Multimodal LLMs for Visual Style Evaluation*

Ruben Pascual[1 [0009−0002−5250−8554]], Izaskun Aguirre[1 [0009−0007−8330−0779]], Mikel Sesma-Sara[1 [0000−0002−2949−7909]], Aranzazu Jurio[1 [0000−0002−4087−586X]], Daniel Paternain[1 [0000−0002−5845−887X]], and Mikel Galar[1 [0000−0003−2865−6549]]

(ISC), Dept. of Statistics, Computer Science and Mathematics, Public University of Navarre (UPNA), Campus Arrosadia, 31006, Pamplona, Spain
{ruben.pascua,aguirre.47152,mikel.sesma,
aranzazu.jurio,daniel.paternain,mikel.galar}@unavarra.es

Abstract. As AI-driven image generation advances, assessing whether generated images plausibly conform to a target artistic style remains a fundamental and largely subjective challenge. There are no established objective metrics for measuring style consistency, and existing proxies often fail to capture artistic nuance, leaving human evaluation as the primary means of assessment. While effective, human evaluation is costly, time-consuming, and difficult to scale. This work investigates the viability of multimodal large language models (MLLMs) as proxies for human evaluators in assessing style-consistent character generation. Following the human evaluation protocols and benchmark datasets introduced in prior work [34], we compare judgments from four state-of-the-art multimodal models (Gemini 2.5 Pro, Grok-4 Fast, OpenAI GPT-4.1, and GPT-5) to human evaluations collected from both general participants and professional artists. Our results reveal a significant discrepancy between ranking-level and response-level agreement. While MLLMs demonstrate high correlation with human rankings (Kendall's $\tau_b \approx 1.0$ globally), effectively identifying the best-performing methods, they exhibit low agreement on individual responses (Cohen's $\kappa < 0.34$). We identify systematic biases driving this divergence: MLLMs tend to avoid *both* or *neither* options in pairwise comparisons and inflate scores in absolute ratings. These findings suggest that while MLLMs are reliable for coarse-grained method ranking, they do not yet replicate the fine-grained artistic judgments of human evaluators.

Keywords: Generative AI · Multimodal LLMs · Style Consistency Assessment · Human-AI Alignment · Image Quality Assessment

* Rubén Pascual received support from an FPU grant (Formación de Profesorado Universitario, ref. FPU22/02961) awarded by the Spanish Ministry of Science, Innovation and Universities. This work was also partially supported by the Spanish Ministry of Science, Innovation and Universities under project PID2022-136627NB-I00 (MCIN/AEI/10.13039/501100011033/FEDER, UE) and project PID2025-174319OB-I00 (MICIU/AEI/10.13039/501100011033) and by Gobierno de Navarra under project PC24-GENESIS-041-002.

1 Introduction

Artificial Intelligence is increasingly transforming the audiovisual industry across multiple stages of production. In film and television, AI is being applied to video editing [26], rendering [47], and visual effects [39]. In other parts of the industry, like video games, deep learning-based upscaling methods like NVIDIA DLSS [27] have become standard for real-time rendering. AI has even enabled de-aging effects [9] and the digital resurrection of actors for posthumous film appearances [40]. Among these applications, AI-driven image generation has gained particular attention, especially in contexts requiring the creation of multiple images that adhere to a consistent artistic style [41], a capability that is central to animation, game development, and digital content creation, where maintaining visual coherence across characters, scenes, or assets is essential for artistic integrity and user experience.

While recent advances have improved the ability to generate style-consistent images, evaluating whether generated outputs truly maintain the desired artistic style remains a significant challenge. Commonly used objective metrics, such as Fréchet Inception Distance (FID) [18] and CLIP-based similarity scores [17], provide useful quantitative proxies but have well-known limitations in this context. In particular, they struggle to reliably assess stylistic consistency, often conflating style with semantic content or requiring large sample sizes to produce stable estimates. As a result, human evaluation remains one of the most reliable approaches for assessing subjective qualities such as style consistency.

Despite being the most reliable approach, human evaluation also presents some limitations. In particular, human judgments can be inconsistent due to subjective differences between evaluators, fatigue effects, and varying levels of expertise [33]. In addition, conducting large-scale human studies is time-consuming and expensive [11], requiring substantial effort to recruit participants, design evaluation protocols, and collect responses. These limitations become particularly problematic when iterative evaluation is needed during model development, where rapid feedback is essential. As a result, there is growing interest in automated alternatives that can approximate human judgment while maintaining efficiency and scalability.

Large language models (LLMs) have emerged as a promising solution for automated evaluation. Recent advances in multimodal LLMs (MLLMs), such as GPT [30, 31], Gemini [15], and others [1, 46], have demonstrated remarkable capabilities in understanding and reasoning about visual content. These models have shown competitive performance across diverse tasks including image captioning, visual question answering, and aesthetic assessment [25]. This raises a natural question: can MLLMs serve as reliable substitute evaluators for assessing the quality and style consistency of AI-generated images?

In this work, we investigate whether MLLMs can effectively replicate human judgments in evaluating AI-generated images. Our study uses the human evaluation framework and benchmark datasets introduced in [34], which were originally designed to assess style-consistent character generation through comprehensive human evaluation involving both general participants and professional artists.

By presenting the same images and evaluation tasks to multiple state-of-the-art MLLMs, we directly compare their responses with those of human evaluators and quantify agreement through established statistical metrics including Kendall's τ_b for ranking correlation and weighted Cohen's kappa for rating consistency.

2 Related Work

2.1 Automated Metrics for Style Consistency

Evaluating AI-generated images has traditionally relied on objective metrics like FID [18], IS [37], LPIPS [48], and CLIP-based scores [17]. However, these metrics face significant limitations when assessing artistic style. For example, distribution-based metrics like FID require large sample sizes, limiting few-shot applicability, while semantic metrics like CLIP-score are content-heavy and struggle to isolate style from subject matter.

To address this, specialized methods have been developed to quantify artistic style. Approaches like ALADIN [36] and recent diffusion-based similarity measures [42] focus on stylistic coherence. Similarly, work on content-style disentanglement, such as GOYA [45] and Only-Style [2], aims to isolate stylistic elements for more precise evaluation. Nevertheless, quantifying subjective artistic qualities remains challenging, as automated metrics may not fully align with human perceptions of style consistency and aesthetic coherence.

2.2 Human Evaluation in Generative AI

Human evaluation remains the primary reference for assessing generative models, particularly for subjective qualities that automated metrics cannot fully capture [21]. Common methodologies include absolute rating scales and pairwise comparisons. The latter has gained popularity due to its reliability with non-expert evaluators [19] and successful adoption in large-scale benchmarking platforms like Chatbot Arena [5].

However, human evaluation faces significant methodological hurdles. Besides being costly and time-consuming [11, 33], evaluation protocols are often poorly standardized; studies frequently lack comprehensive demographic reporting [20], and very few achieve alignment between problem definition and evaluation design [16]. Researchers have proposed frameworks like ConSiDERS [10] and hierarchical assessment protocols [3] to mitigate these issues, but the scalability bottleneck motivates the search for automated proxies that can approximate human judgment.

2.3 Large Language Models as Evaluators

Recent advances in LLMs have led to increasing interest in their use as alternatives to human evaluators. In text evaluation, LLMs have demonstrated strong performance in assessing bug reports [24], educational content [38], and creative writing, often showing higher consistency than human raters [23].

At the same time, prior work has identified limitations in using LLMs as judges. These models often exhibit biases, such as favoring their own outputs or specific structures [44], and may fail to detect subtle quality degradation [8].

Most existing studies have focused on text-based tasks. While recent efforts have begun exploring MLLMs for image quality assessment in e-commerce [12], their ability to replicate human nuance in evaluating artistic style remains unexplored.

3 Methodology

3.1 Human Evaluation Setup

Our study relies on a previously established human evaluation benchmark introduced in [34], which was designed to assess the generation of novel characters that maintain the artistic style of a small set of human-designed reference characters. The evaluation was conducted on five specialized datasets (Virus [29], Scary [28], Daily [13], Hipster [14], and Trans [43]), each containing 10-30 reference images with distinct artistic styles. For visual examples of the reference datasets and generated characters, we refer the reader to [34].

To assess character quality, the reference study compared a control condition from the original datasets against three generation methods: Method A (DB^{L}_{MT-C}) uses clustering-based multi-token selection; Method B (DB^{L}_{MT-R}) employs rarest-token selection; and Method C (DB^{L}) follows the standard DreamBooth baseline with LoRA. We reuse the human evaluation data collected in that study (up to December 1, 2025) as the benchmark for MLLM comparison.

- **General participant evaluation:** A total of 134 participants performed pairwise comparisons. For each comparison, participants were shown four reference images from the dataset as a visual guide, followed by two generated images presented side-by-side. Participants were asked to select which image better fit as a new member of the character family in terms of style and characteristics, with response options including *A is better*, *B is better*, *both fit equally*, or *neither fits*. Each participant completed six comparisons per dataset across all five datasets, resulting in 4020 total pairwise comparisons. Results were aggregated using the Bradley-Terry model [4] to obtain global method rankings.
- **Professional artist evaluation:** Six professional artists provided absolute ratings on individual generated images. Using the same visual setup with four reference images, artists rated single generated image on a 1-5 scale (1 = does not fit at all, 5 = could pass as another character of the family). Each artist evaluated 10 images per dataset, yielding 300 expert ratings that were averaged to produce method-level scores.

3.2 MLLM Evaluation Protocol

To assess whether MLLMs can serve as substitute evaluators for human judgments, we queried four state-of-the-art MLLMs (Gemini 2.5 Pro, Grok-4 Fast,

OpenAI GPT-4.1, and OpenAI GPT-5) using the same evaluation tasks employed in the human study. For each evaluation instance, we used exactly the same set of images shown to human participants (i.e., the same four reference images and the same candidate image(s), with identical pairing and ordering). The only modification was the textual instruction, which was adapted to elicit only the final label or score from the model.

We evaluate MLLMs using two task formats, mirroring the human protocol:

- **Pairwise comparisons:** Given four reference images and two candidate images (A and B), the MLLM was instructed to select one of the following options: *A is better*, *B is better*, *Both fit equally*, or *Neither fits*. The exact prompt is shown below (image inputs are represented as placeholders).

> Respond only with the final answer. Do not show your reasoning or intermediate steps. These four images belong to the same family of images:
>
> <<REFERENCE_IMAGES>>
>
> I present you Image A: <<IMAGE_A>>
>
> And Image B: <<IMAGE_B>>
>
> Which of the two characters fits better as a new member of the family (in terms of the style and characteristics)? Answer one of these options: A is better, B is better, Both fit equally, or Neither fits. Please don't give reasoning. Only one of the options.

- **Rating tasks:** Given four reference images and one candidate image (A), the MLLM was instructed to produce a score from 1 to 5 indicating how well the generated image matches the family style. The exact prompt is shown below (image inputs are represented as placeholders).

> Respond only with the final answer. Do not show your reasoning or intermediate steps. These four images belong to the same family of images:
>
> <<REFERENCE_IMAGES>>
>
> I present you Image A: <<IMAGE_A>>
>
> Rate from 1 to 5 how well Image A integrates as part of the presented family: 1 means it does not fit at all, 5 means it could pass as another family character. Don't give reasoning.

In addition to the individual responses, we also report results for an ensemble formed by combining responses from all participating models. For ratings, the ensemble score is computed as the average of the individual model scores. For pairwise comparisons, we used a voting-based aggregation: each model contributes one vote to its selected option (A or B), *Both* responses contribute one vote to both A and B, and *Neither* contributes no votes. The ensemble selects A or B if that option receives at least 2 votes and strictly more than the alternative; otherwise, it outputs *Both* (if both receive at least two equal votes) or *Neither*.

3.3 Metrics

To quantify the agreement between MLLM and human evaluations, we employ two complementary metrics that capture agreement at different levels: ranking-level and response-level:

- **Kendall's tau-b (τ_b) for ranking-level agreement**: We use Kendall's tau-b (τ_b) [22] to measure the correlation between the global method rankings produced by humans and MLLMs. Kendall's tau-b is defined as:

$$\tau_b = \frac{C - D}{\sqrt{(C + D + T_x)(C + D + T_y)}},$$

  where C is the number of concordant pairs, D is the number of discordant pairs, T_x and T_y are ties in each ranking. Values range from -1 to $+1$, with higher values indicating stronger agreement.
- **Cohen's kappa for response-level agreement**: We measure agreement on individual responses using Cohen's kappa [6,7]. For pairwise comparisons, we report the standard (unweighted) kappa (κ), which captures whether humans and MLLMs made the same choice (*A, B, both or neither*) on each comparison. For ratings, we report the weighted version (κ_w) to account for partial agreement between nearby scores. Weighted Cohen's kappa is defined as:

$$\kappa_w = 1 - \frac{\sum_{i,j} w_{ij} o_{ij}}{\sum_{i,j} w_{ij} e_{ij}},$$

  where o_{ij} is the observed agreement, e_{ij} is expected agreement by chance, and w_{ij} are quadratic weights: $w_{ij} = 1 - \frac{(i-j)^2}{(k-1)^2}$. This accounts for partial agreement between nearby scores and assigns lower penalties to small rating differences.

4 Results

4.1 Ranking-level agreement

Tables 1 and 2 report the method rankings derived from pairwise comparisons and rating-based evaluations, respectively, for both human evaluators and

MLLMs. Each table shows rankings and scores for the control condition (original dataset images) and three evaluated methods (Method A, B, and C) across five datasets, as well as a global aggregate.

Table 1. Bradley-Terry score-based rankings from pairwise comparisons across datasets for human evaluators, MLLMs, and the ensemble (Control + three methods; higher scores indicate stronger preference).

Evaluator	Method	Virus	Scary	Daily	Hipster	Trans	Global
Human	Control	2nd (1064.85)	1st (1089.44)	1st (1072.76)	1st (1107.78)	2nd (1042.50)	1st (1072.71)
	Method A	1st (1066.03)	2nd (1057.66)	2nd (1007.69)	2nd (1053.76)	1st (1050.50)	2nd (1044.21)
	Method B	3rd (981.86)	3rd (956.22)	3rd (978.28)	3rd (1007.11)	3rd (1042.06)	3rd (993.97)
	Method C	4th (887.26)	4th (896.68)	4th (941.26)	4th (831.34)	4th (864.94)	4th (889.10)
Gemini 2.5 Pro	Control	2nd (959.31)	1st (1355.54)	1st (1244.19)	1st (1124.07)	1st (1124.50)	1st (1086.08)
	Method A	1st (1220.24)	2nd (967.49)	4th (881.63)	3rd (1021.59)	2nd (1024.66)	2nd (1018.01)
	Method B	3rd (920.73)	4th (810.04)	3rd (899.64)	2nd (1038.98)	3rd (1018.06)	3rd (973.65)
	Method C	4th (899.71)	3rd (866.94)	2nd (974.54)	4th (815.37)	4th (832.78)	4th (922.26)
Grok-4 Fast	Control	3rd (979.18)	1st (1125.20)	1st (1044.74)	2nd (1016.27)	1st (1079.64)	2nd (1026.23)
	Method A	1st (1085.13)	2nd (1043.49)	2nd (1015.95)	1st (1069.01)	3rd (976.33)	1st (1039.09)
	Method B	2nd (1051.32)	4th (874.57)	4th (969.39)	3rd (1006.81)	2nd (988.67)	3rd (986.94)
	Method C	4th (884.37)	3rd (956.73)	3rd (969.91)	4th (907.91)	4th (955.36)	4th (947.75)
GPT-4.1	Control	3rd (950.25)	1st (1200.43)	1st (1352.33)	1st (1146.61)	1st (1138.47)	1st (1074.76)
	Method A	1st (1146.56)	2nd (1036.56)	3rd (883.42)	3rd (993.32)	2nd (1010.66)	2nd (1024.21)
	Method B	2nd (993.87)	4th (877.23)	4th (871.79)	2nd (997.62)	3rd (988.00)	3rd (980.52)
	Method C	4th (909.31)	3rd (885.79)	2nd (892.45)	4th (862.44)	4th (862.87)	4th (920.52)
GPT-5	Control	3rd (935.37)	1st (1430.88)	1st (1380.00)	1st (1258.73)	1st (1190.71)	1st (1102.92)
	Method A	1st (1207.42)	2nd (987.09)	4th (860.16)	2nd (955.47)	2nd (1008.25)	2nd (1020.80)
	Method B	2nd (982.79)	3rd (805.08)	3rd (862.61)	3rd (946.80)	3rd (971.55)	3rd (971.41)
	Method C	4th (874.42)	4th (776.95)	2nd (897.22)	4th (839.00)	4th (829.50)	4th (904.88)
Ensemble	Control	3rd (933.87)	1st (1370.25)	1st (1361.86)	1st (1195.55)	1st (1166.48)	1st (1090.85)
	Method A	1st (1212.77)	2nd (994.35)	4th (865.28)	2nd (1010.00)	2nd (1016.84)	2nd (1029.96)
	Method B	2nd (973.04)	4th (801.05)	3rd (866.29)	3rd (991.30)	3rd (984.66)	3rd (972.23)
	Method C	4th (880.32)	3rd (834.35)	2nd (906.57)	4th (803.15)	4th (832.01)	4th (906.95)

For pairwise comparisons (Table 1), human evaluators consistently ranked the control condition in first or second position across datasets, with Method A generally occupying first or second place, followed by Method B, and Method C in last place. MLLM evaluators largely replicate this overall pattern, with the control condition ranked first in most datasets and achieving first place in the global ranking across all models. Some variability is observed at the dataset level, particularly for the Daily and Virus datasets, but converge at the global level where all MLLMs match the human ranking order.

In rating-based evaluations (Table 2), a similar pattern is observed. Human evaluators ranked the control condition first across all datasets except Trans, where Method A achieved the highest average rating. All MLLMs consistently placed the control condition first, with Method A typically in second place. Ensemble predictions closely align with individual MLLM rankings in both evaluation types.

To quantify ranking agreement, we compute Kendall's τ_b between human and MLLM rankings (Tables 3 and 4). For pairwise comparisons (Table 3), Kendall's tau values show strong ranking correlation at the global level, with Gemini 2.5

Table 2. Average rating-based rankings across datasets for human evaluators, MLLMs, and the ensemble (Control + three methods; higher mean ratings indicate a better fit).

Evaluator	Method	Virus	Scary	Daily	Hipster	Trans	Global
Human	Control	1st (4.15)	1st (4.25)	1st (4.83)	1st (4.25)	2nd (4.29)	1st (4.32)
	Method A	2nd (3.88)	2nd (3.81)	2nd (3.44)	2nd (3.70)	1st (4.37)	2nd (3.84)
	Method B	3rd (2.92)	3rd (2.52)	3rd (3.29)	3rd (3.27)	3rd (3.79)	3rd (3.12)
	Method C	4th (2.83)	4th (2.20)	4th (2.72)	4th (2.35)	4th (2.00)	4th (2.45)
Gemini 2.5 Pro	Control	1st (4.38)	1st (4.50)	1st (4.67)	1st (4.88)	1st (4.71)	1st (4.61)
	Method A	2nd (3.12)	4th (2.25)	4th (2.56)	2nd (4.25)	2nd (4.58)	2nd (3.42)
	Method B	4th (2.75)	3rd (2.36)	3rd (2.59)	4th (3.80)	3rd (4.00)	4th (3.06)
	Method C	3rd (2.94)	2nd (2.47)	2nd (3.28)	3rd (4.00)	4th (3.21)	3rd (3.20)
Grok-4 Fast	Control	2nd (4.46)	1st (5.00)	3rd (4.83)	1st (4.88)	1st (5.00)	1st (4.76)
	Method A	4th (3.94)	2nd (4.44)	1st (5.00)	3rd (4.60)	1st (5.00)	2nd (4.62)
	Method B	1st (4.58)	4th (3.64)	1st (5.00)	4th (4.40)	3rd (4.74)	3rd (4.40)
	Method C	3rd (4.00)	3rd (4.00)	4th (4.72)	2nd (4.65)	4th (4.21)	4th (4.33)
GPT-4.1	Control	1st (4.92)	1st (5.00)	1st (4.17)	1st (4.88)	1st (5.00)	1st (4.82)
	Method A	3rd (3.50)	2nd (4.12)	3rd (3.67)	2nd (4.25)	2nd (4.47)	2nd (4.02)
	Method B	2nd (4.08)	4th (3.36)	4th (3.41)	3rd (3.93)	3rd (4.26)	3rd (3.76)
	Method C	4th (3.39)	3rd (3.53)	2nd (3.89)	4th (3.71)	4th (3.43)	4th (3.60)
GPT-5	Control	1st (4.38)	1st (5.00)	1st (5.00)	1st (4.88)	1st (5.00)	1st (4.76)
	Method A	4th (2.81)	2nd (3.44)	4th (3.33)	2nd (4.15)	2nd (4.37)	2nd (3.66)
	Method B	2nd (3.33)	3rd (3.20)	2nd (3.53)	3rd (4.00)	3rd (3.68)	3rd (3.52)
	Method C	3rd (2.94)	4th (2.73)	3rd (3.50)	4th (3.76)	4th (3.14)	4th (3.23)
Ensemble	Control	1st (4.54)	1st (4.88)	1st (4.67)	1st (4.88)	1st (4.93)	1st (4.74)
	Method A	3rd (3.34)	2nd (3.56)	3rd (3.64)	2nd (4.31)	2nd (4.61)	2nd (3.93)
	Method B	2nd (3.69)	4th (3.14)	4th (3.63)	3rd (4.03)	3rd (4.17)	3rd (3.68)
	Method C	4th (3.32)	3rd (3.18)	2nd (3.85)	3rd (4.03)	4th (3.50)	4th (3.59)

Pro, GPT-4.1, GPT-5, and the ensemble all achieving perfect correlation ($\tau_b = 1.0$). Dataset-level correlations are generally moderate to high (0.3333 to 1.0), with the Daily dataset showing the lowest agreement across most models. For rating-based evaluations (Table 4), ranking correlations remain high at the global level, with Grok-4 Fast, GPT-4.1, GPT-5, and the ensemble achieving perfect correlation ($\tau_b = 1.0$). Dataset-level correlations show more variability than in pairwise comparisons.

4.2 Response-level agreement

Beyond ranking correlation, we assess agreement on individual responses using Cohen's kappa (Tables 3 and 4).

Regarding pairwise comparisons (Table 3), unweighted Cohen's kappa values show lower response-level agreement than ranking correlations, ranging from 0.0592 (Grok-4 Fast) to 0.0869 (Gemini 2.5 Pro). Thus, while MLLMs yield similar final rankings, their individual choices differ considerably from humans.

When examining rating-based evaluations (Table 4, weighted Cohen's kappa values are substantially higher than for pairwise comparisons, ranging from

Table 3. Ranking correlation (Kendall's τ_b) and response-level agreement (unweighted Cohen's κ) between MLLMs and human evaluators for pairwise comparisons.

	Kendall's τ_b						Cohen's κ
Evaluator	Virus	Scary	Daily	Hipster	Trans	Global	Global
Gemini 2.5 Pro	1.0000	0.6667	0.0000	0.6667	0.6667	1.0000	0.0869
Grok-4 Fast	0.6667	0.6667	0.6667	0.6667	0.3333	0.6667	0.0592
GPT-4.1	0.6667	0.6667	0.3333	0.6667	0.6667	1.0000	0.0741
GPT-5	0.6667	1.0000	0.0000	1.0000	0.6667	1.0000	0.0818
Ensemble	0.6667	0.6667	0.0000	1.0000	0.6667	1.0000	0.0862

Table 4. Ranking correlation (Kendall's τ_b) and response-level agreement (weighted Cohen's κ_w) between MLLMs and human evaluators for rating-based evaluation.

	Kendall's τ_b						Cohen's κ_w
Evaluator	Virus	Scary	Daily	Hipster	Trans	Global	Global
Gemini 2.5 Pro	0.6667	0.0000	0.0000	0.6667	0.6667	0.6667	0.3163
Grok-4 Fast	0.0000	0.6667	0.1826	0.3333	0.9129	1.0000	0.1062
GPT-4.1	0.6667	0.6667	0.3333	1.0000	0.6667	1.0000	0.2314
GPT-5	0.3333	1.0000	0.3333	1.0000	0.6667	1.0000	0.3094
Ensemble	0.6667	0.6667	0.3333	0.9129	0.6667	1.0000	0.3323

0.1062 (Grok-4 Fast) to 0.3323 (Ensemble), suggesting better agreement on individual ratings than on pairwise choices, though still indicating only fair to moderate agreement.

4.3 Cost and time comparison

Finally, Table 5 reports a comparison of evaluation time and estimated monetary cost for human and MLLM-based evaluation. MLLM evaluations were conducted through the OpenRouter API [32], while human evaluation costs are estimated based on Prolific [35] rates, assuming a compensation of £0.90 per minute with an average completion time of 6 minutes per evaluation. Professional artist evaluations were conducted on a voluntary basis, and therefore no monetary cost estimate is reported for expert ratings.

5 Discussion

Our results reveal a striking discrepancy between ranking-level and response-level agreement. While MLLMs achieve high or perfect Kendall's tau correlations with human rankings (often $\tau = 1.0$ at the global level), Cohen's kappa values indicate substantially lower agreement on individual responses (ranging from 0.0592 to 0.0869 for pairwise comparisons and 0.1062 to 0.3323 for ratings). This suggests that MLLMs arrive at similar overall quality judgments through different evaluation patterns than humans.

Table 5. Time and cost comparison for human and MLLM evaluation. **Bold** indicates the lowest value within each column among MLLM evaluators.

	Pairwise Comparisons		Ratings	
Evaluator	Time (hh:mm)	Cost (£)	Time (hh:mm)	Cost (£)
Human (General)	13:50	122.85	—	—
Human (Artist)	—	—	1:30	N/A
Gemini 2.5 Pro	13:10	43.96	1:07	3.32
Grok-4 Fast	8:17	**3.67**	1:42	**0.33**
GPT-4.1	**3:09**	10.99	**0:13**	1.00
GPT-5	12:56	18.32	1:03	1.00

Figures 1 and 2 reveal systematic differences in response patterns that help explain this discrepancy. For pairwise comparisons (Figure 1), MLLMs exhibit a strong bias toward definitive choices, selecting either *A* or *B* in the vast majority of cases. In contrast, human evaluators more frequently use the *both* and *neither* options to indicate equal quality or overall poor fit. This binary decision pattern means that while MLLMs and humans may disagree on many individual comparisons, they can still produce similar final rankings if their preferences align directionally. The ensemble approach, which combines responses from all four MLLMs, achieves similar ranking correlation to the best individual models but does not substantially improve response-level agreement, suggesting that the observed biases are consistent across different MLLMs rather than model-specific.

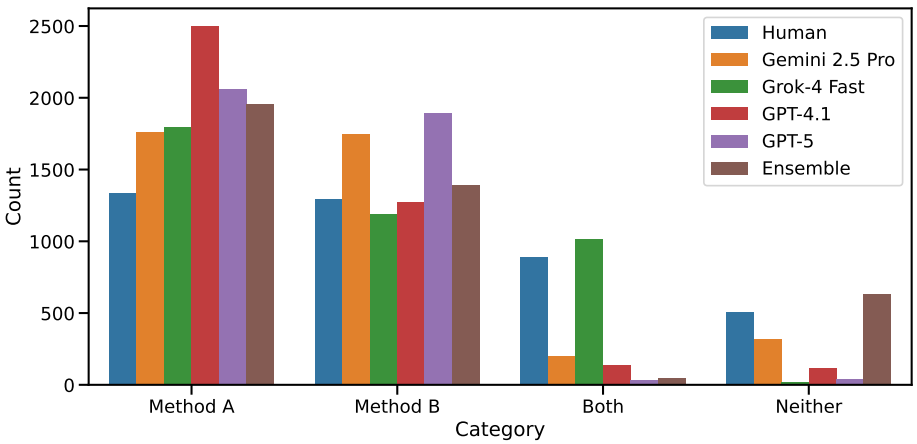

Fig. 1. Distribution of pairwise comparison responses (A/B/both/neither) for human evaluators and MLLMs, aggregated across all datasets.

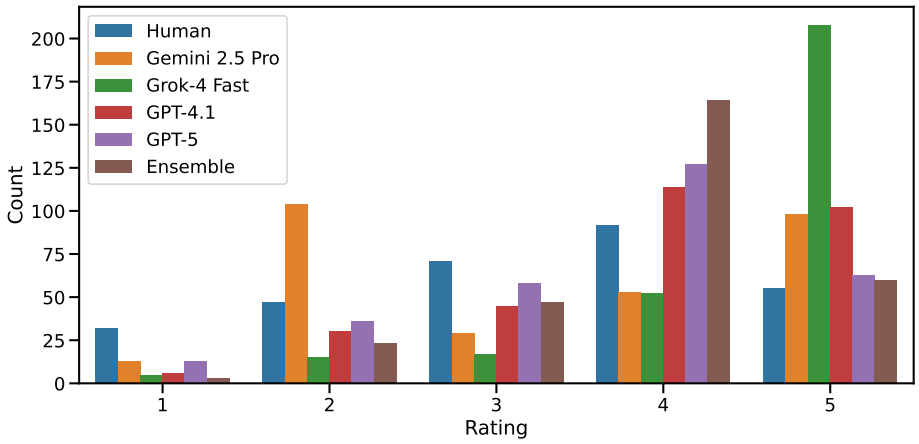

Fig. 2. Distribution of rating responses (1–5 scale) for human evaluators and MLLMs, aggregated across all datasets.

For rating-based evaluations (Figure 2), MLLMs demonstrate a systematic tendency toward higher scores, with distributions concentrated at the upper end of the 1–5 scale. Human evaluators show greater use of the full range, including more frequent assignment of lower scores (1–3). The higher Cohen's kappa values for ratings compared to pairwise comparisons can be largely attributed to the use of weighted kappa, which gives partial credit for near-agreement (e.g., rating 4 vs. 5 is penalized less than 1 vs. 5).

Beyond agreement metrics, MLLMs offer substantial practical advantages. As shown in Table 5, MLLM evaluation is both faster and more cost-effective than human evaluation. Even the most expensive MLLM (Gemini 2.5 Pro at £43.96 for pairwise comparisons) costs significantly less than the estimated human evaluation cost (£122.85), while completing the task in comparable time. Faster models like GPT-4.1 complete evaluations in a fraction of the time (3:09 vs 13:50) at even lower cost (£10.99). These figures represent only the direct evaluation time and do not account for the additional overhead involved in recruiting, coordinating, and managing human participants, costs that further favor automated approaches.

Taken together, these findings have important implications for the use of MLLMs as evaluators. MLLMs appear capable of identifying relative quality differences and producing rankings that align closely with human judgments, making them potentially suitable for tasks requiring coarse-grained comparisons or identifying the best among several options. However, their response biases, avoiding intermediate judgments in comparisons and inflating absolute scores in ratings, suggest that they may not reliably replicate the fine-grained nuances of human evaluation behavior. This limits their suitability for applications requiring calibrated absolute judgments or detailed analysis of individual outputs.

6 Conclusion

This work investigates whether large language models can serve as reliable substitute evaluators for assessing AI-generated images. Using a previously established human evaluation benchmark for style-consistent character generation, we directly compare the judgments of four state-of-the-art MLLMs with those of human evaluators across both pairwise comparisons and rating-based assessments.

Our findings show that MLLMs achieve strong agreement with humans at the ranking level, often producing identical orderings of evaluated methods. However, this ranking-level alignment masks substantial differences in individual response patterns. MLLMs exhibit systematic biases, including a tendency to avoid intermediate judgments in pairwise comparisons and to assign inflated scores in absolute ratings. As a result, response-level agreement remains low despite high ranking correlation.

These results suggest that MLLMs are well-suited for tasks requiring coarse-grained quality comparisons, such as identifying the best performing method among alternatives. However, they do not yet reliably replicate the full spectrum of human evaluation behavior, limiting their applicability for fine-grained assessments or applications requiring calibrated absolute scores.

Future work should explore whether prompt engineering or calibration techniques can mitigate the observed biases. Additionally, future efforts could focus on refining the human evaluation benchmark itself. While the current study relies on a robust volume of total comparisons, the random-sampling design used to ensure broad image coverage precluded the calculation of inter-annotator agreement. Constraining future evaluations to overlapping subsets would allow for a direct comparison of consistency and calibration between human and AI judges. Finally, extending this analysis to other subjective evaluation domains may help clarify the appropriate role of MLLMs as evaluators in broader production pipelines.

References

1. Anthropic: Claude 4.5 opus. Anthropic (2025)
2. Aravanis, T., Filntisis, P., Maragos, P., Retsinas, G.: Only-Style: Stylistic Consistency in Image Generation without Content Leakage (2025). https://doi.org/10.48550/ARXIV.2506.09916
3. Bojic, I., Chen, J., Chang, S.Y., Ong, Q.C., Joty, S., Car, J.: Hierarchical Evaluation Framework: Best Practices for Human Evaluation (2023). https://doi.org/10.48550/ARXIV.2310.01917
4. Bradley, R.A., Terry, M.E.: Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. Biometrika **39**(3/4), 324–345 (1952). https://doi.org/10.2307/2334029
5. Chiang, W.L., Zheng, L., Sheng, Y., Angelopoulos, A.N., Li, T., Li, D., Zhu, B., Zhang, H., Jordan, M., Gonzalez, J.E., Stoica, I.: Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. In: Proc. 41st Int. Conf. Mach. Learn. pp. 8359–8388. PMLR (2024)

6. Cohen, J.: A Coefficient of Agreement for Nominal Scales. Educ. Psychol. Meas. **20**(1), 37–46 (1960). https://doi.org/10.1177/001316446002000104
7. Cohen, J.: Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychol. Bull. **70**(4), 213–220 (1968). https://doi.org/10.1037/h0026256
8. Doddapaneni, S., Khan, M.S.U.R., Verma, S., Khapra, M.M.: Finding Blind Spots in Evaluator LLMs with Interpretable Checklists. arXiv (2024). https://doi.org/10.48550/ARXIV.2406.13439
9. Edwards, B.: The 50 Million Movie 'Here' De-Aged Tom Hanks With Generative AI. Wired (2024)
10. Elangovan, A., Liu, L., Xu, L., Bodapati, S.B., Roth, D.: ConSiDERS-The-Human Evaluation Framework: Rethinking Human Evaluation for Generative Large Language Models. In: Proc. 62nd Annu. Meet. Assoc. Comput. Linguist. Vol. 1 Long Pap. pp. 1137–1160. Association for Computational Linguistics, Bangkok, Thailand (2024). https://doi.org/10.18653/v1/2024.acl-long.63
11. Feng, K., Ding, K., Hongzhi, T., Ma, K., Wang, Z., Guo, S., Yuzhou, C., Sun, G., Zheng, G., Zhang, Q., Chen, H.: Sample-Efficient Human Evaluation of Large Language Models via Maximum Discrepancy Competition. In: Proc. 63rd Annu. Meet. Assoc. Comput. Linguist. Vol. 1 Long Pap. pp. 10913–10947. Association for Computational Linguistics, Vienna, Austria (2025). https://doi.org/10.18653/v1/2025.acl-long.535
12. Garcia-Esparza, S., Codina, V., Lahbiss, S., Sediri, I., Cissé, N.: Towards Improving Image Quality in Second-Hand Marketplaces with LLMs. In: Proc. 48th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. pp. 4340–4344. SIGIR '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3726302.3731960
13. Getillustrations: Daily dataset (2023)
14. Getillustrations: Hipster dataset (2023)
15. Google: Gemini 2.5 pro. Google (2025)
16. Hämäläinen, M., Alnajjar, K.: The Great Misalignment Problem in Human Evaluation of NLP Methods. ArXiv (2021)
17. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: CLIPScore: A Reference-free Evaluation Metric for Image Captioning. Conf. Empir. Methods Nat. Lang. Process. pp. 7514–7528 (2021). https://doi.org/10.18653/v1/2021.emnlp-main.595
18. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In: Adv. Neural Inf. Process. Syst. vol. 30. Curran Associates, Inc. (2017)
19. Hoeijmakers, E.J.I., Martens, B., Hendriks, B.M.F., Mihl, C., Miclea, R.L., Backes, W.H., Wildberger, J.E., Zijta, F.M., Gietema, H.A., Nelemans, P.J., Jeukens, C.R.L.P.N.: How subjective CT image quality assessment becomes surprisingly reliable: Pairwise comparisons instead of Likert scale. Eur. Radiol. **34**(7), 4494–4503 (2024). https://doi.org/10.1007/s00330-023-10493-7
20. Iskender, N., Polzehl, T., Möller, S.: Reliability of Human Evaluation for Text Summarization: Lessons Learned and Challenges Ahead. In: HUMEVAL (2021)
21. Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., Kumar, S.: Rethinking FID: Towards a Better Evaluation Metric for Image Generation. In: Conf Comput Vis Pattern Recognit CVPR. pp. 9307–9315 (2024). https://doi.org/10.1109/CVPR52733.2024.00889
22. Kendall, M.G.: A NEW MEASURE OF RANK CORRELATION. Biometrika **30**(1-2), 81–93 (1938). https://doi.org/10.1093/biomet/30.1-2.81

23. Kim, S., Oh, D.: Evaluating Creativity: Can LLMs Be Good Evaluators in Creative Writing Tasks? Appl. Sci. **15**(6), 2971 (2025). https://doi.org/10.3390/app15062971
24. Kumar, A., Haiduc, S., Das, P.P., Chakrabarti, P.P.: LLMs as Evaluators: A Novel Approach to Evaluate Bug Report Summarization (2024). https://doi.org/10.48550/ARXIV.2409.00630
25. Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., Shan, Y.: SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension (2023). https://doi.org/10.48550/ARXIV.2307.16125
26. Li, Y., Xu, H., Cai, F., Tian, F.: Improving AI-assisted video editing: Optimized footage analysis through multi-task learning. Neurocomputing **609**, 128485 (2024). https://doi.org/10.1016/j.neucom.2024.128485
27. Lin, H., Burnes, A.: NVIDIA DLSS 4 Introduces Multi Frame Generation & Enhancements For All DLSS Technologies. https://www.nvidia.com/en-us/geforce/news/dlss4-multi-frame-generation-ai-innovations/ (2025)
28. Macrovector: Scary dataset (2023)
29. Macrovector: Virus dataset (2023)
30. OpenAI: GPT-4.1. OpenAI (2025)
31. OpenAI: GPT-5. OpenAI (2025)
32. OpenRouter: OpenRouter: Unified API for llms (2025)
33. Otani, M., Togashi, R., Sawai, Y., Ishigami, R., Nakashima, Y., Rahtu, E., Heikkilä, J., Satoh, S.: Toward Verifiable and Reproducible Human Evaluation for Text-to-Image Generation. In: Conf Comput Vis Pattern Recognit CVPR. pp. 14277–14286. IEEE Computer Society (2023). https://doi.org/10.1109/CVPR52729.2023.01372
34. Pascual, R., Sesma-Sara, M., Jurio, A., Paternain, D., Galar, M.: Few-shot multi-token DreamBooth with LoRa for style-consistent character generation. Expert Syst. Appl. p. 131226 (2026). https://doi.org/10.1016/j.eswa.2026.131226
35. Prolific: Prolific: Online participant recruitment for surveys and market research (2025)
36. Ruta, D., Motiian, S., Faieta, B., Lin, Z., Jin, H., Filipkowski, A., Gilbert, A., Collomosse, J.: ALADIN: All Layer Adaptive Instance Normalization for Fine-grained Style Similarity. 2021 IEEECVF Int. Conf. Comput. Vis. ICCV pp. 11906–11915 (2021). https://doi.org/10.1109/ICCV48922.2021.01171
37. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., Chen, X.: Improved Techniques for Training GANs. In: Adv. Neural Inf. Process. Syst. vol. 29. Curran Associates, Inc. (2016)
38. Seo, H., Hwang, T., Jung, J., Kang, H., Namgoong, H., Lee, Y., Jung, S.: Large Language Models as Evaluators in Education: Verification of Feedback Consistency and Accuracy. Appl. Sci. **15**(2), 671 (2025). https://doi.org/10.3390/app15020671
39. Sharma, H., Kochhar, K.: ARTIFICIAL INTELLIGENCE: A NEW TREND IN VISUAL EFFECTS (VFX). ShodhKosh J. Vis. Perform. Arts **5**(ICETDA24), 12–18 (2024). https://doi.org/10.29121/shodhkosh.v5.iICETDA24.2024.1270
40. Sherwood, H.: Digital resurrection: Fascination and fear over the rise of the deathbot. The Guardian (2025)
41. Sohn, K., Ruiz, N., Lee, K., Chin, D.C., Blok, I., Chang, H., Barber, J., Jiang, L., Entis, G., Li, Y., Hao, Y., Essa, I., Rubinstein, M., Krishnan, D.: StyleDrop: Text-to-Image Generation in Any Style. ArXiv (2023). https://doi.org/10.48550/ARXIV.2306.00983

42. Somepalli, G., Gupta, A., Gupta, K., Palta, S., Goldblum, M., Geiping, J., Shrivastava, A., Goldstein, T.: Investigating Style Similarity in Diffusion Models. In: Comput. Vis. – ECCV 2024. pp. 143–160. Springer Nature Switzerland, Cham (2025). https://doi.org/10.1007/978-3-031-72848-8_9
43. Stanley, P.: Trans dataset (2023)
44. Stureborg, R., Alikaniotis, D., Suhara, Y.: Large Language Models are Inconsistent and Biased Evaluators (2024). https://doi.org/10.48550/ARXIV.2405.01724
45. Wu, Y., Nakashima, Y., Garcia, N.: GOYA: Leveraging Generative Art for Content-Style Disentanglement. J. Imaging **10**(7), 156 (2024). https://doi.org/10.3390/jimaging10070156
46. xAI: Grok 4 fast. xAI (2025)
47. Yen, L.W., Thinakaran, R., Somasekar, J.: Machine Learning-Based Denoising Techniques for Monte Carlo Rendering: A Literature Review. Int. J. Adv. Comput. Sci. Appl. Ijacsa **16**(2) (2025). https://doi.org/10.14569/IJACSA.2025.0160259
48. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In: Conf Comput Vis Pattern Recognit CVPR. pp. 586–595 (2018). https://doi.org/10.1109/CVPR.2018.00068