroom for different content groups. There is a visible difference which can show that preferred render depends on content. Blue line is mean value and blue area is 0.25 and 0.75 quartile.

Rendering preferences prediction

There are numerous existing methods addressing the IQA task. However, none of them were specifically designed for IRQA. In this section, we analyze state-of-the-art IQA models and train new models tailored for IRQA on the proposed REPID.

We've selected several of the most promising NR-IQA methods representing both deep learning-based and statistical approaches. Since NR-IQA models operate on a single image at a time, and our dataset is first computed quality scores for each image independently and then trained a linear regression model on these scores.

For evaluation, we allocated approximately 25% of the images to the test set. The test set comprised $n = 1,283$ scenes, resulting in $N = 1,283 \times 15 = 19,245$ image pairs. Some images caused invalid preprocessing issues, thus we filtered out 250 image pairs, resulting in 18,995 image pairs.

We adopt the following notation: the PMOS value for an image pair i and j within scene s is denoted as p_{ij}^s , with the corresponding predicted score given by $\hat{p}_{ij}^s$. The aggregated pairwise score obtained via the Bradley-Terry procedure⁵³ is denoted as q_i^s , with its predicted counterpart $\hat{q}_i^s$.

We then evaluated performance using the following metrics:

- *Accuracy:*

$$\frac{1}{N} \sum_{s,i,j,i \neq j} \mathbf{1}_{p_{ij}^s > 0.5} (\hat{p}_{ij}^s > 0.5) + \mathbf{1}_{p_{ij}^s \leq 0.5} (\hat{p}_{ij}^s \leq 0.5) \quad (3)$$

- *Top-1 accuracy:*

$$\frac{1}{n} \sum_s \mathbf{1}_{\arg\max_i(q_i^s) = \arg\max_i(\hat{q}_i^s)} \quad (4)$$

- *Spearman correlation:*

$$\frac{1}{n} \sum_s \rho[R_i[q_i^s], R_i[\hat{q}_i^s]], \quad (5)$$

where $R_i[\cdot]$ is ranking operation for list of scores for i index.

- *Kendall correlation:*

$$\frac{1}{n} \sum_s \tau[R_i[q_i^s], R_i[\hat{q}_i^s]] \quad (6)$$

Aesthetic preference prediction

In the first block in Table 2 there are results for evaluating total variation (TV) and pretrained models: BRISQUE³⁰, PIQE³², ARNIQA⁴², CLIP IQA⁴⁴, LARIQA⁴⁶ and LIQE¹. It can be seen that simple TV performs comparable with foundation-based models like LIQE. This means that current IQA models are not applicable to an IRQA task without finetuning these data.

For the MIT-Adobe FiveK dataset, metadata describing the applied rendering settings (e.g., saturation, contrast, color temperature) is available. This allows for score prediction based solely on these settings. Additionally, we extracted several handcrafted features directly from the images, including standard deviation, lightness, saturation, hue, as well as u and v components from the UV color space. All features were normalized to the [0, 1] range. Moreover, due to similarity of the TV method to standard deviation (SD), we used linear regression for SD only. The results for this experiment are shown in the second block of Table 2. It is worth noting that this simple combination of image settings and handcrafted features outperforms all conventional IQA metrics. Even using only SD as a feature yields a better result than most existing IQA methods.

Finally, we trained several models specifically for the score prediction task. Because we work with image pairs, we explored two strategies: the Siamese approach (“model_{siam}”), i.e., forwarding both images independently through the base model and combining the outputs with a linear regression, and ConCat approach (“model_{cc}”), i.e., concatenating the two images along the channel dimension (resulting in a tensor with a depth dimension equal to 6), and directly predicting the score. It is worth noting that for the CLIP model we freeze the vision transformer (ViT) weights training linear regression only. The results are shown in the third block of Table 2. Our experiments reveal a consistent advantage of ConCat approach over the Siamese one across most architectures, with ConvNeXtV2 as a notable exception. The next paragraph provides a brief discussion of the possible reasons underlying this phenomenon.

	Model	Accuracy ↑	Spearman ↑	Kendall ↑	Top 1 accuracy ↑
Pretrained	TV	0.73	0.56	0.47	0.50
	BRISQUE	0.64	0.35	0.29	0.35
	PIQE	0.58	0.20	0.17	0.27
	ARNIQA	0.65	0.39	0.32	0.38
	CLIP IQA	0.68	0.45	0.38	0.43
	LARIQA	0.63	0.33	0.27	0.32
	LIQE	0.73	0.56	0.48	0.48
Feature based	Linear Regression	0.76	0.62	0.53	0.52
	Random Forest	0.74	0.60	0.51	0.50
	SD Linear Regression	0.71	0.49	0.42	0.44
Image based	Resnet152 _{cc}	0.81	0.74	0.64	0.57
	Resnet50 _{cc}	0.81	0.74	0.64	0.57
	MobileNetV3(L) _{cc}	0.81	0.75	0.65	0.58
	MobileNetV2 _{cc}	0.80	0.73	0.62	0.56
	ConvNeXtV2 _{cc}	0.82	0.78	0.68	0.61
	Resnet152 _{siam}	0.79	0.71	0.61	0.53
	Resnet50 _{siam}	0.79	0.70	0.60	0.52
	MobileNetV3(L) _{siam}	0.79	0.71	0.61	0.55
	MobileNetV2 _{siam}	0.78	0.68	0.58	0.53
	ConvNeXtV2 _{siam}	0.83	0.79	0.69	0.61
	CLIP ViT-B/32 _{siam}	0.78	0.69	0.58	0.53
Image Feature based	Resnet152 _{cc}	0.80	0.74	0.64	0.57
	MobileNetV3(L) _{cc}	0.81	0.75	0.64	0.56
	Resnet152 _{siam}	0.79	0.71	0.61	0.54
	MobileNetV3(L) _{siam}	0.80	0.72	0.61	0.55

Table 2. Evaluated metrics for different methods over the pairwise comparison task. The **best-performing** models for each metric **in bold** is highlighted in **bold**. The *best-performing* and *second-best* models for each metric are highlighted in bold and italic and italic, respectively.

The superiority of CC stems from early fusion: by processing a 6-channel tensor, the model captures low-level cross-image differences from the first convolutional layer. The Siamese paradigm, by contrast, enforces a late-fusion bottleneck where comparison is deferred to the embedding level — a constraint that disproportionately affects lower-capacity architectures such as MobileNetV2, whose embeddings lack the capacity to encode both per-image content and pairwise difference signals simultaneously. As for exception, ConvNeXtV2 incorporates Global Response Normalization (GRN), which performs channel-wise normalization over the global feature map. Under CC, the 6-channel input deviates from the 3-channel distribution the model was pretrained on, and we hypothesize that GRN introduces adverse gradient dynamics during fine-tuning by normalizing over a channel space it was not originally optimized for. The Siamese variant preserves the expected input distribution, allowing GRN to operate within its trained regime. Combined with ConvNeXtV2's high representational capacity — which makes late-fusion embeddings sufficiently rich for comparison without early fusion — this explains its systematic deviation from the general trend.

We further enhanced these models by incorporating metadata and the handcrafted features from the previous experiment; this extended setup is presented in the fourth block of Table 2. For the CLIP model we used pretrained CLIP ViT-B/32 to generate the embeddings. Using these embeddings and the extracted features, we trained a linear regression model, where we concatenated CLIP embedding and extracted features. *Note:* this model has a significant downside, as it implies that the best render for a given scene consists of min-max slider values.

Despite the additional information, models with additional information yield even worse results than the image-based only approaches. Possible reasons for this performance degradation include the fact that the relationship between slider parameters and perceived image quality is complex and highly non-linear, making it difficult to capture with a linear formulation. Furthermore, different combinations of slider values may lead to visually similar renderings, introducing ambiguity and reducing the identifiability of optimal parameter configurations. A promising direction for further investigation is the use of large language models and vision-language models, which could better capture high-level semantic context and provide more expressive, non-linear mappings between scene content, editing parameters, and perceived quality. Such models may also enable reasoning about editing intent and user preference in a more structured manner, potentially overcoming the limitations of purely parametric regression approaches.

A range of illustration demonstrating quality evaluation for diverse rendering can be found in [Supplementary Materials](#). Detailed technical information on training and testing procedures is available in project GitHub repository.

The rejection curve (see Fig. 9) shows that accuracy is not an optimal metric for IRQA, as it implicitly assumes binary labels. In our case, pairwise PMOS values are concentrated near 0.5, for example 36% of all answers are falling in range 0.4...0.6. As a result, accuracy increases almost linearly as uncertain samples are rejected.

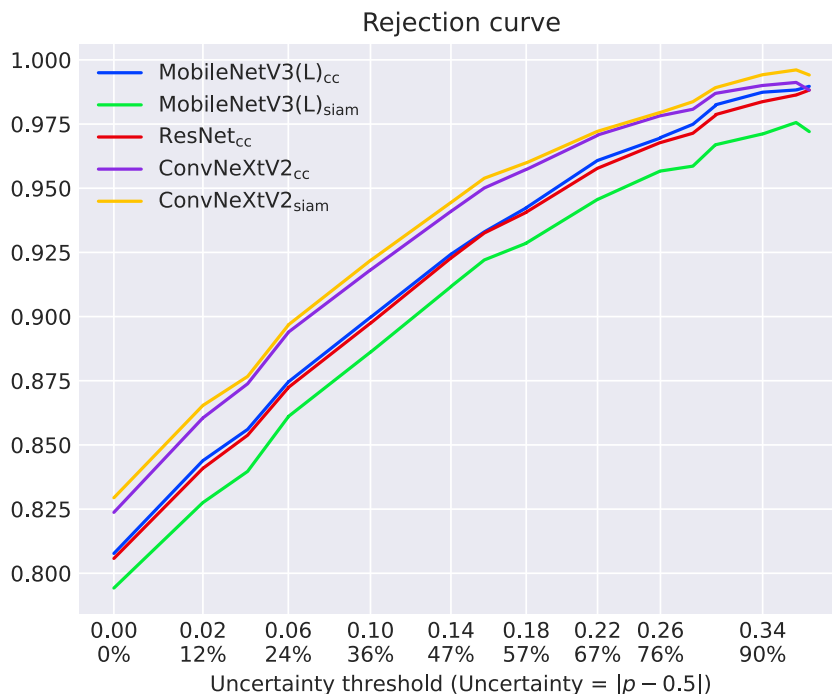


Fig. 9. This rejection curve shows the dependence of accuracy on the certainty of the PMOS p_{ij}^s . Data points are aligned by the proportion of rejected samples, which is also indicated by the tick labels on the x-axis.

	Model	Accuracy ↑	Spearman ↑	Kendall ↑	Top 1 accuracy ↑
Pretrained	TV	0.67	0.35	0.35	0.58
	BRISQUE	0.68	0.36	0.36	0.58
	PIQE	0.67	0.34	0.33	0.57
	ARNIQA	0.66	0.32	0.31	0.56
	CLIP IQA	0.66	0.32	0.31	0.56
	LARIQA	0.66	0.32	0.31	0.56
	LIQE	0.68	0.36	0.35	0.57
Trained	SD Linear Regression	0.67	0.36	0.36	0.58
	Linear Regression	0.71	0.43	0.39	0.59
	Random Forest	0.70	0.38	0.35	0.56
	CLIP ViT-B/32 _{siam}	0.70	0.26	0.24	0.29

Table 3. Evaluated metrics for different methods obtained on the pairwise comparison task for the personalized IRQA for the “first” most active voter. Pretrained block was used without training on labels from voter, trained one consists of models that were trained on votes. The **best-performing** model for each metric is highlighted in **bold**.

	Model	Accuracy ↑	Spearman ↑	Kendall ↑	Top 1 accuracy ↑
Pretrained	TV	0.52	0.01	0.01	0.50
	BRISQUE	0.49	– 0.04	– 0.04	0.46
	PIQE	0.48	0.00	0.01	0.47
	ARNIQA	0.51	– 0.03	– 0.03	0.44
	CLIP IQA	0.52	– 0.05	– 0.06	0.42
	LARIQA	0.49	– 0.04	– 0.04	0.47
	LIQE	0.48	– 0.08	– 0.08	0.43
Trained	SD Linear Regression	0.87	0.78	0.77	0.86
	Linear Regression	0.87	0.76	0.71	0.87
	Random Forest	0.87	0.75	0.70	0.86
	CLIP ViT-B/32 _{siam}	0.88	0.53	0.52	0.87

Table 4. Evaluated metrics for different methods obtained on the pairwise comparison task for the personalized IRQA for the “second” most active voter. Pretrained block was used without training on labels from voter, trained one consists of models that were trained on votes. The **best-performing** model for each metric is highlighted in **bold**.

Personalized IRQA

In the previous section, we analyzed the aggregated preferences of all voters. In this experiment, we focus on analyzing and predicting particular voters. We observed significant variability between voters, with the mean standard deviation between individual votes and the aggregated PMOS reaching 0.41 ± 0.045 . Given that the maximum possible standard deviation is 0.5 for our data, these results indicate that individual preferences often diverge sharply from the average, complicating the task of personalized modeling. To the first individual voter provided 8, 540 votes, covering 2338 scenes in which he/she voted on any available pair of renders at least once. Given the relatively small size of this subset, we restricted our evaluation to simple models based on handcrafted image features and CLIP embeddings. The results are reported in Table 3. It can be seen that individual prediction performs worse than similar methods in Table 2 except for top-1 accuracy. This may be the result of a well-known issue, where individuals are confident about the best and the worst examples and are indifferent to anything in between^{54,55}.

The second most promising voter marked up 6, 170 tasks, covering 1, 962 scenes. The results for models predicting him can be found in Table 4.

The results for these two evaluators differ substantially. This discrepancy can likely be explained by their voting behavior: the first evaluator selected “Both images are equally preferable” in 41% of cases, whereas the second did so in 79% of cases, indicating a significantly higher tendency toward indifference. The overall mean standard deviation between the aggregated votes and an individual evaluator is 0.41. These findings suggest that sensitivity to rendering differences varies considerably across individuals.

Given that training machine learning models requires large-scale annotated data, relying on a single annotator is impractical for datasets comprising tens of thousands of samples. To address this, we suggest the use of active learning strategies. One promising approach is to prioritize samples based on inter-annotator disagreement—for example, by measuring the standard deviation of votes across evaluators for each scene – and selectively sending

Dataset	model							
	TV	BRISQUE	PIQE	ARNIQA	CLIP-IQA	LIQE	LARIQA	IRQAM
VCLFER	0.14	0.90	0.79	0.90	0.56	0.90	0.66	0.72
LIVE	0.30	0.96	0.85	0.91	0.73	0.92	0.72	0.79
CSIQ	0.27	0.55	0.54	0.89	0.87	0.93	0.61	0.56
TID2013	0.02	0.39	0.36	0.75	0.61	0.74	0.38	-0.25
KADID-10k	0.12	0.35	0.25	0.93	0.56	0.79	0.49	0.41
PDAP-HDDS	0.16	0.41	0.32	0.64	0.42	0.62	0.46	0.46
REPID	0.61	0.34	0.07	0.43	0.53	0.49	0.29	0.79

Table 5. Spearman correlation for different models on different datasets. IRQAM is the best model, trained on REPID. The **best-performing** model for each dataset is highlighted in **bold**.

only “non-trivial” (i.e., high-uncertainty) samples for annotation. This strategy can significantly reduce labeling effort while preserving informative training data.

Comparison of distortion-based IQA on an IRQA task

A final experiment was conducted to compare IQA tasks for distortions with IRQA. The key idea was to evaluate a set IQA models and benchmark datasets, and then cross-validate their performance with REPID and our proposed Image Rendering Quality Assessment Model (IRQAM).

For this comparison, we selected MobileNetV3(L)_{cc} as the representative implementation of IRQAM, since it demonstrated the strongest performance across most metrics on the REPID benchmark.

Because our model operates on image pairs, we adapted existing IQA datasets by using distorted–reference pairs as inputs. As the evaluation metric, we computed the Spearman correlation between predicted scores and subjective ratings for all images associated with the same reference, and then averaged these correlations across reference scenes. The results of this experiment are summarized in Table 5.

It can be observed that all existing models, with the exception of IRQAM, achieve relatively low correlations on the REPID benchmark. At the same time, IRQAM exhibits comparatively weaker performance on distortion-oriented IQA tasks, while providing substantially higher correlation for IRQA. These results indicate that IRQA is fundamentally different from traditional distortion-based IQA and that each task benefits from models specifically trained to address its unique objectives.

Discussion

As demonstrated in the previous sections, the REPID introduces a novel and challenging problem setting within the field of IQA: namely, IRQA. Existing IQA methods exhibit substantially lower performance on REPID compared to their results on traditional IQA datasets, as well as relative to the maximum achievable accuracy. Even models trained specifically for the REPID task reach only 0.83 accuracy, whereas the empirically found practical upper bound appears to be about 0.89. This gap suggests that predicting the quality of rendering remains a difficult problem, likely due to the subjective and fine-grained nature of aesthetics judgments.

One factor contributing to the difficulty of the task is the prevalence of image pairs with preference scores near 0.5, indicating ambiguous or subtle differences between renders. These “hard” pairs reduce the separability of classes and limit the performance of both classical and deep learning–based methods.

Although metadata describing the applied rendering parameters (e.g., saturation, contrast, color temperature) have information useful for prediction to some extent, its use alone does not yield state-of-the-art performance. Moreover, simply concatenating this metadata with deep learning embeddings does not significantly improve the results. A promising future direction may involve explicitly modeling the effect of each retouching parameter — potentially through submodules that approximate the transformations applied by Adobe Lightroom — and integrating these parameter-conditioned transformations into the network. Such an approach, however, would require partial reverse engineering of Adobe’s rendering pipeline or learning differentiable approximations of these adjustments.

An important advantage of REPID is its potential for extensibility. The dataset can be expanded by collecting additional votes, which would reduce noise and improve the stability of PMOS. In particular, increasing the number of evaluators per image would help mitigate the impact of individual biases and provide a more robust representation of collective aesthetic preference. Such expansions would make REPID even more valuable as a benchmark for fine-grained aesthetic prediction.

Despite its advantages, REPID has several limitations. The current version contains only six rendering styles, which represent a limited portion of the possible parameter space. This excludes many extreme or intentionally poor combinations of parameters (e.g., overly saturated or severely underexposed images) that could help models learn a wider range of aesthetic judgments. Future versions of the dataset could incorporate a richer and more diverse set of retouching styles to broaden its coverage and challenge models to generalize across a wider aesthetic spectrum.

It is worth noting that, for markup processing, we implemented a procedure similar to those used in well-established works¹⁵ on vote collection. While this methodology is widely accepted in the scientific community, there remains potential for improvement in vote aggregation by explicitly accounting for the reliability and quality of annotations provided by individual users.

Beyond its immediate utility for benchmarking, REPID opens opportunities for several research directions. It can serve as a benchmark for developing models capable of learning subtle perceptual differences, as well as for studying human aesthetic biases across image categories and renders. Furthermore, the availability of evaluator identifiers allows investigation of personalization and user-specific preference modeling — an area of growing importance in content curation, recommendation systems, and generative AI. Finally, the dataset provides a bridge between traditional IQA research and the broader study of image aesthetics, motivating the development of models that move beyond distortion-based quality prediction toward understanding visual appeal.

Conclusions

In this work, we introduced the REPID—a novel benchmark for studying aesthetic preferences prediction and render evaluation through the IRQA task. REPID provides pairwise preference annotations for 5000 images, each rendered into multiple styles, making it uniquely suited for fine-grained analysis of human aesthetic judgments. Through extensive experiments, we demonstrated that existing IQA methods perform poorly on the IRQA task compared to their results on standard quality-assessment datasets, highlighting the need for new approaches that explicitly model aesthetic factors. Even purpose-trained models achieved only 0.83 accuracy, leaving a substantial gap from the achievable upper bound of 0.89.

We further explored the predictive power of image metadata and handcrafted features, finding that they offer useful but limited information, and we showed that integrating them with deep learning models does not trivially improve performance. These findings emphasize the complexity of aesthetic perception and suggest that future models may need to incorporate parameter-aware or differentiable post-processing modules to better capture the underlying retouch transformations.

REPID not only provides a challenging benchmark, but also opens several promising research directions, including personalized preference prediction, render ranking, and generative modeling of preferred styles. As a dataset, REPID can be extended with additional votes and new renders, improving the robustness of mean opinion scores and expanding the diversity of aesthetic scenarios. We hope that REPID will foster the development of next-generation aesthetic prediction models and serve as a common ground for comparing approaches at the intersection of IQA, computational aesthetics, and human-AI co-creation.

Future work will include generating new renders that cover parameter field more broadly in order to obtain more generalized model that can be used as loss for automatically creating the most preferable styles. Moreover, generating parameter-conditioned transformations and integrating them into the network might reduce the gap with respect to the maximum achievable performance.

Data availability

The dataset is available at <https://huggingface.co/datasets/vsevolodpl/REPID>.

Received: 24 March 2026; Accepted: 15 May 2026

Published online: 27 June 2026

References

1. Zhang, W., Zhai, G., Wei, Y., Yang, X. & Ma, K. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. *Paper Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 18–23 June 2023*. <https://doi.org/10.48550/arXiv.2303.14968> (2023).
2. Zhai, G. & Min, X. Perceptual image quality assessment: A survey. *Sci. China Inf. Sci.* **63**, 211301. <https://doi.org/10.1007/s11432-019-2757-1> (2020).
3. Hosu, V., Agnolucci, L., Wiedemann, O., Iso, D. & Saupe, D. UHD-IQA benchmark database: Pushing the boundaries of blind photo quality assessment. Preprint at [arXiv:2406.17472](https://arxiv.org/abs/2406.17472) (2024).
4. Ershov, E. et al. Reliability and stability of mean opinion score for image aesthetic quality assessment obtained through crowdsourcing. *Paper Presented at the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, Rome, Italy, 27–29 February 2024*. <https://doi.org/10.5220/0012462000003660> (2024).
5. Vladimir, B., Sylvain, P., Eric, C. & Frédo, D. Learning photographic global tonal adjustment with a database of input / output image pairs. *Paper presented at the Twenty-Fourth IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 20–25 June 2011*. <https://doi.org/10.1109/CVPR.2011.5995413> (2011).
6. Sheikh, H., Wang, Z., Cormack, L. & Bovik, A. Live image quality assessment database release 2. Dataset source <http://live.ece.utexas.edu/research/quality> (2005).
7. Andela, Z. et al. VCL@FER image quality assessment database. *Automatika* **53**, 344–354. <https://doi.org/10.7305/automatika.53-4.241> (2012).
8. Ruikar, J. & Chaudhury, S. NITS-IQA database: A new image quality assessment database. *Sensors* **23**, 2279. <https://doi.org/10.3390/s23042279> (2023).
9. Chandler, D. M. Most apparent distortion: Full-reference image quality assessment and the role of strategy. *J. Electron. Imaging* **19**, 011006. <https://doi.org/10.1117/1.3267105> (2010).
10. Ponomarenko, N. et al. Image database TID2013: Peculiarities, results and perspectives. *Image Commun.* **30**, 57–77. <https://doi.org/10.1016/j.image.2014.10.009> (2015).
11. Wen, S., Fei, Z. & Qingmin, L. MDID: A multiply distorted image database for image quality assessment. *Pattern Recognit.* **61**, 153–168. <https://doi.org/10.1016/j.patcog.2016.07.033> (2017).
12. Tsung, J. L., Hsin, H. L., Soo, C. P. & Kuan, H. L. A high-definition diversity-scene database for image quality assessment. *IEEE Access* **6**, 45427–45438. <https://doi.org/10.1109/ACCESS.2018.2864514> (2018).
13. Hanhe, L., Vlad, H. & Dietmar, S. DeepFL-IQA: Weak Supervision for Deep IQA Feature Learning. Preprint at [arXiv:2001.08113](https://arxiv.org/abs/2001.08113) (2020).
14. Siahhaan, E., Hanjalic, A. & Redi, J. A reliable methodology to collect ground truth data of image aesthetic appeal. *IEEE Trans. Multimed.* **18**, 1338–1350. <https://doi.org/10.1109/TMM.2016.2559942> (2016).
15. Hosu, V., Lin, H. & Saupe, D. Expertise screening in crowdsourcing image quality. *Paper Presented at the 2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX), Cagliari, Italy, 29 May - 01 June 2018*. <https://doi.org/10.1109/QoMEX.2018.8463427> (2018).

16. Vlad, H., Hanhe, L., Tamas, S. & Dietmar, S. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Trans. Image Process.* **29**, 4041–4056. <https://doi.org/10.1109/TIP.2020.2967829> (2020).
17. Murray, N., Marchesotti, L. & Perronnin, F. Ava: A large-scale database for aesthetic visual analysis. *Paper Presented at the 2012 IEEE 2012 IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, 16–21 June 2012. <https://doi.org/10.1109/CVPR.2012.6247954> (2012).
18. Kairanbay, M., See, J. & Wong, L.-K. Towards demographic-based photographic aesthetics prediction for portraits. *Paper Presented at the MultiMedia Modeling: 24th International Conference, MMM 2018, Bangkok, Thailand, 5–7 February 2018*. https://doi.org/10.1007/978-3-319-73603-7_43 (2018).
19. Kang, C., Valenzise, G. & Dufaux, F. Eva: An explainable visual aesthetics dataset. *Paper Presented at the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020*. <https://doi.org/10.1145/3423268.3423590> (2020).
20. Yang, Y. et al. Personalized image aesthetics assessment with rich attributes. *Paper Presented at the 2012 IEEE 2012 IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, Louisiana, USA, 21–24 June 2022*. <https://doi.org/10.48550/arXiv.2203.16754> (2022).
21. Mantiuk, R. K., Tomaszewska, A. & Mantiuk, R. Comparison of four subjective methods for image quality assessment. *Comput. Graph. Forum* **31**, 2478–2491. <https://doi.org/10.1111/j.1467-8659.2012.03188.x> (2012).
22. Wu, X. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. Preprint at [arXiv:2306.09341](https://arxiv.org/abs/2306.09341) (2023).
23. Chunyi, L. et al. AGIQA-3K: An open database for AI-generated image quality assessment. *IEEE Trans. Circuits Syst. Video Technol.* **34**, 6833–6846. <https://doi.org/10.1109/TCSVT.2023.3319020> (2023).
24. Li, C. et al. Aigiqa-20k: A large database for ai-generated image quality assessment. *Paper Presented at the 2024 IEEE Conference on Computer Vision and Pattern Recognition, Seattle WA, USA, 17–21 June 2024*. <https://doi.org/10.48550/arXiv.2404.03407> (2024).
25. Yuan, J. Pku-i2iaqa: An image-to-image quality assessment database for ai generated images. Preprint at [arXiv:2311.15556](https://arxiv.org/abs/2311.15556) (2023).
26. Ak, A., Goswami, A., Hauser, W., Le Callet, P. & Dufaux, F. RV-TMO: Large-scale dataset for subjective quality assessment of tone mapped images. *IEEE Trans. Multimed.* **25**, 6013–6025. <https://doi.org/10.1109/TMM.2022.3203211> (2022).
27. Jiang, Q., Li, X., Wang, X., Wang, Z. & Zhai, G. Dataset and metric for quality assessment of HDR tone mapping: Detail visibility, color naturalness, and overall quality. *IEEE Trans. Multimed.* <https://doi.org/10.1109/TMM.2025.3535338> (2025).
28. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **13**, 600–612. <https://doi.org/10.1109/TIP.2003.819861> (2004).
29. Zhang, R., Isola, P., Efros, A. A., Shechtman, E. & Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. <https://doi.org/10.48550/arXiv.1801.03924> (2018).
30. Anish, M., Anush, K. M. & Alan, C. B. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.* **21**, 4695–4708. <https://doi.org/10.1109/TIP.2012.2214050> (2012).
31. Schölkopf, B., Smola, A. J., Williamson, R. C. & Bartlett, P. L. New support vector algorithms. *Neural Comput.* **12**, 1207–1245. <https://doi.org/10.1162/089976600300015565> (2000).
32. Venkatanath, N., Praneeth, D., Bh, M. C., Channappayya, S. S. & Medasani, S. S. Blind image quality evaluation using perception based features. *Paper Presented at the 2015 Twenty First National Conference on Communications (NCC), Mumbai, India, 27 February - 01 March 2015*. <https://doi.org/10.1109/NCC.2015.7084843> (2015).
33. Saad, M. A., Bovik, A. C. & Charrier, C. Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE Trans. Image Process.* **21**, 3339–3352 (2012).
34. Liu, L., Liu, B., Huang, H. & Bovik, A. C. No-reference image quality assessment based on spatial and spectral entropies. *Signal Process. Image Commun.* **29**, 856–863. <https://doi.org/10.1016/j.image.2014.06.006> (2014).
35. Mittal, A., Soundararajan, R. & Bovik, A. C. Making a completely blind image quality analyzer. *IEEE Signal Process. Lett.* **20**, 209–212 (2012).
36. Liu, L., Hua, Y., Zhao, Q., Huang, H. & Bovik, A. C. Blind image quality assessment by relative gradient statistics and adaboosting neural network. *Signal Process. Image Commun.* **40**, 1–15 (2016).
37. Bianco, S., Celona, L., Napoletano, P. & Schettini, R. On the use of deep learning for blind image quality assessment. *Signal Image Process.* **12**, 355–362. <https://doi.org/10.1007/s11760-017-1166-8> (2018).
38. Ke, J., Wang, Q., Wang, Y., Milanfar, P. & Yang, F. Musiq: Multi-scale image quality transformer. *Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition*, 19–25 June 2021. <https://doi.org/10.48550/arXiv.2108.05997> (2021).
39. Ying, Z. et al. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. <https://doi.org/10.48550/arXiv.1912.10088> (2020).
40. Madhusudana, P. C., Birkbeck, N., Wang, Y., Adsumilli, B. & Bovik, A. C. Image quality assessment using contrastive learning. *IEEE Trans. Image Process.* **31**, 4149–4161. <https://doi.org/10.1109/TIP.2022.3181496> (2022).
41. Saha, A., Mishra, S. & Bovik, A. C. Re-iqa: Unsupervised learning for image quality assessment in the wild. <https://doi.org/10.48550/arXiv.2304.00451> (2023).
42. Agnolucci, L., Galteri, L., Bertini, M. & Del Bimbo, A. Arniqa: Learning distortion manifold for image quality assessment. *Paper presented at the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, Hawaii, 3–7 January 2023*. <https://doi.org/10.1109/WACV57701.2024.00026> (2023).
43. Wan, D., Jiang, X. & Yu, Q. Blind HDR image quality assessment based on aggregating perception and inference features. *Sci. Rep.* **15**, 10808. <https://doi.org/10.1038/s41598-025-94005-1> (2025).
44. Wang, J., Chan, K. C. & Loy, C. C. Exploring clip for assessing the look and feel of images. *Paper Presented at the 37th AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023*. <https://doi.org/10.1609/aaai.v37i2.25353> (2023).
45. Agnolucci, L., Galteri, L. & Bertini, M. Quality-Aware Image-Text Alignment for Opinion-Unaware Image Quality Assessment. Preprint at [arXiv:2403.11176](https://arxiv.org/abs/2403.11176) (2024).
46. Avnaki, N. J., Ghildyal, A., Barman, N. & Zadtootaghaj, S. LAR-IQA: A lightweight, accurate, and robust no-reference image quality assessment model. *Paper Presented at the 18th European Conference on Computer Vision ECCV, MiCo, Milano, 29 September – 04 October 2024*. <https://doi.org/10.48550/arXiv.2408.17057> (2024).
47. Liu, Z. et al. Kan: Kolmogorov-arnold networks. *Paper Presented at the Twelfth International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024*. <https://doi.org/10.48550/arXiv.2404.19756> (2024).
48. Chen, P. RAM-VQA: Restoration Assisted Multi-modality Video Quality Assessment. *IEEE Transactions on Image Processing* (2026).
49. Hu, Z., Liu, L., Gu, K., Li, L. & Bovik, A. Distortion-sensitive masked autoencoder for omnidirectional video quality assessment. *IEEE Trans. Multimed.* <https://doi.org/10.1109/TMM.2026.3651128> (2026).
50. Zhang, Z. et al. Large multimodal models evaluation: A survey. *Sci. China Inf. Sci.* <https://doi.org/10.1007/s11432-025-4676-4> (2025).
51. Zhang, Z. et al. Towards versatile multimedia quality assessment for visual communications. *Sci. China Inf. Sci.* <https://doi.org/10.1007/s11432-025-4631-2> (2026).
52. Li, J., Li, D., Xiong, C. & Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *Paper presented at the Thirty-Ninth International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022*. <https://doi.org/10.48550/arXiv.2201.12086> (2022).
53. Bradley, R. A. & Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika* **39**, 324–345. <https://doi.org/10.2307/2334029> (1952).

54. Parducci, A. Category judgment: A range-frequency model. *Psychol. Rev.* **72**, 407–418. <https://doi.org/10.1037/h0022602> (1965).
55. Tversky, A. & Kahneman, D. Judgment under uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science* **185**, 1124–1131. <https://doi.org/10.1126/science.185.4157.1124> (1974).

Acknowledgements

We would like to thank Kuznetsov Maksim, Broymtan Aleksandra, Provotorova Kseniia Andreevna, Sarafanova Ekaterina, Illarionova Kira, Pevnaya Elizaveta, Cook Nikol-Maria Thomas, Myasnikova Ksenia Dmitrievna, Mamontova Ekaterina, Bilichenko Vera, Rizaev Anton, Ulanov Konstantin Vladimirovich, Esaulkova Anastasia Mikhailovna, Kozyreva Daria Alexandrovna from AXXX for helping with the markup process.

Author contributions

V.P. was responsible for dataset markup and subsequent analysis of annotation consistency and quality. Moreover, he developed Model for Evaluating the Aesthetics of Rendering (IRAAM). A.P. conducted the analysis of existing image quality metrics using both public datasets and the proposed benchmark dataset. Also he trained new neural networks for the Rendering Evaluation of Photographic Image Dataset (REPID). N.B. handled data preprocessing for markup, implemented experiments related to rendering style selection, and performed statistical tests for dataset analysis. S.B. defined the technical requirements and supervised the development and validation of the proposed benchmark. M.F. acted as the leading expert on aesthetic rendering, guiding the selection of rendering styles and defining the most informative rendering parameters. E.E. originated the idea of assessing the aesthetic quality of rendered images using DMOS and contributed to the experimental design.

Funding

The work was supported by the Ministry of Economic Development of the Russian Federation (agreement No. 139-15-2025-013, dated June 20, 2025, IKG 000000C313925P4B0002).

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

This research was carried out in accordance with the Declaration of Helsinki, and includes ethical approval from the Ethics Committee of the Federal State Budgetary Institution of Science — the Kharkevich Institute for Information Transmission Problems of the Russian Academy of Sciences (IITP RAS) 2026-03-05.

Additional information

Electronic supplementary material The online version of this article (<https://doi.org/10.1038/s41598-026-54744-1>) contains supplementary material, which is available to authorized users.

Correspondence and requests for materials should be addressed to S.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026