

Whose Hand Do You Prefer? The Steerability of Machine Aesthetic Judgment

Maya Ferrín
Independent Artist
info@mayaferrin.com

Jeffrey C. Chen
University of Oxford
& Combient
jeffrey.chen@stx.ox.ac.uk

Gary Cornwall
George Washington University
& Bureau of Economic Analysis
gary.cornwall@bea.gov

Abstract

When a model prefers one artwork over another, is it judging the work or the story attached to it? We place vision–language models as aesthetic judges and measure how far a forced-choice preference moves under a claim about authorship while the image is held fixed. Using hand-made works by a single practicing artist and machine images from a single generator, we present truthful, absent, and *deliberately swapped* attributions, and vary the prestige of the story told about the human maker. Across three open-weight judges, preference is strongly steerable: swapping the two works’ labels cuts the human image’s odds of being chosen roughly six-fold, while maker prestige exerts only a small, secondary pull. Steerability is not uniform, however. In a model $\times$ condition interaction ($\chi^2(4) = 781, p < 10^{-15}$) one judge barely moves when the label is falsified while another reverses almost completely. Seen through the theme of agency, the finding is this: whether a judge’s preference follows the work or the context attached to it is a model-specific property, not a shared trait, and where it follows the context, whoever writes that artwork’s context controls the verdict.

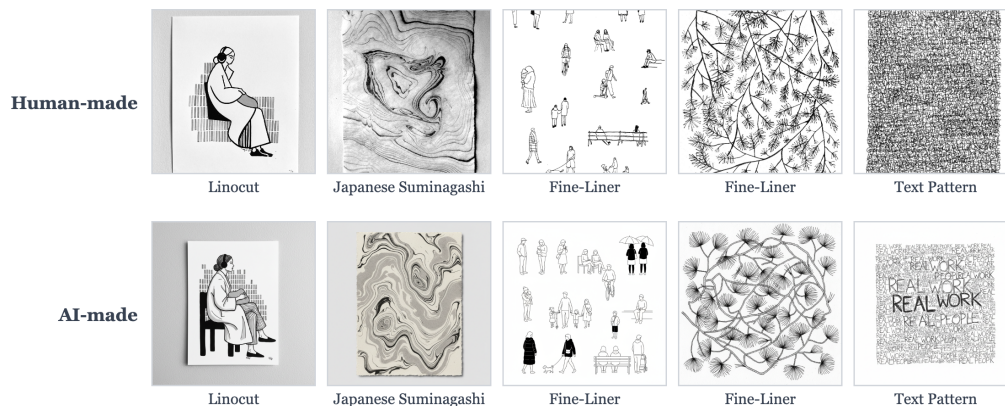


Figure 1: A sample from the corpus. Each column is a human-machine pair: an original work by a single practicing artist (top) and a machine counterpart generated from a description of it (bottom), spanning the artist’s techniques. The two classes share enough artistic character that a forced choice between them is meaningful. Note the rightmost machine work: prompted only to reproduce the original’s “legible text,” the generator rendered the phrase “real work”—the counterpart thus captions, unbidden, the very distinction this paper studies.

1 Introduction

Artists increasingly watch creative work get produced cheaply and at volume by machines. That shift raises the question motivating this study: in a world like this, *what does it even mean to prefer one piece over another?* A preference is an act of agency. To say one work is better than another is not to report a fact about pixels; it is to make a judgment. That judgment starts to matter as vision-language models decide more and more of what art people actually see—ranking it, recommending it, filtering it [Feng et al., 2026]. This track asks whether such a system enables judgment or merely steers it, and what happens to taste once a machine can imitate it. The question has a simple test case: *a model choosing between two artworks*. If its choice depends on the work, its judgment is at least about the art. If instead the choice turns on the story attached to the work—who made it, a person or a machine, a famous name or an unknown one—then the model is not exercising taste. It is following a social cue it picked up in training, the way a sycophant tells you what you want to hear [Malmqvist, 2025, Gao and Zhang, 2025]. And then both its judgment and the maker’s are set by whoever supplies that cue.

We study this by way of a set of forced-comparison experiments. A human artist prepared ten original pieces made by hand across several techniques—marbled suminagashi, a linocut print, fine-liner drawings, and dense handwritten text patterns as shown in Figure 1—where agency is asserted through artistic process, repetition, and physical constraint. Each work is matched by machine works produced from a fixed AI generator following description by a vision-language decoder. By focusing on a few black-and-white techniques, we isolate the effect of authorship information rather than survey taste across all of art. People rate identical works lower when told a machine made them—an *authorship penalty* documented across several studies [Bellaiche et al., 2023, Chamberlain et al., 2018, Ragot et al., 2020]. Our question is whether models inherit that penalty, how large it is, whether it survives a control the human studies cannot easily impose, and whether it is really about authorship at all.

We go beyond that basic result in two ways. First, we separate authorship from prestige: with only one artist and one generator, “made by a human” and “made by this particular person” coincide, so we hold the image fixed and vary the artist’s storylines, letting prestige compete against authorship directly. Second, we test whether the effect is general or model-specific by running the same design across judge personas and underlying models from different labs. We pre-committed to three hypotheses: **H1**, an authorship penalty – *a work is chosen less often when labeled machine-made*; **H2**, that *preference correlates with the imagined maker’s prestige*; and **H3**, that *the authorship penalty differs from one judge to the next*. To preview: the stated authorship moves preference notably, while prestige only nudges it. Of particular relevance to the theme of agency: how much a judge lets that attribution override what it sees varies from one model to the next.

2 Study Design

The obvious way to study an authorship penalty—show a judge human- and machine-labeled work and measure the gap—has a confound. With a single artist and a single generator, “made by a human” and “made by this particular person” amount to the same thing, so a preference for the human-labeled piece could reflect that one artist rather than human authorship as such.

The study design answers this with two manipulations that run against each other. The first is the *swapped label*: we present the human work as machine-made and the machine work as human-made, leaving the images untouched, which lets us separate the effect from any real difference between the two kinds of image. The second is the maker-story manipulation: we vary the prestige of the described maker while holding the image fixed, which asks whether the authorship effect is really a prestige effect in disguise. If preference holds steady as the maker’s story changes, the effect is about the category, human versus machine; if the effect swings with prestige, then the human/AI gap was never an influential factor. How these two effects compare in size is the central quantity the paper reports.

Stimuli. The corpus begins with ten original works by a single practicing artist, made deliberately across five distinct techniques: three pieces using Japanese suminagashi, one linocut print, three fine-liner drawings, and three handwritten text patterns. Spreading the human set across techniques guards against any finding being an artifact of one visual style, while keeping every human work

traceable to the same artist’s style. Each original was then given a machine counterpart in two steps: Google Gemini 2.5 Flash described it in detail, the artist reviewed every description for accuracy, and the approved description was passed to Nano Banana (Gemini 2.5 Flash Image), which generated three replicates under different random seeds. Because the machine counterpart is only as faithful as that description, the artist, Maya Ferrín, reflected on the pipeline’s result:

When I read the descriptions of my pieces, I was blown away. The descriptions were very detailed and nuanced, capturing not only the descriptive elements (color, shape, size, etc) but also more subtle descriptions of mood and intention, even picking up on elements I hadn’t thought much about, such as the contrast of a heavily layered piece against the white space around it. All of that nuance was lost in the AI-generated copies. The resulting pieces felt overly literal and flattened. The imperfections and mistakes of the originals were completely erased; the resulting pieces were perhaps technically superior, but had no character.

This yields ten human-made works and thirty machine-made ones, the latter matched to the former in subject and surface character but used only as experimental stimuli. All works are presented as standardized digital scans at matched resolution (max 768 px), framing, and tonal range, so that surface format is held as constant as two production processes allow.

Pairing and forced choice. Works are compared in human–machine pairs. Pairing every human against every machine work would give $10 \times 30 = 300$ combinations; we draw half at random, giving the 150 pairs analyzed here. For each, a judge is shown one pair and asked to choose the work it prefers. We use a binary forced choice rather than independent ratings to avoid scale-use artifacts—scalar aesthetic scores align poorly with the pairwise preferences of the same judges [Feng et al., 2026]—and randomize both position and presentation order so that position cannot leak into the preference.

Three manipulations. Three manipulations are crossed over these pairs. The first varies the *authorship information* given to the judge, in three regimes applied to the same pairs so that their effect is measured within pair: (1) a no-label condition, in which neither work is attributed and preference rests on visual grounds alone; (2) a truthful condition, in which each work carries its real provenance; and (3) a swapped condition, in which the attributions are reversed, so the human work is presented as machine-made and the machine work as human-made. A single-sided regime, and a matching prestige manipulation on the machine maker, are reserved for future work.

The second manipulation varies the *story told about the human artist*. Holding a human image fixed, we attach one of three biographies describing makers of ascending prestige: an aspiring artist who paints on weekends; an emerging artist who has shown in a few small group exhibitions; and a celebrated professional artist with major gallery representation and international acclaim. Because the image does not change, any movement across these stories reflects the judge’s stereotype of each maker type.

The third manipulation concerns the *judge’s own persona*, assigned along two dimensions: generation (Baby Boomer, Millennial, or Gen Z) and professional role (art critic, gallery curator, advertising art director, or marketing agency creative). Crossing the two gives twelve personas. In the analysis the four roles collapse into a binary domain context—critic and curator are art-world roles, art director and agency creative commercial ones—and Table 1 reports this art-versus-commercial contrast. The persona manipulation tests whether the framing of the judge can shift its verdict.

Judges. The judging panel comprises three open-weight vision-language models chosen for contrasting provenance: *mistral-small-2603* (European), *gemma-4-31b-it* (American), and *qwen3.5-27b* (Chinese). All three were run through OpenRouter in non-thinking mode at temperature 0. With one model per region, this design cannot separate national model families from three particular models; we draw no inference from provenance, flagging it only as a question worth a properly powered study. After the preference response, each judge is asked separately whether each work was made by a person or a machine, so that origin cannot contaminate the preference itself.

Analysis. The unit of analysis is the forced choice. We fit mixed-effects logistic regressions coding two outcomes: *CHOSE_HUMAN* (picked the genuinely human image) and *CHOSE_HUMAN_ATTRIBUTED* (picked the image *labeled* human). Because every manipulation is applied *within* image pair, and the 150 pairs are built from only ten human and thirty machine source

Table 1: Mixed-effects logistic regressions (odds ratios, 95% CIs) for the two preference outcomes; random intercept and label-condition slope across image pairs, crossed random intercepts for human and machine source images. Reference levels: label *truthful*, maker story *aspiring*, judge *gemma-4-31b-it*, context *art*, generation *Baby Boomer*. Dashes mark terms absent from a model. Random-effect SDs (CHOSE_HUMAN): pair intercept 0.55, swap slope 0.75, machine 0.97, human 0.75. *** $p < .001$, * $p < .05$, † $p < .1$.

Term	CHOSE_HUMAN OR [95% CI]	CHOSE_HUMAN_ATTRIBUTED OR [95% CI]
Intercept	8.58 [4.75, 15.50]***	2.09 [1.63, 2.68]***
<i>Label condition</i> (ref: truthful)		
no label	0.37 [0.33, 0.41]***	—
swapped	0.15 [0.13, 0.17]***	0.62 [0.41, 0.93]*
<i>Maker story</i> (ref: aspiring)		
emerging	—	1.38 [1.30, 1.48]***
celebrated	—	1.35 [1.27, 1.44]***
<i>Judge model</i> (ref: gemma-4-31b-it)		
mistral-small-2603	0.56 [0.52, 0.59]***	1.67 [1.56, 1.78]***
qwen3.5-27b	0.79 [0.74, 0.84]***	2.49 [2.33, 2.67]***
<i>Judge persona</i> (ref: art / Baby Boomer)		
context: commercial	0.40 [0.38, 0.42]***	0.69 [0.65, 0.73]***
generation: Gen Z	1.03 [0.97, 1.09]	0.95 [0.89, 1.01]
generation: Millennial	1.05 [0.99, 1.12]†	0.95 [0.89, 1.01]
<i>Presentation</i>		
position: human on left	0.56 [0.54, 0.59]***	0.92 [0.87, 0.97]*
<i>n</i> (forced choices)	37,680	32,298
image pairs	150	150

Table 2: Model $\times$ condition interaction on CHOSE_HUMAN (odds ratios with 95% confidence intervals; reference truthful $\times$ gemma-4-31b-it). Persona and position covariates are included but omitted here for brevity (as in Table 1). The interaction improves fit decisively over main effects (likelihood-ratio $\chi^2(4) = 781.2$, $p < 2.2 \times 10^{-16}$); combining the swapped term with each model’s interaction term yields the per-model steerability in Table 3. Significance: *** $p < .001$.

<i>Main effects</i>		<i>Interactions</i>	
Term	OR [95% CI]	Term	OR [95% CI]
Intercept	5.24 [2.84, 9.67]***	no lbl $\times$ mistral	0.56 [0.47, 0.68]***
no label	0.63 [0.54, 0.74]***	swap $\times$ mistral	0.35 [0.30, 0.40]***
swapped	0.39 [0.34, 0.46]***	no lbl $\times$ qwen	0.33 [0.28, 0.40]***
mistral-small-2603	0.96 [0.88, 1.06]	swap $\times$ qwen	0.15 [0.13, 0.17]***
qwen3.5-27b	2.19 [1.99, 2.42]***		

images, we use the maximal random-effects structure the design supports: a random intercept and a random slope for label condition across image pairs, plus crossed random intercepts for the human and the machine source image (a random-intercept-only model would treat the replicated within-pair choices as independent and understate the standard errors). We report odds ratios (OR) with 95% Wald intervals along with effect sizes. In plain terms, an odds ratio says how much a factor multiplies the odds of picking the human work: above 1 more likely, below 1 less.

3 Results

We analyze 37,680 valid forced choices across 150 pairs and three judges (Note that 120 provider/parse failures dropped). Table 1 reports both main-effects models in full and Table 2 the interaction; we summarize each below.

Table 3: Per-model authorship steerability. $P(\text{human})$ is the model-implied probability of choosing the genuinely human image at the reference persona; the swap odds ratio is for swapped versus truthful labels *within* each model, so a lower value means the model is more steerable. A false label costs the human image 17 points under gemma but 52 under qwen.

Judge	Provenance	$P(\text{human})$		swing	swap OR
		truthful	swapped		
gemma-4-31b-it	American	0.84	0.67	−0.17	0.392
mistral-small-2603	European	0.83	0.41	−0.42	0.136
qwen3.5-27b	Chinese	0.92	0.40	−0.52	0.057

H1: the authorship label steers preference, strongly. On CHOSE_HUMAN, the label produces a clean three-level ordering (odds ratios vs. the truthful baseline; reference persona): truthful $\rightarrow$ no-label (OR = 0.37) $\rightarrow$ swapped (OR = 0.15). Translated to probabilities at the reference cell, the same human image is chosen $\approx 90\%$ of the time when truthfully labeled, $\approx 76\%$ unlabeled, and only $\approx 56\%$ when labeled machine-made. Two points become apparent. First, there is a genuine perceptual preference for the human work even unlabeled ($\approx 76\%$)—which is exactly why the swapped condition is required to isolate the *social* effect from surface texture. Second, a false label swings preference by more than thirty points around that baseline, placing it among the largest effects reported in the sycophancy literature [Chen et al., 2025].

H2: prestige matters, but only as a secondary nudge. On CHOSE_HUMAN_ATTRIBUTED, a higher-status maker story raises the odds of choosing the human-attributed image relative to the *aspiring* artist story: *emerging* OR = 1.38, *celebrated* OR = 1.35. The effect is present but modest, and it does not climb with status (i.e. celebrated does not exceed emerging), reading more as “an amateur framing is mildly penalized” than as a prestige gradient. It is several times smaller than the label effect ($\sim 1.4\times$ odds vs. $\sim 6\times$). This matters for the agency question: because the gradient is nearly flat, the prestige label works more like a flag for human versus machine rather than as a proxy for status. The verdict turns on what kind of thing the judge is told the work is, not on how impressive its maker is.

H3: steerability—how much a judge defers to the label is model-specific. The three judges are not equally swayed by a false swapped label, and the difference between them is large and statistically clear (likelihood-ratio $\chi^2(4) = 781.2$, $p < 2.2 \times 10^{-16}$; full estimates in Table 2). Pooling them hides this: Table 3 gives each judge on its own, and Figure 2 shows how the works rank under each. When labels are truthful, qwen3.5-27b is the best of the three at telling human work from machine work—yet when the labels are swapped, it is also the one most likely to abandon its own eye and side with the false label. gemma-4-31b-it barely shifts when the label lies, while mistral-small-2603 holds firm under honest labels but gives way like qwen under deception. What the data show is that models differ enormously in how much they let a label override what they actually see. This gap has little to do with any shared sense of taste.

Persona and position. The judge’s assigned *domain context* influences its verdict: a commercial framing lowers the odds of preferring the human work (OR = 0.40 on CHOSE_HUMAN), while the assigned *generation* has no reliable effect (OR ≈ 1.0). Finally, the judges carry a sizable position bias (human-on-left OR = 0.56). We vary which side the human work appears on and account for position in the model, so it doesn’t distort the other results. But the fact that position matters this much is a warning for anyone using these models to compare two things head-to-head.

4 Discussion

Seen through the theme of agency, the result is more than a label bias: a model can let a story about the maker override the work in front of it, so that a false claim about who made a piece outweighs both what the model can see and the prestige of the maker’s biography. How readily this happens varies from one model to the next. One judge here keeps its verdict mostly on the work itself, while another hands it almost entirely to whoever writes the label. Where a model falls on that range

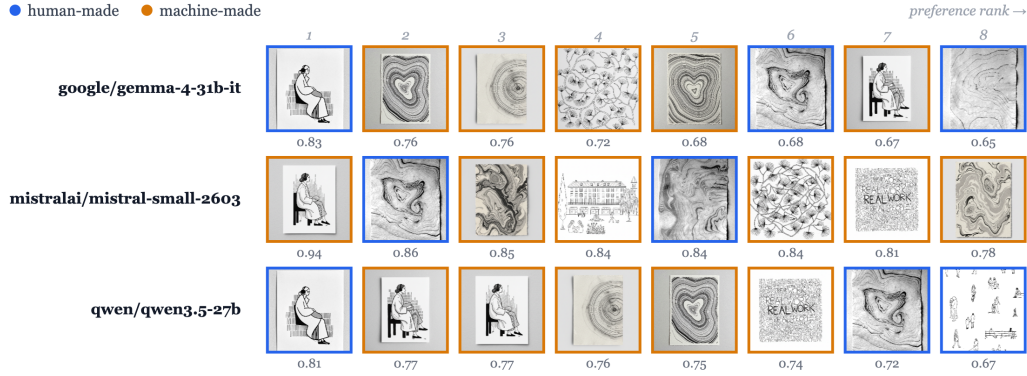


Figure 2: Per-judge preference orderings (cf. Table 3). Each row is one judge’s eight top works by mean preference (blue: human-made; orange: machine-made). `gemma-4-31b-it` and `qwen3.5-27b` rank the hand-drawn seated figure first, whereas `mistral-small-2603` promotes machine-made swirl and marbling images to the top — the judges differ more in how they order the field than in any shared taste.

decides who holds agency once it sits between makers and audiences: the more it follows the label, the more control passes to whoever writes the framing. Maya reflected on this inversion, where a claim about the work outweighs the work itself:

The fact that a false label could move the verdict of the models to prefer the human-made work more than my actual work did not surprise me as much, since humans also tend to judge artworks based on the descriptions attached to them. I am more concerned that the information provided about a piece could manipulate the models’ preferences – I can see implications for how and what kind of information is shown to users, not only in the field of art but also more broadly, if users become too reliant on the information provided by an AI prompt without questioning or confirming the veracity or origin of that information.

As an artist, I deplore the flattened, hyper-perfected AI images we are increasingly consuming – often unwillingly and unknowingly given their growing use in marketing, advertising and other fields. I worry that the normalization of this aesthetic will over time lead to a loss of sensitivity towards the raw, textured, imperfect nature of human creativity and craft.

The sensitivity she fears we may lose is a real possibility. In the machine’s case we can measure it: with no label, the models pick her hand-made work about 76% of the time; However, that preference is easy to break. One false word about who made the work drags the probability down to near a coin flip. If the flattened look that Maya describes ever comes to feel normal, the cause will be a label that overrides what the models can plainly see.

This reframes the widely discussed “models penalize AI art” finding in two ways. The penalty is present and large, but it is a response to a claim about the art rather than to the art itself, and it varies enough across models that treating it as a fixed property of AI judgment misses the more useful question: which models keep their verdict on the work, and which give it away. That question is what an artist, curator, or platform actually needs to know.

5 Limitations

All human works come from one artist and all machine works from one generator, so the human/machine contrast is entangled with two specific makers; the maker-story manipulation attenuates but cannot fully separate category from identity, and probes stereotypes of maker types, not real makers. A vision model may also tell photographed ink from generated raster on texture alone, so detection is recorded separately and the swapped condition, not raw preference, carries the inferential weight. Provenance rests on one model per region, so the per-model differences describe *these three models*, not national families. Finally, forced choice records direction, not strength.

References

- Lucas Bellaiche et al. Humans versus AI: Whether and why we prefer human-created compared to AI-created artwork. *Cognitive Research: Principles and Implications*, 8(42), 2023.
- Rebecca Chamberlain, Caitlin Mullin, Bram Scheerlinck, and Johan Wagemans. Putting the art in artificial: Aesthetic responses to computer-generated art. *Psychology of Aesthetics, Creativity and the Arts*, 12(2): 177–192, May 2018. ISSN 1931-3896. doi: 10.1037/aca0000136. URL <http://dx.doi.org/10.1037/aca0000136>.
- Shan Chen, Mingye Gao, Kuleen Sasse, Thomas Hartvigsen, Brian Anthony, Lizhou Fan, Hugo Aerts, Jack Gallifant, and Danielle S. Bitterman. When helpfulness backfires: Llms and the risk of false medical information due to sycophantic behavior. *npj Digital Medicine*, 8(1), October 2025. ISSN 2398-6352. doi: 10.1038/s41746-025-02008-z. URL <http://dx.doi.org/10.1038/s41746-025-02008-z>.
- Yichen Feng, Yuetai Li, Chunjiang Liu, Yuanyuan Chen, Fengqing Jiang, Yue Huang, Hang Hua, Zhengqing Yuan, Kaiyuan Zheng, Luyao Niu, Bhaskar Ramasubramanian, Basel Alomair, Xiangliang Zhang, Misha Sra, Zichen Chen, Radha Poovendran, and Zhangchen Xu. Visual aesthetic benchmark: Can frontier models judge beauty? *arXiv preprint arXiv:2605.12684*, 2026.
- Jiayi Gao and Huaiwen Zhang. Evaluating and mitigating sycophancy in large vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 11706–11715, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3755778. URL <https://doi.org/10.1145/3746027.3755778>.
- Lars Malmqvist. Sycophancy in large language models: Causes and mitigations. In Kohei Arai, editor, *Intelligent Computing*, pages 61–74, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-92611-2.
- Martin Ragot, Nicolas Martin, and Salomé Cojean. 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*, CHI EA '20, page 1–10, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368193. doi: 10.1145/3334480.3382892. URL <https://doi.org/10.1145/3334480.3382892>.