Art and Design Review, 2026, 14(4), 333-358
https://www.scirp.org/journal/adr
ISSN Online: 2332-2004
ISSN Print: 2332-1997

A Computational Comparison of Human Artworks and Manually Curated DALL·E 2 Images with Matched Styles

Yakun Gu, Yiying Zhang, Tianling Huang*

School of Art, Nantong University, Nantong, China
Email: 15862006368@163.com, *htl828@ntu.edu.cn

How to cite this paper: Gu, Y. K., Zhang, Y. Y., & Huang, T. L. (2026). A Computational Comparison of Human Artworks and Manually Curated DALL·E 2 Images with Matched Styles. *Art and Design Review, 14,* 333-358.
https://doi.org/10.4236/adr.2026.144017

Received: September 28, 2026
Accepted: November 16, 2026
Published: November 19, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

Abstract

This exploratory secondary study compares 50 human artworks with 50 independently generated, manually style-matched DALL·E 2 images from an official stimulus archive. Thirteen color, grayscale, local-transition, and frequency features were computed after common RGB decoding and resizing. Paired estimation used 100,000 BCa bootstrap resamples and Holm correction across 13 tests. The AI images showed higher mean green intensity ($\Delta = 0.0537$), blue-channel dispersion ($\Delta = 0.0448$), and grayscale entropy ($\Delta = 0.241$ bits), and lower high-frequency energy ($\Delta = -0.0305$). Blue dispersion and entropy retained corrected support under equal artist weighting. Green and frequency estimates barely changed, but their adjusted tests were less conclusive with 28 artist means. Matching advantages in grayscale cosine and RGB histograms were attenuated but remained supported under genre-restricted permutations; subgroup evidence was less precise for still life. A ten-pair external variation sample yielded no corrected effects. Common archived JPEG encoding did not resolve unknown upstream color and compression histories. These findings describe selected visual differences in a curated corpus and do not establish transformations of individual artworks or explain aesthetic preference.

Keywords

Generative Artificial Intelligence, Visual Art, DALL·E 2, Image Statistics, Paired Comparison, Curation

1. Introduction

Generative visual systems have moved from adversarial image synthesis to diffusion-based, language-conditioned production. Generative adversarial networks

established a trainable framework for synthetic imagery, and style-based architectures improved control over appearance and scale (Goodfellow et al., 2014; Karras et al., 2019; Karras et al., 2020). Denoising diffusion and score-based models subsequently offered stable, high-fidelity generation (Ho et al., 2020; Nichol & Dhariwal, 2021; Rombach et al., 2022; Saharia et al., 2022; Song et al., 2021). In parallel, large image-text models and text-to-image systems made natural-language descriptions operational inputs for visual production (Radford et al., 2021; Ramesh et al., 2022; Ramesh et al., 2021). These developments allow a prompt to invoke subject matter, medium, genre, and artist-associated conventions without requiring a source image.

The relation between computation and artistic style predates contemporary diffusion systems. Neural style transfer separated learned content and texture-like style representations, while feed-forward variants accelerated their application (Gatys et al., 2016; Johnson et al., 2016). Art-specific generative systems then explored categorical styles and deliberate deviation from learned norms (Elgammal et al., 2017; Tan et al., 2017). This technical trajectory has sharpened older questions about machine creativity, autonomy, authorship, and intention (Hertzmann, 2018; Mazzone & Elgammal, 2019; McCormack et al., 2019). It also changes the practical status of style: historically situated conventions can be rendered as prompt conditions and recombined at scale.

Empirical work has examined how viewers respond to this change. Studies compare aesthetic responses to computer-generated and human-made works, perceived emotion, origin discrimination, and the effect of an AI authorship label (Agudo et al., 2022; Chamberlain et al., 2018; Chiarella et al., 2022; Gangadharbatla, 2022; Hong & Curran, 2019; Ragot et al., 2020). Judgments are not uniform. Beliefs about whether a system can count as an artist, preferences for human authorship, and anthropocentric bias influence evaluation (Bellaiche et al., 2023; Horton et al., 2023; Mikalonytė & Kneer, 2022; Millet et al., 2023). A recent systematic review documents the expanding empirical field (Oksanen et al., 2023). van Hees et al. (2025) found that observers preferred selected DALL·E 2 images above chance while also identifying their origin above chance, demonstrating that preference and detectability need not oppose each other.

These behavioral findings do not by themselves identify the image properties that differ. Computational aesthetics has described photographs and artworks through composition, color, complexity, and learned quality representations (Datta et al., 2006; Dhar et al., 2011; Ke et al., 2006; Murray et al., 2012; Oliva & Torralba, 2001; Talebi & Milanfar, 2018). Related work classifies artistic style and locates artworks in learned historical representations (Elgammal et al., 2018; Karayev et al., 2014; Saleh & Elgammal, 2016). Perceptual image metrics and distributional measures provide further ways to compare generated and reference images (Heusel et al., 2017; Wang et al., 2004; Zhang et al., 2018). Such methods can make formal differences auditable, but their interpretation must remain narrow. A change in entropy, color variance, or spatial frequency is not a direct measure of

beauty, creativity, intention, or meaning. Psychological and neuroaesthetic models treat art experience as an interaction among stimulus structure, knowledge, context, affect, and the viewer (Chatterjee & Vartanian, 2014; Graham & Redies, 2010; Leder et al., 2004; Pelowski et al., 2016; Redies, 2015; Van Geert & Wagemans, 2020).

The present study is a secondary analysis of the published stimulus corpus of van Hees et al. (2025), rather than an independent collection of artworks or generated images. Their image-statistics work already included GIST, intensity, root-mean-square contrast, and entropy. The contribution here is paired effect estimation within a transparent 13-feature family, multiplicity control, BCa intervals, artist averaging, multiresolution checks, and restricted mismatch baselines. The questions are which visual features differ within designated pairs, whether the matching advantage remains when broad genre is held fixed, and how estimates depend on weighting, resolution, and a small external variation sample.

Throughout, a style-matched AI image means an independently generated text-to-image output that the source-study authors manually paired with a human artwork according to artist, genre, and broad visual properties. The paired human image was not supplied as an image-to-image input. The estimand is therefore a difference between selected human and AI images under this matching procedure. It does not isolate generation from subsequent human selection.

2. Materials and Methods

2.1. Study Design and Analysis Unit

This was a secondary, paired computational study. For image pair i and feature j, the contrast was

$$d_{ij} = x_{ij}^{\mathrm{AI}} - x_{ij}^{\mathrm{human}} \tag{1}$$

The image pair, rather than the pixel, was the statistical unit. Positive differences indicate higher values in the synthetic image. The primary analysis was exploratory and was not preregistered; multiplicity control and sensitivity analyses were therefore used to limit overinterpretation. **Figure 1** summarizes the analysis sequence.

2.2. Primary Paired Stimulus Corpus

We used the public stimulus set accompanying van Hees et al. (2025). The corpus comprises 50 human-made artworks and 50 DALL·E 2 images arranged into 50 matched pairs. According to the source study, the human images were lesser-known works by established representational artists, obtained from Wikimedia Commons and WikiArt and chosen to remain representative of the relevant artist without appearing in WikiArt's "famous work" section. The stimuli span Baroque, Romantic, Impressionist, and Post-impressionist traditions.

The source authors generated the synthetic images with 36 unique prompts that combined an artist's name and an artwork category; some prompts were reused.

Each prompt returned several candidates. The authors manually compared candidates with catalogued works, retained images considered comparable in color, style, or composition and reasonably representative of the artist, excluded conspicuous outliers, and cropped the DALL·E 2 logo from selected outputs. Thus, the source procedure includes human curation. It cannot isolate an automatic model effect.

We downloaded the official resized archive from the authors' OSF repository (https://osf.io/n7w32/; direct archive: https://osf.io/download/e2mr4/). It contains 100 JPEG files: 99 are 512 × 512 pixels and one is 512 × 511. The source analysis code classifies stimulus identifiers 1 - 50 as human-made and 51 - 100 as AI-generated and defines a match by an index difference of 50. Human stimulus i was therefore paired with AI stimulus i + 50, for $i = 1, \cdots, 50$. SHA-256 hashes, stimulus identifiers, prompts, paths, and dimensions were recorded in a manifest. Figure 2 illustrates four designated pairs.

An audit found that all 100 archived files were RGB JPEGs without embedded ICC profiles. All shared 4:4:4 chroma sampling and identical quantization tables. Two human files retained Photoshop EXIF metadata, including an uncalibrated color-space tag; the other files did not retain a comparable upstream record. These observations characterize final archived encoding only. Equivalent source-profile conversion, earlier compression, resizing, and re-export cannot be established from the archive. Cropping of the DALL·E 2 logo is documented. Such upstream operations may contribute to color, entropy, and frequency differences.

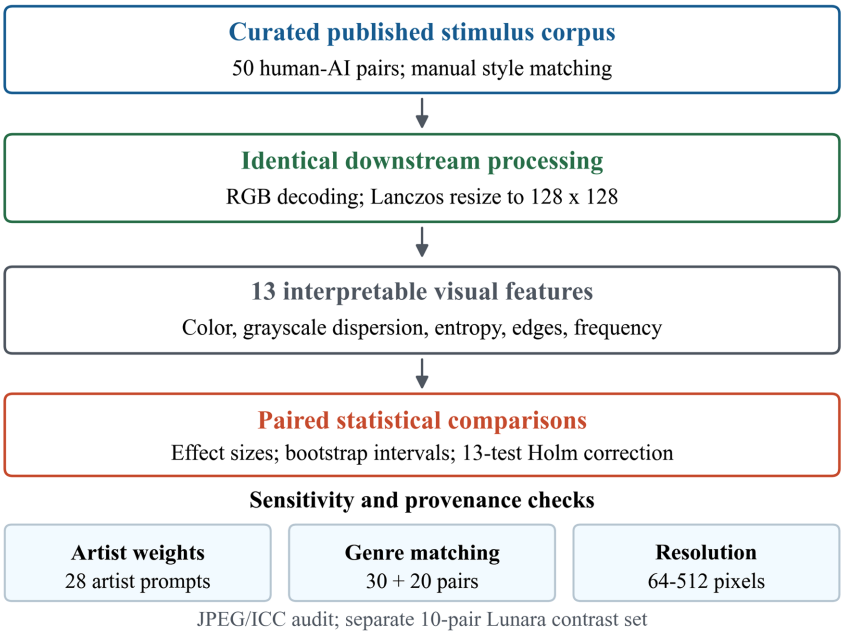


Figure 1. Analysis workflow for a secondary comparison of 50 independently created, manually style-matched human-AI pairs. Downstream preprocessing was shared; upstream image-processing histories were not fully recoverable. All 13 features formed the primary Holm family. Genre-restricted matching, artist weighting and resolution analyses assess sensitivity; the 10-pair Lunara contrast set remains separate.

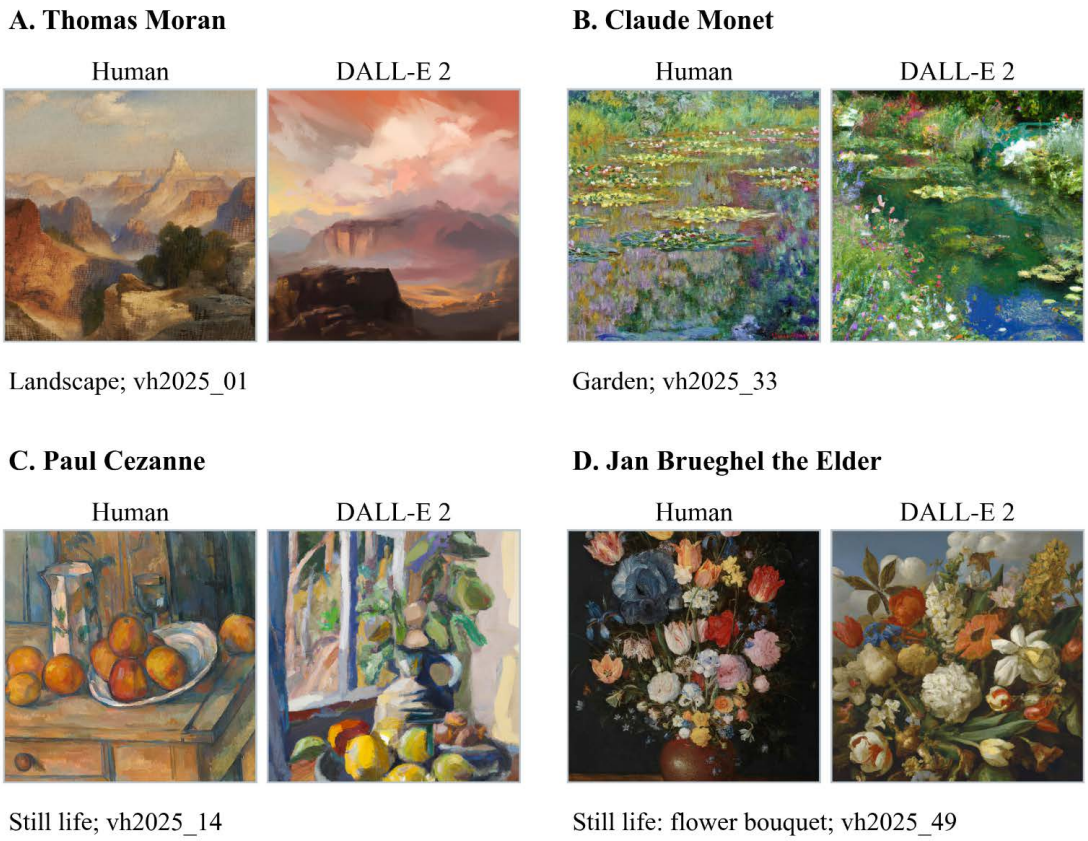


Figure 2. Four illustrative pairs from the official published stimulus archive, shown without crop modification: (A) Thomas Moran, landscape; (B) Claude Monet, garden; (C) Paul Cezanne, still life; (D) Jan Brueghel the Elder, flower bouquet. Labels identify the artist/style prompt attached to each curated human-AI pair; the AI image was generated independently and manually selected, rather than transformed from the human image. Examples are illustrative, not a random subsample.

2.3. External Image-Variation Pilot and Excluded Source Pool

As a limited process-control analysis, we used 10 anchor-variant pairs from Moonworks Lunara Aesthetic II (Wang et al., 2026), traced to Hugging Face commit 33f418207770cb5f0aad7f4f67f376b7bf41267f. The full dataset contains original photographs and artworks together with generated contextual variants. The local convenience sample contained six nature/landscape, two city/architecture, one everyday-life, and one religion/spirituality anchor. Its prompts deliberately changed illumination, weather, scene composition, mood, viewpoint, or color tone. Because it contains only 10 heterogeneous pairs and is dominated by photographic scenes, it was treated as a stress test rather than a replication.

Twelve public-domain Art Institute of Chicago images were also preserved in the workspace as a future controlled-generation pool. They had no corresponding AI outputs and were not analyzed. This exclusion prevents unpaired originals from being misrepresented as completed human-AI comparisons. Table 1 summarizes these distinct data roles.

Table 1. Data sources and their distinct analytical roles.

Dataset	Size	Role	Provenance and limits
van Hees official stimuli	50 pairs	Primary comparison	Official OSF archive; human artworks and curated DALL·E 2 images
Lunara convenience subset	10 pairs	External variation check	Mainly photographs; contextual changes; not pooled with the art corpus
Art Institute of Chicago	12 images	Excluded future source pool	No AI counterparts; not included in statistical analyses

Note. The primary corpus and external convenience sample are analyzed separately. The unpaired museum pool contributes no results.

2.4. Preprocessing and Visual Features

Each archived file was decoded with Pillow, converted to RGB mode, and resized to 128 × 128 pixels with Lanczos interpolation. All archived images were already RGB and lacked ICC profiles, so this was a common mode-and-size pipeline, not a profile-aware conversion to a verified colorimetric space. Values were analyzed as decoded RGB codes; no linear-light conversion was performed. No color augmentation, denoising, sharpening, or random crop was applied. The same deterministic function processed both members of every pair. Nearly square primary images required negligible geometric alteration; heterogeneous Lunara images underwent direct square resizing.

The 13-feature family provides a lightweight description of channel level and dispersion, tonal distribution, local transitions, and spectral allocation. RGB means and standard deviations separate color balance from within-channel variation; grayscale summaries describe complementary tonal properties; channel range measures chromatic spread; gradient and Fourier measures summarize spatial structure. These choices follow image-statistics and computational-aesthetics practice (Datta et al., 2006; Graham & Redies, 2010; van Hees et al., 2025). The measures are partly correlated descriptive summaries, not independent aesthetic constructs. All 13 remain in the correction family.

RGB values were normalized to [0, 1]. The 13 features comprised: mean red, green, and blue intensity; within-image standard deviation of each channel; mean and standard deviation of a grayscale proxy Y' = 0.299R + 0.587G + 0.114B; mean channel range max(R, G, B)-min(R, G, B) as a saturation proxy; grayscale histogram entropy; edge density; mean gradient magnitude; and high-frequency energy. Because Y' was calculated on gamma-encoded RGB values, it is called grayscale intensity, not physical luminance.

Grayscale entropy was calculated from a 256-bin eight-bit histogram,

$$H = -\sum_{k=0}^{255} p_k \log_2 p_k \tag{2}$$

Horizontal and vertical first differences of normalized grayscale were combined into gradient magnitude, with a zero-padded final row or column. Edge density

was the proportion of pixels with magnitude greater than 0.12. This operational cutoff suppresses relatively small fluctuations on the [0, 1] intensity scale; it is not a calibrated perceptual boundary. Mean gradient provides a threshold-free companion measure. Fourier power was calculated after mean subtraction. Radial DFT distance was divided by N/2, where N is image width, so a radius of 1 equals the axis Nyquist frequency. High-frequency energy was the fraction of power outside normalized radius 0.25, equivalent to radial frequency above 0.125 cycles per pixel. These fixed cutoffs were retained from the existing implementation; no preregistration or documented data-blind threshold choice is claimed. A post hoc threshold grid is reported separately.

Two additional pair-structure measures were analyzed separately from the 13-feature outcome family. Global grayscale layout was the cosine similarity between flattened 128 × 128 grayscale arrays. Color-distribution distance was the Jensen-Shannon divergence between 8 × 8 × 8 RGB histograms. These are low-level alignment measures rather than semantic embeddings.

2.5. Statistical Analysis

For each primary feature, we report the human and AI mean and standard deviation, mean paired difference, 95% bias-corrected and accelerated (BCa) bootstrap interval from 100,000 pair-level resamples, and standardized paired effect

$$d_z = \frac{\bar{d}}{s_d} \tag{3}$$

Effect sizes were interpreted as estimates rather than labels (Lakens, 2013). Two-sided paired t tests assessed a zero mean difference. Wilcoxon signed-rank tests served as a distribution-robust sensitivity analysis. Holm's sequential method controlled family-wise error across all 13 features separately for the primary and Lunara datasets (Holm, 1979). The decision threshold was adjusted $p < 0.05$; marginal bootstrap intervals were not treated as simultaneous intervals.

We formed complete 50 × 50 grayscale-cosine and RGB-histogram Jensen-Shannon divergence (JSD) matrices. The diagonal contained designated matches; the 2450 off-diagonal cells contained unrestricted mismatches. For each human-image row, the matching advantage was designated cosine minus mean mismatched cosine, or mean mismatched JSD minus designated JSD, so both advantages are positive for closer matching. One-to-one AI-label permutations tested the observed mean designated statistic, using 200,000 assignments with fixed points allowed. Directional alternatives were higher cosine and lower JSD; Monte Carlo p values used (b + 1)/(B + 1), where b counts assignments at least as extreme as observed. These are curation checks rather than reconstruction tests.

We repeated assignments within landscape/garden (n = 30) and still life (n = 20), and pooled the 50 pairs while preserving these groups during permutation. Same-genre off-diagonal candidates defined the restricted mismatch means. We report raw p values and Holm correction across four subgroup tests; the two

pooled restricted tests form a separate family. All assignment tests used 200,000 permutations. Marginal BCa intervals used 100,000 resamples of row advantages; pooled restricted resampling retained the 30:20 genre weights. Rows share AI candidates, so these intervals are conditional descriptive summaries rather than independent superpopulation effects. Assignment permutation is the principal matching test and need not agree with the row-bootstrap interval.

Exploratory genre analysis collapsed 27 landscape and three garden pairs into landscape/garden (n = 30) and grouped 13 still-life, six flower-bouquet, and one fruit still-life pair as still life (n = 20). Welch tests compared feature differences between categories, with Holm correction across 13 contrasts. Because 50 pairs represented 28 artists unequally, an artist-equal-weight sensitivity analysis first averaged each feature difference within artist and repeated the 13-test family on 28 artist means. A resolution sensitivity analysis recomputed all features at 64, 128, 256, and 512 pixels, applying Holm correction within each resolution.

Analyses used Python 3.13.7, NumPy 2.3.3, pandas 2.3.2, SciPy 1.16.2, Pillow 11.3.0, and Matplotlib 3.10.6. Fixed seeds were used for resampling. Manifests contain SHA-256 hashes for all images; derived tables and scripts follow FAIR-oriented provenance principles (Wilkinson et al., 2016).

Supporting analyses are presented in **Figure S1** and **Figure S2** and **Table S1**, **Table S2**, **Table S3**, and **Table S4**. These document the external sample, resolution dependence, artist averaging, genre contrasts, and post hoc threshold checks. The new encoding and restricted-permutation outputs accompany the replication package.

2.6. Licensing and Ethics

The Frontiers article and its published figures are distributed under CC BY 4.0. That article-level license is not evidence that every independently hosted Wikimedia Commons or WikiArt reproduction carries the same license. The official OSF archive was used for analysis, and the source location is preserved rather than recasting each underlying artwork as CC BY. Lunara's data repository reports Apache-2.0. This study collected no new human-participant or personal data; institutional human-subject approval was therefore not applicable to the present image-only secondary analysis.

3. Results

3.1. Corpus Characteristics and Matching Structure

All 50 primary pairs were complete. The 100 files had unique SHA-256 hashes, and the manifest-to-file checks found no duplicate image. The corpus represented 28 artist prompts: 11 artists appeared once, 13 twice, three three times, and one four times. The category labels comprised 27 landscapes, three gardens, 13 still lifes, six flower-bouquet still lifes, and one fruit still life.

The designated pairs were closer than unrestricted mismatches on both low-

level measures (Table 2 and Figure 3). Mean grayscale cosine was 0.8303 for designated pairs and 0.8094 for mismatches; the oriented advantage was 0.0209 (95% BCa CI [0.0051, 0.0356]), with permutation $p = 0.000325$. Mean histogram JSD was 0.4973 and 0.5880, respectively; its oriented advantage was 0.0907 [0.0557, 0.1209], with $p < 0.00001$. Both remain supported after two-test Holm correction. These tests describe resemblance achieved by manual curation, without establishing semantic equivalence or reconstruction.

Restricting mismatch candidates to the same genre reduced the pooled cosine advantage to 0.0135 [0.0004, 0.0278] and the histogram advantage to 0.0743 [0.0393, 0.1044] (Table 3). Permutation p values were 0.006525 and 0.000010, with Holm-adjusted values 0.006525 and 0.000020. Within landscape/garden, adjusted p values were 0.03159 for cosine and 0.000020 for JSD. Still-life adjusted values were 0.17478 and 0.18137. This attenuation is consistent with a contribution from broad genre matching, while the pooled advantage remains supported. It is not evidence of a genre interaction: the still-life cosine point advantage (0.0151) slightly exceeded the landscape/garden estimate (0.0125) but was less precise. The landscape/garden cosine row-bootstrap interval included zero despite its permutation result, reflecting different statistics and conditioning assumptions.

Table 2. Designated matching compared with unrestricted mismatch candidates.

Metric	Matched M (SD)	Mismatch M	Advantage	Permutation p
Grayscale cosine	0.8303 (0.0918)	0.8094	0.0209	0.000325
RGB histogram JSD	0.4973 (0.1318)	0.5880	0.0907	<0.00001

Note. n = 50. Positive advantage is matched minus mismatch cosine, or mismatch minus matched JSD. Directional p values use 200,000 one-to-one assignments and (b + 1)/(B + 1); both meet separate two-test Holm correction. BCa intervals are in the text and Figure 3.

Table 3. Genre-restricted matching checks.

Allowed group	Metric	Matched M	Other M	Advantage	Raw p	Holm p
Landscape/garden (30)	Cosine	0.8753	0.8628	0.0125	0.01053	0.03159
Landscape/garden (30)	RGB JSD	0.5021	0.6121	0.1100	<0.00001	0.00002
Still life (20)	Cosine	0.7629	0.7478	0.0151	0.08739	0.17478
Still life (20)	RGB JSD	0.4900	0.5109	0.0209	0.18137	0.18137
Pooled same genre	Cosine	0.8303	0.8168	0.0135	0.00652	0.00652
Pooled same genre	RGB JSD	0.4973	0.5716	0.0743	<0.00001	0.00002

Note. Other M is the mean of off-diagonal candidates in the allowed genre. Pooled n = 50 with fixed 30:20 genre weights. Positive advantages denote closer designated pairs. Four subgroup tests and two pooled tests have separate Holm families. Descriptive row-bootstrap intervals and exact seeds are in Supplementary File 1; permutation tests are the principal assignment checks. Different subgroup significance does not establish a genre interaction.

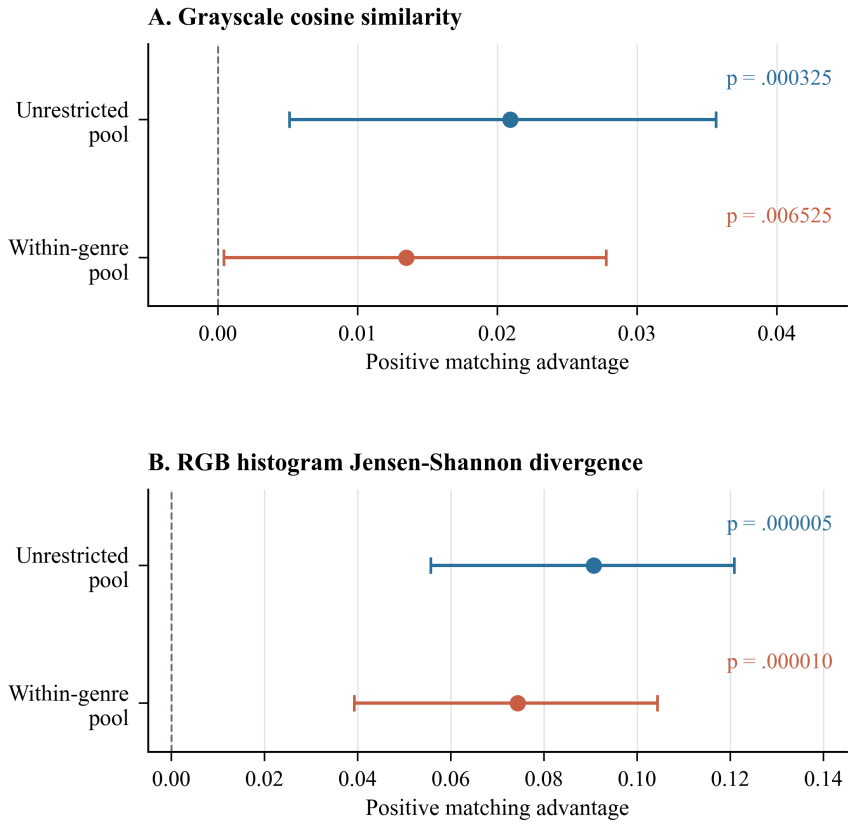


Figure 3. Pooled matching advantages before and after restricting candidate AI images to the same broad genre. Positive values indicate matched minus mismatched grayscale cosine, or mismatched minus matched RGB-histogram Jensen-Shannon divergence. The within-genre pool combines 30 landscape/garden pairs and 20 still-life pairs. Points are observed advantages and bars are marginal 95% BCa intervals from 100,000 row-advantage resamples (stratified with fixed genre weights for the restricted pool), conditional on the fixed candidate corpus. The displayed one-sided assignment-permutation *p* values use 200,000 one-to-one permutations with fixed points allowed and a plus-one correction. Bootstrap intervals and permutation tests concern different sampling schemes. Genre restriction attenuates both advantages, but residual pooled matching support remains.

3.2. Primary Paired Feature Differences

Four of the 13 visual features differed after Holm correction (**Figure 4** and **Table 4**). The DALL·E 2 images had higher mean green-channel intensity than the human images (human M = 0.3871, AI M = 0.4408; Δ = 0.0537, 95% BCa CI [0.0199, 0.0838], t(49) = 3.27, d_z = 0.46, Holm-adjusted p = 0.0196). They also had greater blue-channel dispersion (0.1726 vs. 0.2174; Δ = 0.0448 [0.0210, 0.0680], t(49) = 3.71, d_z = 0.52, adjusted p = 0.0064).

Grayscale entropy was higher in the AI images (7.087 vs. 7.328 bits; Δ = 0.241 bits [0.136, 0.362], t(49) = 4.17, d_z = 0.59, adjusted p = 0.0016). High-frequency energy was lower (0.1247 vs. 0.0942; Δ = −0.0305 [−0.0494, −0.0149], t(49) = −3.46, d_z = −0.49, adjusted p = 0.0125). Holm-adjusted Wilcoxon tests supported

the same four pair-level conclusions (adjusted p = 0.0113, 0.0074, 0.00039, and 0.0113, respectively).

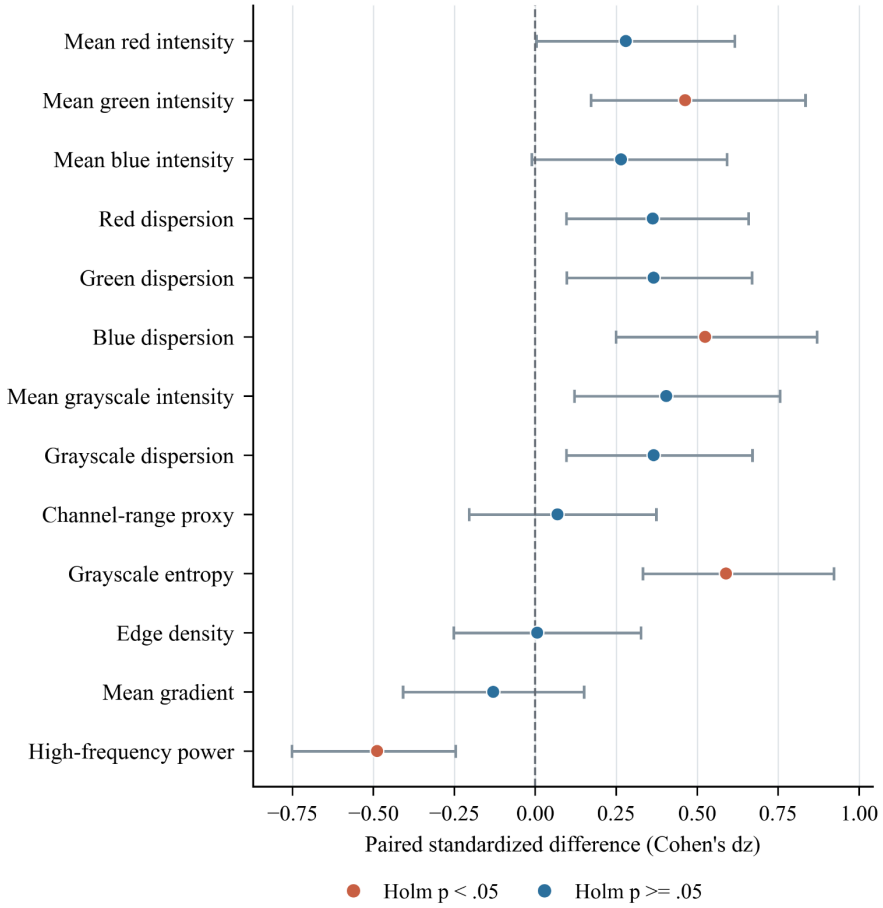


Figure 4. All 13 primary paired standardized differences (AI minus human; n = 50). Bars show 95% percentile bootstrap intervals for Cohen's d_z from 100,000 pair resamples, taken directly from the frozen analysis table. Orange points identify two-sided paired-t tests retaining support after Holm correction across all 13 features. Effect-size intervals are marginal; family-wise inference follows Holm-adjusted tests.

Table 4. Primary paired feature estimates for the official 50-pair corpus.

Feature	Human M (SD)	AI M (SD)	AI-human [95% BCa CI]	d_z	Holm p
red-intensity	0.436 (0.109)	0.467 (0.112)	+0.031 [−0.001, +0.061]	+0.28	0.268
green-intensity	0.387 (0.115)	0.441 (0.098)	+0.054 [+0.020, +0.084]	+0.46	0.020
blue-intensity	0.301 (0.112)	0.338 (0.098)	+0.037 [−0.003, +0.074]	+0.26	0.269
Red-channel contrast	0.205 (0.047)	0.229 (0.042)	+0.024 [+0.006, +0.043]	+0.36	0.102
Green-channel contrast	0.182 (0.052)	0.207 (0.047)	+0.026 [+0.007, +0.045]	+0.37	0.102
Blue-channel contrast	0.173 (0.068)	0.217 (0.051)	+0.045 [+0.021, +0.068]	+0.52	0.006
gray intensity	0.392 (0.106)	0.437 (0.094)	+0.045 [+0.013, +0.075]	+0.40	0.056
Gray-level contrast	0.182 (0.050)	0.207 (0.044)	+0.025 [+0.007, +0.044]	+0.37	0.102

Continued

Saturation	0.186 (0.072)	0.191 (0.090)	+0.005 [−0.017, +0.026]	+0.07	1.000
Gray-level entropy	7.087 (0.351)	7.328 (0.259)	+0.241 [+0.136, +0.362]	+0.59	0.002
Edge density	0.124 (0.079)	0.124 (0.074)	+0.000 [−0.020, +0.017]	+0.01	1.000
gradient	0.057 (0.020)	0.055 (0.020)	−0.002 [−0.007, +0.002]	−0.13	1.000
High-frequency energy	0.125 (0.088)	0.094 (0.060)	−0.031 [−0.049, −0.015]	−0.49	0.013

Note. n = 50. d_z is the paired standardized mean difference. Intervals are marginal BCa mean-difference intervals from 100,000 resamples. Holm correction covers all 13 two-sided paired t tests. Normalized intensity units are used except entropy (bits) and unitless frequency proportion.

The remaining nine features did not meet the corrected paired-t criterion. Mean grayscale intensity increased by 0.0451, but the adjusted value was p = 0.0558; the separately corrected Wilcoxon sensitivity test was p = 0.0415. Because these inferential routes straddled the threshold, grayscale intensity was not classified as a primary finding. Red- and green-channel contrast also had positive marginal intervals but adjusted p = 0.1021. Saturation, edge density, and mean gradient showed little pair-level evidence of a systematic shift at the primary resolution. Figure 5 shows the four primary difference distributions.

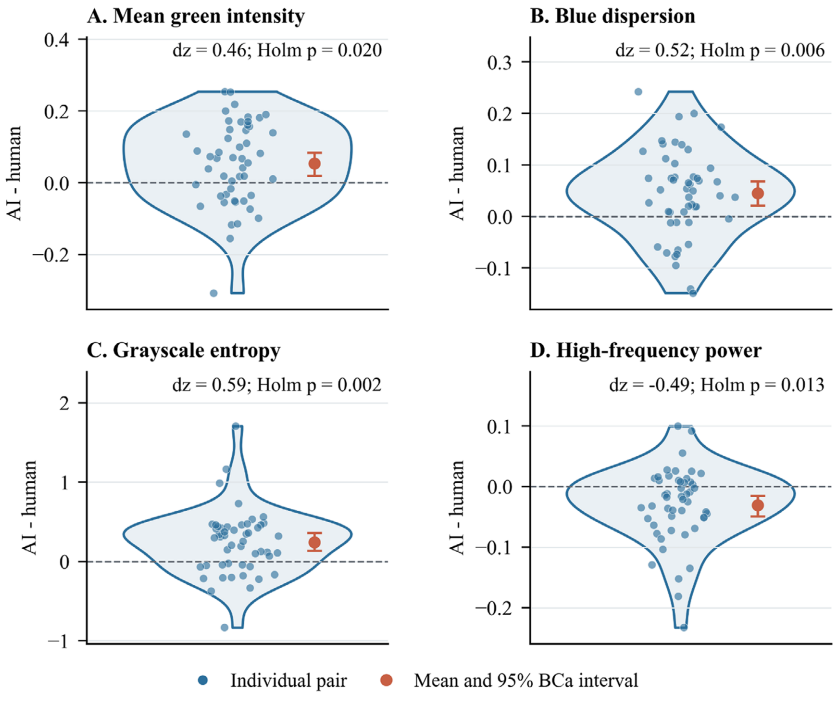


Figure 5. Distributions of paired differences for the four features retaining support in the primary 13-test Holm family (n = 50). Blue dots are individual AI-minus-human differences, with deterministic horizontal jitter; the violin is a descriptive density estimate. Orange points and bars show the mean and marginal 95% BCa intervals from the frozen 100,000-resample table. Green intensity, blue dispersion and high-frequency power use normalized image units; grayscale entropy uses bits. Artist-equal weighting retains corrected support for blue dispersion and entropy only.

3.3. Artist Weighting and Resolution Sensitivity

With equal artist weighting, blue-channel dispersion remained higher (Δ = 0.0510, 95% BCa CI [0.0225, 0.0826], Holm-adjusted $p = 0.0351$), as did grayscale entropy (Δ = 0.290 bits [0.173, 0.486], adjusted $p = 0.0107$). Green intensity and high-frequency energy did not meet the corrected threshold (adjusted $p = 0.0908$ and 0.0995). Their estimates, however, barely changed: green intensity was 0.0537 for pairs and 0.0525 for artist means; high-frequency energy was −0.0305 and −0.0302. Artist averaging changes weights and reduces the analysis units from 50 pairs to 28 artist means (Table S1). These p-value changes are compatible with reduced precision and altered analysis structure; they do not establish that sample composition caused the effects.

Three patterns—higher green-channel intensity, greater blue-channel dispersion, and higher grayscale entropy—survived Holm correction at 64, 128, 256, and 512 pixels. The frequency result was absent at 64 pixels but present at 128 pixels and above. Edge density and mean gradient changed inferential status and, for edge density, direction across resolutions. These scale-dependent quantities should not be read as model-invariant texture signatures. The complete heatmap is shown in Supplementary Figure S2. In a separately labeled post hoc grid, edge thresholds 0.08, 0.12, and 0.16 yielded no corrected effects. Frequency estimates remained negative at normalized cutoffs 0.15, 0.25, and 0.35, but 13-feature replacement-family adjusted p values were 0.1148, 0.0125, and 0.0017 (Table S4). Directional stability therefore coexists with threshold-sensitive statistical support.

3.4. Exploratory Genre Differences

None of the 13 contrasts between landscape/garden and still-life pairs survived Holm correction (Figure 6; Table S2). The largest adjusted contrast concerned green-channel dispersion, but it remained nonsignificant ($p = 0.282$). The category analysis therefore provides no corrected evidence that the observed low-level shifts differed by these two genre groupings.

3.5. External Lunara Stress Test

No feature in the 10-pair Lunara sample survived Holm correction across 13 tests. The smallest adjusted value was 0.087 for blue-channel dispersion. Several estimated directions differed from the primary art corpus: Lunara variants tended toward lower channel dispersion and lower entropy, whereas the DALL·E 2 corpus showed higher blue dispersion and entropy. The point-estimate pattern did not transfer consistently to this small sample. The analysis cannot isolate model, prompt, or domain effects and does not test a universal AI-image signature. Supplementary Figure S1 displays all effect estimates. Complete native-unit results are reported in Table S3.

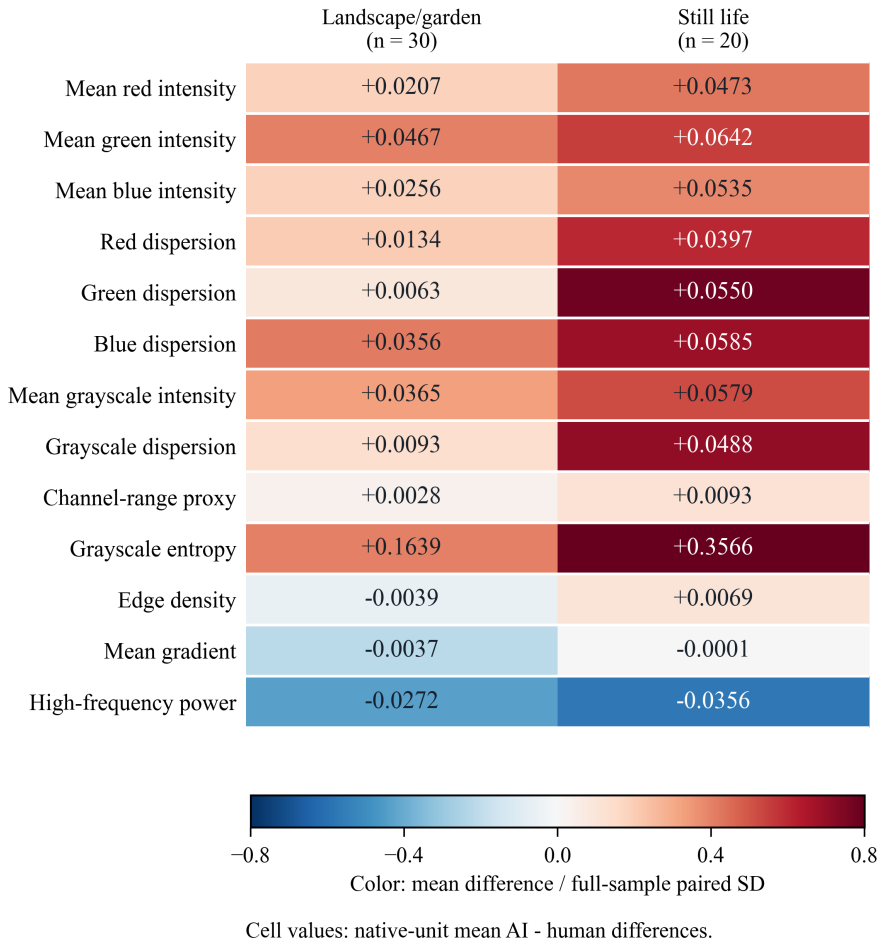


Figure 6. Descriptive genre-specific mean paired differences for all 13 features, copied from the frozen genre-results table. Cell numbers retain native feature units (entropy in bits; other normalized image descriptors). Shading divides each genre mean by the full 50-pair standard deviation of that feature's paired differences to permit a common visual scale; this is a descriptive normalization, not a within-genre effect-size estimate. These means do not establish a genre interaction; none of the original 13 between-genre tests survived Holm adjustment.

4. Discussion

4.1. Visual Differences within the Examined Corpus

The pair-level analysis identified selected differences in color and tonal statistics within this curated corpus. In the official van Hees corpus, style-matched DALL·E 2 images had higher green-channel intensity, greater within-image blue-channel dispersion, higher grayscale entropy, and less high-frequency energy at the primary 128 × 128 resolution. There was no comparable primary evidence for saturation, edge density, or mean gradient. The effect-size pattern therefore does not support a simple account in which AI images are uniformly more colorful, more detailed, or smoother.

Two effects deserve the greatest weight. Blue-channel dispersion and grayscale entropy remained significant when repeated examples from the same artist were

averaged before inference, and both were stable across all four analysis resolutions. They indicate wider variation in the blue channel and greater uncertainty in the grayscale histogram within the selected AI outputs. Entropy here is a histogram statistic. It does not index semantic richness, originality, compositional intelligence, or cultural complexity. Likewise, blue-channel dispersion cannot by itself establish more harmonious or expressive color use.

The entropy contrast should be read alongside the source study. In that study, van Hees et al. (2025) already calculated luminance, RMS contrast, and entropy a posteriori. Their supplementary figure shows an AI entropy distribution shifted toward higher values, directionally consistent with our paired estimate. That figure reports distributions and image values rather than a numerical paired effect with multiplicity correction. Controlling these measures left the image-wise preference-detection correlation virtually unchanged (reported r = 0.42, $p = 0.003$, compared with r = 0.42, $p = 0.002$ before adjustment). The present reanalysis estimates the paired entropy difference with uncertainty, controls the full 13-feature family, and checks artist weighting and resolution. The source supplement does not specify its entropy implementation, so absolute entropy values are not directly equated. These additions clarify stability within the same corpus; they do not establish that entropy explains preference or origin detection.

Evidence for green intensity and high-frequency energy was less conclusive under artist averaging, although their estimated magnitudes and directions changed little. Replacing 50 pair observations with 28 artist means changes both weighting and uncertainty. Loss of corrected significance cannot identify sample composition as the source of an effect. Frequency evidence additionally depended on resolution and the operational spectral cutoff. Balanced artist sampling and repeated outputs per prompt are needed for broader generalization.

4.2. Curation, Resemblance, and the Meaning of a Pair

Unrestricted mismatches yielded larger pooled advantages than same-genre mismatches, consistent with a contribution from broad genre matching. However, both pooled advantages retained corrected support when assignments were restricted within genre. The source curation therefore achieved resemblance beyond that explained by the two broad genre groups alone. Subgroup significance patterns do not establish a genre interaction, and still-life uncertainty must not be interpreted as evidence of absent matching. The source authors selected candidates using color, style, and composition, so the remaining advantage cannot be attributed to the generator alone.

This distinction changes the conceptual interpretation of "AI recreation". In the present design, recreation occurs at the level of artist-associated and genre-associated conventions. A controlled claim about transformation of a specific work would require image-to-image input, fixed generation parameters, all candidate outputs, and a documented counterfactual design. Without those elements, causal phrases such as "AI made the original brighter" are not warranted.

4.3. Relation to Human Judgments and Computational Aesthetics

The source behavioral study found above-chance preference for the selected AI images and above-chance origin discrimination (van Hees et al., 2025). The present analysis did not link image-level feature differences to participant-level responses. It therefore cannot claim that entropy, color dispersion, or spatial frequency caused preference or enabled detection. That question requires behavioral outcome data joined at the image level, mixed-effects models for participants and stimuli, and out-of-sample validation.

More broadly, empirical responses to AI art depend on labels, beliefs about agency, and assumptions about human creativity (Chiarella et al., 2022; Gangadharbatla, 2022; Horton et al., 2023; Millet et al., 2023). The current results complement that literature by describing image statistics while keeping aesthetic judgment outside the measurement model. This separation is necessary because formal regularity is only one component of art experience; prior knowledge, emotion, context, and attribution also shape evaluation (Chatterjee & Vartanian, 2014; Leder et al., 2004; Pelowski et al., 2016).

4.4. Model and Domain Specificity

The Lunara sample did not reproduce the primary directions. The difference in point estimates may reflect changes in generator, source domain, and contextual prompts; this heterogeneous ten-pair sample cannot isolate those explanations. The contrast shows why a single "AI look" should not be inferred from one model or corpus. Model architecture, training data, prompt language, candidate selection, content domain, output postprocessing, JPEG encoding, and analysis resolution can each alter low-level statistics.

The resolution analysis reinforces this point. Channel means and entropy were relatively stable, while edge and gradient measures were scale-sensitive. A fixed gradient threshold does not have the same perceptual meaning at every spatial scale. Future comparisons should preregister the working resolution, use scale-normalized edge operators, retain color profiles, and report multiscale texture estimates.

4.5. Authorship, Rights, and Responsible Reuse

Visual similarity does not resolve authorship or copyright. Generative systems raise questions about training data, attribution, artist consent, and the legal status of outputs that require institutional and jurisdiction-specific analysis (Epstein et al., 2023; Henderson et al., 2023; Samuelson, 2023). The present paper therefore distinguishes the article's CC BY license from the rights status of independently hosted source reproductions. A submission package can distribute code, manifests, hashes, and derived statistics while directing users to the official archive. High-resolution redistribution of each underlying human artwork should follow image-specific rights verification.

4.6. Limitations

This study has several limitations. The corpus was manually curated and is not a random sample of artworks or DALL·E 2 outputs. The pairs are artist/style matches, not before-and-after transformations, and unavailable candidate images prevent estimation of selection effects. Repeated artists and small genre groups limit precision; different subgroup p values do not demonstrate interaction. The 13 correlated low-level measures do not operationalize creativity, meaning, authenticity, or artistic value. Comparable final JPEG encoding does not establish comparable pre-archive color-management, resampling, compression, or export histories. The chosen resolution and thresholds also condition measurements. Row-bootstrap matching intervals are conditional on shared candidate images rather than independent superpopulation estimates. The ten-pair Lunara sample is heterogeneous, predominantly photographic, and context-manipulating. Finally, no semantic-embedding or participant-response model was fitted.

4.7. Future Work

A confirmatory extension should begin with license-verified, high-resolution human artworks and a factorial generation protocol that records the model checkpoint, full prompt, seed, sampler, guidance scale, and every candidate. Multiple current generators should produce repeated outputs for each prompt, with selection either eliminated or modeled explicitly. Multiscale low-level features can be combined with frozen semantic embeddings, and designated-pair similarity should always be compared against restricted random-pair baselines. Human evaluations should then test preference, origin detection, perceived creativity, and semantic fidelity using participant and artwork random effects. Preregistration, power analysis, held-out validation, and a public repository would convert this pilot framework into a stronger test of model, prompt, and genre interactions.

5. Conclusion

In this curated corpus of 50 pairs, DALL·E 2 images differed from human artworks in four of 13 features after Holm correction: mean green intensity, blue-channel dispersion, grayscale entropy, and high-frequency energy. Blue dispersion and entropy also met the corrected threshold with equal artist weighting. Green and frequency point estimates changed little under artist averaging, but evidence was less conclusive with 28 artist means. Frequency support also depended on scale and cutoff. Genre-restricted assignments attenuated, but did not eliminate, the pooled matching advantages. The external variation sample produced no corrected effects. These findings describe the selected corpus, generation procedure, and curation choices; they establish neither a transformation of individual artworks nor an explanation of artistic quality or preference.

Data and Code Availability

The primary materials are available at https://osf.io/n7w32/ and the exact archive

at https://osf.io/download/e2mr4/. Lunara is available at https://huggingface.co/datasets/moonworks/lunara-aesthetic-image-variations. Supplementary File 1 supplies the replication package, including stimulus manifests and hashes, derived features, JPEG audits, permutation results, analysis and figure scripts, software requirements, and manuscript source. Source images remain subject to their original terms and can be obtained through the cited repositories. No permanent public DOI has been assigned to this replication package.

Ethics Statement

This secondary computational study used existing publicly accessible image materials. It involved no recruitment, intervention, or collection of personal or participant-level data. Informed consent was not applicable to the present image-only analysis.

Generative AI Assistance

OpenAI Codex assisted with code generation, statistical workflow implementation, manuscript drafting and revision, figure preparation, and document formatting. The primary AI images were existing DALL·E 2 stimuli from the source study; no new artwork images were generated for this reanalysis. Responsibility for the study and the submitted content remains with the named authors.

Author Contributions

Yakun Gu: conceptualization, data curation, software, formal analysis, visualization, and writing of the original draft. Yiying Zhang: revise the manuscript, rewrite, repair and modify, and provide creative ideas for the paper. Tianling Huang: supervision, project administration, and review and editing of the manuscript.

Conflicts of Interest

The authors declare no competing interests.

References

Agudo, U., Arrese, M., Liberal, K. G., & Matute, H. (2022). Assessing Emotion and Sensitivity of AI Artwork. *Frontiers in Psychology, 13,* Article 879088. https://doi.org/10.3389/fpsyg.2022.879088

Bellaiche, L., Shahi, R., Turpin, M. H., Ragnhildstveit, A., Sprockett, S., Barr, N. et al. (2023). Humans versus AI: Whether and Why We Prefer Human-Created Compared to AI-Created Artwork. *Cognitive Research: Principles and Implications, 8,* Article 42. https://doi.org/10.1186/s41235-023-00499-6

Chamberlain, R., Mullin, C., Scheerlinck, B., & Wagemans, J. (2018). Putting the Art in Artificial: Aesthetic Responses to Computer-Generated Art. *Psychology of Aesthetics, Creativity, and the Arts, 12,* 177-192. https://doi.org/10.1037/aca0000136

Chatterjee, A., & Vartanian, O. (2014). Neuroaesthetics. *Trends in Cognitive Sciences, 18,* 370-375. https://doi.org/10.1016/j.tics.2014.03.003

Chiarella, S. G., Torromino, G., Gagliardi, D. M., Rossi, D., Babiloni, F., & Cartocci, G. (2022). Investigating the Negative Bias towards Artificial Intelligence: Effects of Prior Assignment of AI-Authorship on the Aesthetic Appreciation of Abstract Paintings. *Computers in Human Behavior, 137,* Article 107406. https://doi.org/10.1016/j.chb.2022.107406

Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2006). Studying Aesthetics in Photographic Images Using a Computational Approach. In A. Leonardis, H. Bischof, & A. Pinz (Eds.), *Computer Vision-ECCV 2006* (Vol. 3953, pp. 288-301). Springer. https://doi.org/10.1007/11744078_23

Dhar, S., Ordonez, V., & Berg, T. L. (2011). High Level Describable Attributes for Predicting Aesthetics and Interestingness. In *CVPR 2011* (pp. 1657-1664). IEEE. https://doi.org/10.1109/cvpr.2011.5995467

Elgammal, A., Liu, B., Elhoseiny, M., & Mazzone, M. (2017). CAN: Creative Adversarial Networks, Generating "Art" by Learning about Styles and Deviating from Style Norms. In *Proceedings of the Eighth International Conference on Computational Creativity* (pp. 96-103). Association for Computational Creativity. https://computationalcreativity.net/iccc2017/ICCC_17_accepted_submissions/ICCC-17_paper_47.pdf

Elgammal, A., Liu, B., Kim, D., Elhoseiny, M., & Mazzone, M. (2018). The Shape of Art History in the Eyes of the Machine. *Proceedings of the AAAI Conference on Artificial Intelligence, 32,* 2183-2191. https://doi.org/10.1609/aaai.v32i1.11894

Epstein, Z., Hertzmann, A., Akten, M., Farid, H., Fjeld, J., Frank, M. R. et al. (2023). Art and the Science of Generative AI. *Science, 380,* 1110-1111. https://doi.org/10.1126/science.adh4451

Gangadharbatla, H. (2022). The Role of AI Attribution Knowledge in the Evaluation of Artwork. *Empirical Studies of the Arts, 40,* 125-142. https://doi.org/10.1177/0276237421994697

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image Style Transfer Using Convolutional Neural Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2414-2423). IEEE. https://doi.org/10.1109/cvpr.2016.265

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 27, pp. 2672-2680). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2014/hash/f033ed80deb0234979a61f95710dbe25-Abstract.html

Graham, D. J., & Redies, C. (2010). Statistical Regularities in Art: Relations with Visual Coding and Perception. *Vision Research, 50,* 1503-1509. https://doi.org/10.1016/j.visres.2010.05.002

Henderson, P., Li, X., Jurafsky, D., Hashimoto, T., Lemley, M. A., & Liang, P. (2023). Foundation Models and Fair Use. *Journal of Machine Learning Research, 24,* 1-79. https://www.jmlr.org/papers/v24/23-0569.html

Hertzmann, A. (2018). Can Computers Create Art? *Arts, 7,* Article 18. https://doi.org/10.3390/arts7020018

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 6626-6637). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-

Abstract.html

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 6840-6851). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics, 6,* 65-70.

Hong, J. W., & Curran, N. M. (2019). Artificial Intelligence, Artists, and Art: Attitudes toward Artwork Produced by Humans vs. Artificial Intelligence. *ACM Transactions on Multimedia Computing, Communications, and Applications, 15,* 1-16. https://doi.org/10.1145/3326337

Horton Jr, C. B., White, M. W., & Iyengar, S. S. (2023). Bias against AI Art Can Enhance Perceptions of Human Creativity. *Scientific Reports, 13,* Article No. 19001. https://doi.org/10.1038/s41598-023-45202-3

Johnson, J., Alahi, A., & Li, F. F. (2016). Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer Vision-ECCV 2016* (Vol. 9906, pp. 694-711). Springer. https://doi.org/10.1007/978-3-319-46475-6_43

Karayev, S., Hertzmann, A., Trentacoste, M., Han, H., Winnemoeller, H., Agarwala, A. et al. (2014). Recognizing Image Style. In *Proceedings of the British Machine Vision Conference 2014* (pp. 122.1-122.11). British Machine Vision Association. https://doi.org/10.5244/c.28.122

Karras, T., Laine, S., & Aila, T. (2019). A Style-Based Generator Architecture for Generative Adversarial Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4396-4405). IEEE. https://doi.org/10.1109/cvpr.2019.00453

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and Improving the Image Quality of StyleGAN. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 8107-8116). IEEE. https://doi.org/10.1109/cvpr42600.2020.00813

Ke, Y., Tang, X., & Jing, F. (2006). The Design of High-Level Features for Photo Quality Assessment. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Vol. 1, pp. 419-426). IEEE. https://doi.org/10.1109/cvpr.2006.303

Lakens, D. (2013). Calculating and Reporting Effect Sizes to Facilitate Cumulative Science: A Practical Primer for T-Tests and ANOVAs. *Frontiers in Psychology, 4,* Article 863. https://doi.org/10.3389/fpsyg.2013.00863

Leder, H., Belke, B., Oeberst, A., & Augustin, D. (2004). A Model of Aesthetic Appreciation and Aesthetic Judgments. *British Journal of Psychology, 95,* 489-508. https://doi.org/10.1348/0007126042369811

Mazzone, M., & Elgammal, A. (2019). Art, Creativity, and the Potential of Artificial Intelligence. *Arts, 8,* Article 26. https://doi.org/10.3390/arts8010026

McCormack, J., Gifford, T., & Hutchings, P. (2019). Autonomy, Authenticity, Authorship and Intention in Computer Generated Art. In *Lecture Notes in Computer Science* (pp. 35-50). Springer International Publishing. https://doi.org/10.1007/978-3-030-16667-0_3

Mikalonytė, E. S., & Kneer, M. (2022). Can Artificial Intelligence Make Art: Folk Intuitions as to Whether AI-Driven Robots Can Be Viewed as Artists and Produce Art. *ACM Transactions on Human-Robot Interaction, 11,* 1-19. https://doi.org/10.1145/3530875

Millet, K., Buehler, F., Du, G., & Kokkoris, M. D. (2023). Defending Humankind: Anthro-

pocentric Bias in the Appreciation of AI Art. *Computers in Human Behavior, 143,* Article 107707. https://doi.org/10.1016/j.chb.2023.107707

Murray, N., Marchesotti, L., & Perronnin, F. (2012). AVA: A Large-Scale Database for Aesthetic Visual Analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2408-2415). IEEE. https://doi.org/10.1109/cvpr.2012.6247954

Nichol, A. Q., & Dhariwal, P. (2021). Improved Denoising Diffusion Probabilistic Models. In M. Meila, & T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 8162-8171). PMLR. https://proceedings.mlr.press/v139/nichol21a.html

Oksanen, A., Cvetkovic, A., Akin, N., Latikka, R., Bergdahl, J., Chen, Y. et al. (2023). Artificial Intelligence in Fine Arts: A Systematic Review of Empirical Research. *Computers in Human Behavior: Artificial Humans, 1,* Article 100004. https://doi.org/10.1016/j.chbah.2023.100004

Oliva, A., & Torralba, A. (2001). Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision, 42,* 145-175. https://doi.org/10.1023/a:1011139631724

Pelowski, M., Markey, P. S., Lauring, J. O., & Leder, H. (2016). Visualizing the Impact of Art: An Update and Comparison of Current Psychological Models of Art Experience. *Frontiers in Human Neuroscience, 10,* Article 160. https://doi.org/10.3389/fnhum.2016.00160

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models from Natural Language Supervision. In M. Meila, & T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 8748-8763). PMLR. https://proceedings.mlr.press/v139/radford21a.html

Ragot, M., Martin, N., & Cojean, S. (2020). AI-Generated vs. Human Artworks. A Perception Bias towards Artificial Intelligence? In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-10). ACM. https://doi.org/10.1145/3334480.3382892

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv*. https://doi.org/10.48550/arXiv.2204.06125

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-Shot Text-to-Image Generation. In M. Meila, & T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 8821-8831). PMLR. https://proceedings.mlr.press/v139/ramesh21a.html

Redies, C. (2015). Combining Universal Beauty and Cultural Context in a Unifying Model of Visual Aesthetic Experience. *Frontiers in Human Neuroscience, 9,* Article 218. https://doi.org/10.3389/fnhum.2015.00218

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 10674-10685). IEEE. https://doi.org/10.1109/cvpr52688.2022.01042

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. et al. (2022). Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Advances in Neural Information Processing Systems 35* (pp. 36479-36494). Neural Information Processing Systems Foundation, Inc. https://doi.org/10.52202/068431-2643

Saleh, B., & Elgammal, A. (2016). Large-Scale Classification of Fine-Art Paintings: Learning

the Right Metric on the Right Feature. *International Journal for Digital Art History, 2,* 70-93.

Samuelson, P. (2023). Generative AI Meets Copyright. *Science, 381,* 158-161. https://doi.org/10.1126/science.adi0656

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2021). Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*. https://openreview.net/forum?id=PxTIG12RRHS

Talebi, H., & Milanfar, P. (2018). NIMA: Neural Image Assessment. *IEEE Transactions on Image Processing, 27,* 3998-4011. https://doi.org/10.1109/tip.2018.2831899

Tan, W. R., Chan, C. S., Aguirre, H. E., & Tanaka, K. (2017). ArtGAN: Artwork Synthesis with Conditional Categorical GANs. In *2017 IEEE International Conference on Image Processing (ICIP)* (pp. 3760-3764). IEEE. https://doi.org/10.1109/icip.2017.8296985

Van Geert, E., & Wagemans, J. (2020). Order, Complexity, and Aesthetic Appreciation. *Psychology of Aesthetics, Creativity, and the Arts, 14,* 135-154. https://doi.org/10.1037/aca0000224

van Hees, J., Grootswagers, T., Quek, G. L., & Varlet, M. (2025). Human Perception of Art in the Age of Artificial Intelligence. *Frontiers in Psychology, 15,* Article 1497649. https://doi.org/10.3389/fpsyg.2024.1497469

Wang, Y., Hassan, P., Sadeka, S., Soliman, N., Abdullah, S., & Hassan, S. (2026). Moon-works Lunara Aesthetic II: An Image Variation Dataset. *arXiv*. https://doi.org/10.48550/arXiv.2602.01666

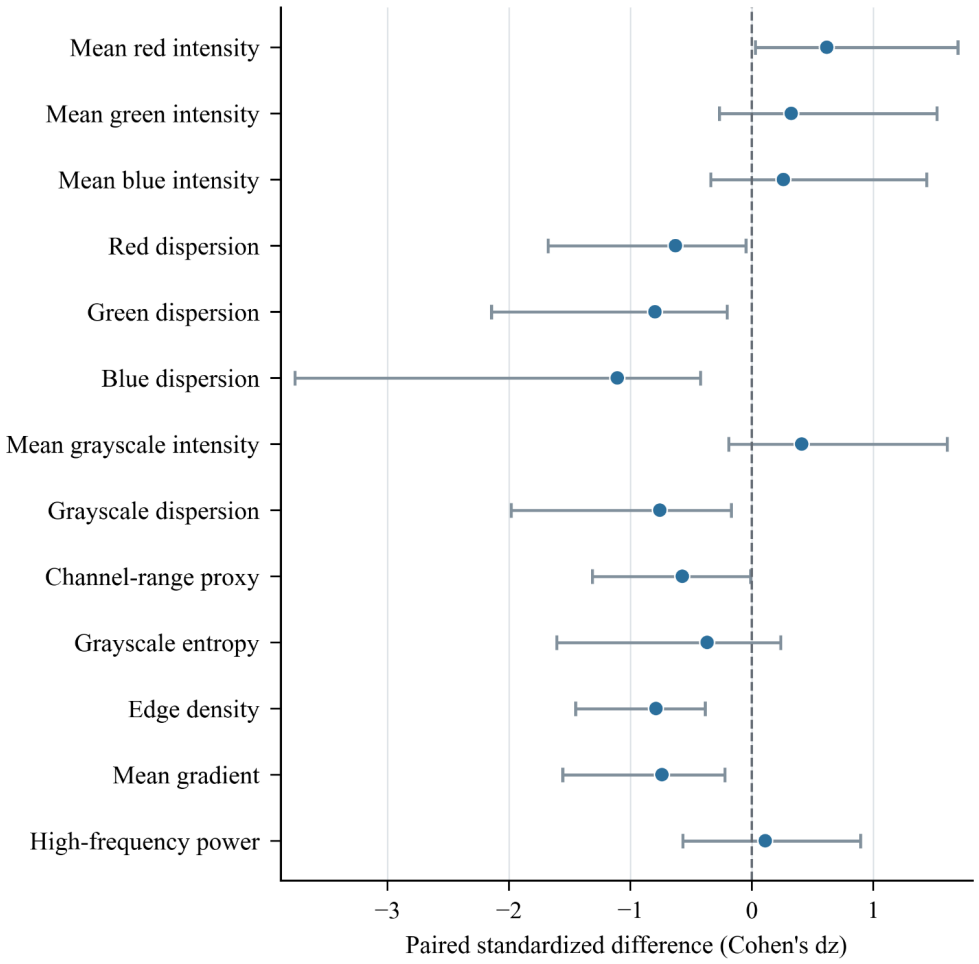
Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing, 13,* 600-612. https://doi.org/10.1109/tip.2003.819861

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A. et al. (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data, 3,* Article 160018. https://doi.org/10.1038/sdata.2016.18

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 586-595). IEEE. https://doi.org/10.1109/cvpr.2018.00068

Supplementary Analyses

Figure S1 reports the external Lunara effect estimates. Figure S2 reports multi-resolution effects. Tables S1-S3 give artist, genre, and external-sample estimates; Table S4 reports the post hoc threshold grid. Machine-readable results and the complete file-encoding audit accompany Supplementary File 1.



Separate contrast set: n = 10; all Holm p >= .05.

Figure S1. Separate ten-pair Lunara contrast-set analysis. Points are paired standardized differences (AI minus original anchor), with 95% percentile bootstrap effect-size intervals from the frozen 100,000-resample analysis. None of the 13 tests retained support after Holm correction. These anchors include non-art photographic scenes, are not pooled with the art corpus, and do not constitute direct replication of the DALL-E 2 art comparison.

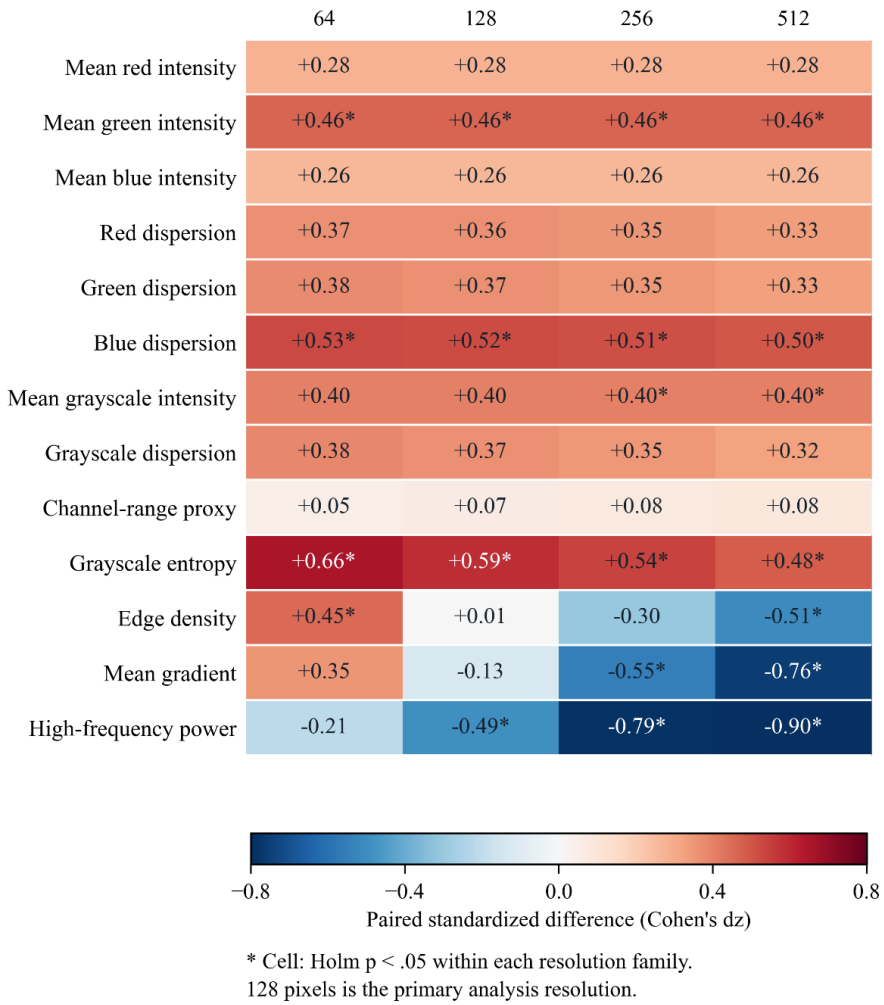


Figure S2. Resolution sensitivity at square image sizes of 64, 128, 256 and 512 pixels, using the frozen resolution table. Cells show paired standardized differences (Cohen's d_z; AI minus human; n = 50). Asterisks inside cells identify Holm-adjusted $p < 0.05$ within the 13-feature family at each resolution. The 128-pixel column is the original primary analysis. Frequency masks are relative to each resolution's Nyquist scale, so their physical image-scale interpretation changes across resizing.

Table S1. Equal artist weighting with 28 artist means.

Feature	Mean difference	95% BCa CI	Holm p
Mean red-channel intensity	+0.038	[+0.004, +0.069]	0.1827
Mean green-channel intensity	+0.052	[+0.015, +0.086]	0.0908
Mean blue-channel intensity	+0.039	[−0.007, +0.083]	0.4255
Red-channel contrast	+0.030	[+0.012, +0.055]	0.0995
Green-channel contrast	+0.029	[+0.008, +0.057]	0.1827
Blue-channel contrast	+0.051	[+0.023, +0.083]	0.0351
Mean gray intensity	+0.047	[+0.011, +0.080]	0.1130
Gray-level contrast	+0.030	[+0.010, +0.057]	0.1339
Saturation	+0.002	[−0.023, +0.027]	1.0000

Continued

Gray-level entropy	+0.290	[+0.173, +0.486]	0.0107
Edge density	+0.005	[−0.017, +0.024]	1.0000
Mean gradient	−0.001	[−0.006, +0.005]	1.0000
High-frequency energy	−0.030	[−0.054, −0.011]	0.0995

Note. Each artist receives one unit after averaging its pair differences. This changes both weighting and precision relative to 50 pair observations. Marginal intervals are not simultaneous intervals.

Table S2. Exploratory differences between broad genre groups.

Feature	Landscape/garden Δ	Still life Δ	Between-group Δ	Holm *p*
Mean red-channel intensity	+0.021	+0.047	−0.027	1.000
Mean green-channel intensity	+0.047	+0.064	−0.018	1.000
Mean blue-channel intensity	+0.026	+0.053	−0.028	1.000
Red-channel contrast	+0.013	+0.040	−0.026	1.000
Green-channel contrast	+0.006	+0.055	−0.049	0.282
Blue-channel contrast	+0.036	+0.059	−0.023	1.000
Mean gray intensity	+0.037	+0.058	−0.021	1.000
Gray-level contrast	+0.009	+0.049	−0.039	0.679
Saturation	+0.003	+0.009	−0.006	1.000
Gray-level entropy	+0.164	+0.357	−0.193	1.000
Edge density	−0.004	+0.007	−0.011	1.000
Mean gradient	−0.004	−0.000	−0.004	1.000
High-frequency energy	−0.027	−0.036	+0.008	1.000

Note. Landscape/garden n = 30; still life n = 20. The between-group contrast is landscape/garden minus still life. Welch tests are Holm-corrected across 13 contrasts; no contrast met 0.05. Full intervals are supplied in the replication data.

Table S3. External Lunara variation estimates in ten convenience-sampled pairs.

Feature	Anchor M	Variant M	Difference [95% BCa CI]	d_z	Holm *p*
Mean red intensity	0.404	0.518	+0.114 [−0.004, +0.216]	+0.62	0.625
Mean green intensity	0.456	0.519	+0.062 [−0.071, +0.160]	+0.33	1.000
Mean blue intensity	0.482	0.529	+0.047 [−0.080, +0.138]	+0.26	1.000
Red-channel dispersion	0.241	0.176	−0.065 [−0.127, −0.005]	−0.63	0.625
Green-channel dispersion	0.241	0.160	−0.081 [−0.137, −0.018]	−0.80	0.393
Blue-channel dispersion	0.282	0.166	−0.116 [−0.166, −0.039]	−1.11	0.087
Mean grayscale intensity	0.444	0.520	+0.076 [−0.050, +0.171]	+0.41	1.000
Grayscale dispersion	0.239	0.163	−0.076 [−0.133, −0.015]	−0.76	0.402
Channel-range proxy	0.166	0.104	−0.062 [−0.132, −0.004]	−0.57	0.625
Grayscale entropy	7.274	7.057	−0.216 [−0.512, +0.192]	−0.37	1.000
Edge density	0.143	0.088	−0.055 [−0.103, −0.020]	−0.79	0.393
Mean gradient	0.058	0.040	−0.018 [−0.033, −0.005]	−0.74	0.402
High-frequency energy	0.098	0.108	+0.010 [−0.042, +0.065]	+0.11	1.000

Note. This sample contains mostly photographs with generated contextual variants. It is separate from the primary corpus and is not a direct replication. No feature met the 13-test Holm threshold.

Table S4. Post hoc checks of edge and spectral thresholds.

Feature	Cutoff	Mean Δ	Raw *p*	Holm 13	Holm 6
edge density	0.08	−0.0060	0.66086	1.00000	1.00000
edge density	0.12	+0.0004	0.96756	1.00000	1.00000
edge density	0.16	+0.0023	0.71753	1.00000	1.00000
high frequency energy	0.15	−0.0332	0.01838	0.11483	0.07353
high frequency energy	0.25	−0.0305	0.00114	0.01251	0.00569
high frequency energy	0.35	−0.0262	0.00014	0.00172	0.00086

Note. Each Holm 13 test replaces only the chosen threshold in the existing 13-feature family; Holm 6 adjusts the six grid tests. These are exploratory sensitivity checks, not a new primary family. The frequency cutoff is normalized DFT radius divided by N/2, not cycles per pixel.

Feature Definitions

RGB means and dispersion. Each channel is normalized to [0, 1]. Means describe channel level; population standard deviations describe within-image dispersion.

Grayscale intensity and contrast. $Y' = 0.299R + 0.587G + 0.114B$ is computed from decoded normalized RGB codes. Its mean and population standard deviation are intensity summaries, not physical luminance.

Saturation proxy. Mean pixel-wise maximum minus minimum RGB channel; this is an operational chromatic-spread proxy.

Gray-level entropy. Shannon entropy in bits of the 256-bin eight-bit grayscale histogram; it measures marginal tonal dispersion rather than spatial or semantic complexity.

Local transitions. First differences are zero-padded in the last row or column. Mean gradient is threshold-free; edge density uses magnitude greater than 0.12.

Frequency energy. Image-mean subtraction precedes the centered Fourier transform. Power beyond radial DFT radius $0.25 \times (N/2)$ forms the numerator. At 128 pixels this begins beyond radius 16 bins, or 0.125 cycles per pixel.

Matching measures. Grayscale cosine uses flattened intensity arrays. RGB JSD uses normalized $8 \times 8 \times 8$ channel histograms. Neither measure is a semantic embedding.

Reproducibility Information

The primary archive is identified by its official URL and per-file SHA-256 hashes. Human stimulus i is designated to AI stimulus i + 50. Pixels are not independent analysis units. Primary feature families contain all 13 outcomes, and the new four subgroup and two pooled pairing tests have separately stated Holm families. Fixed seeds, exact resample counts, software versions, feature CSVs, image audits, and plotting inputs are retained in Supplementary File 1. The package supports a derived-data verification without image downloads and documents the additional downloads needed to repeat pixel extraction.