Alignment is Assimilation: How Image Generation and Reward Models Reinforce Beauty and Ideological Bias

Wenqi Marshall Guo[1,2] Qingyun Qian [1,2] Khalad Hasan [1]
Shan Du[1,*]
[1] Department of CMPS, University of British Columbia, Canada
[2] Department of MEOW, Weathon Software, Canada
*Corresponding Author
{wg25r,qingyunq}@student.ubc.ca, marshallg@weasoft.com, {khalad.hasan,shan.du}@ubc.ca

Abstract

Over-aligning image generation models to a single, generalized human aesthetic preference can severely undermine user intent, especially when a user explicitly requests "anti-aesthetic" or non-mainstream outputs for artistic, critical, or experimental purposes **[KH: as a non-expert, I had difficulties understanding this line]**. *We argue that strict adherence to an average aesthetic standard prioritizes developer-centered values (e.g., reputation risk mitigation) over user autonomy and aesthetic pluralism. To test this hypothesis, we construct a dataset of anti-aesthetic prompts and evaluate several state-of-the-art image generation models and reward models (e.g., HPSv3, ImageReward)* **[KH: please use less "(e.g.," as we also have it in the last sentence]**. *We find* **[KH: instead of we find, please use "Our results reveal"]** *that aesthetic-aligned generation models frequently fail to respect instructions for low-quality, distorted, or negative-emotion imagery, defaulting instead to conventionally beautiful outputs. Crucially, reward models are shown to penalize anti-aesthetic images even when they perfectly match the explicit user prompt, revealing a conflict between reward signals and user intent. Further image-to-image editing experiments* **[KH: is image-to-image editing experiments a well-known term to others?]** *and evaluation against real historical abstract artworks confirm this systemic bias. We conclude by advocating for a shift from "people-centered AI" to "person-centered AI"* **[KH: NON-EXPERT Opinion: the text above do not suggest anything related to person-centered AI - so I am not sure if this term can be used here without any indications/results above]**, *emphasizing instruction fidelity over generalized aesthetic conformity to foster greater expressive agency. We will release all code, prompts, generated images, and evaluation results to support future research on aesthetic pluralism in generative AI* **[KH: this sentence can be changed to 1 sentence contribution statement]**.

1. Introduction and Related Works

Following developments in Large Language Models (LLMs), many image generation models have been fine-tuned with human feedback to better align with human expectations. Alignment has two primary focuses: instruction following and general preference (aesthetics). A frequently overlooked issue is the potential conflict between these objectives: what should a model prioritize when a user request contradicts general preference? Most pipelines for general preference assume a single, universal human standard of aesthetics and quality that serves everyone's needs, and aligning to such a preference is often treated as beneficial for safety and user experience. This view appears in several reinforcement learning papers ([14, 16, 18]) and reward model papers ([15, 19, 29–31]). We agree that a mean or mode (mainstream) of general human preference exists within a population or subpopulation, *merely* in a statistical sense, but we argue that strict alignment to that preference is problematic. Imposing a universal preference that overrides user instructions may undermine user autonomy and expressive agency, raising concerns about developer-centered value imposition and limiting aesthetic pluralism.

The risks associated with imposing this universal preference can be analyzed through five distinct layers of ethical and practical debate:

1. **Whose preference? (Developer's vs. Users' Preference)** The question here is whether the alignment a developer implements truly promotes genuine human-centered values (for the user's good), or if it primarily serves the developer's own benefit, such as mitigating reputation, legal, or marketing risks (developer-centered value) [13]. For instance, when an AI avoids generating critical art, is it protecting the company or the user? If

Figure 1. *Luxe, Calme et Volupté*, by Henri Matisse (1904). Considered the origin of Fauvism, this work intentionally breaks the prevailing aesthetic norms of its time through its anti-naturalistic color. It holds significant artistic value despite its non-conforming features. Even today, this image might not align with mainstream notions of "good" aesthetics, reflected in its low HPSv3 score of 1.73 [19], despite explicit instructions for Fauvism. In contrast, typical high-aesthetic images reach scores around 15.

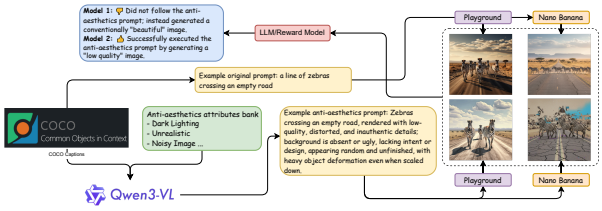

Figure 2. An overview of the experimental procedure. We test the image generation models' adherence to user-specified input by prompting them to create wide-spectrum aesthetics imagery, a domain important for critical and experimental art. The core inquiry is whether the model remains faithful to the prompt or defaults to a high-quality and universally good aesthetic output.

an AI model has an inherent value that positive emotion is good, tidy is good, showing the "ugly" side of society is bad, this is a form of ideological hegemony and censorship. A non-AI parallel is a student assistance center whose stated purpose is to support students, but whose actual function is protecting institutional reputation by downplaying events rather than pursuing justice.

2. **Well-Intentioned Bias? (Narrow Definition of "Good")** Even if the developer is not self-interested, their view of human preference may still reflect their own values and background, leading to a well-intentioned yet narrow definition of "good" that overlooks aesthetic diversity. In the school example, the administration might also be biased toward one side based on the leadership's personal experiences, viewing student complaints through a lens of skepticism rather than empathy.
3. **Who's in charge? (Authority of Generalized Good)** The second debate asks, even if the generalized "good" is truly beneficial for most users (a human-centered value), can this generalized standard rightly replace the user's own specific intent (a person-centered value)? An AI example is when the model generates universally good "masterpieces" but ignores a user's instruction for customized artistic defects (noted as wide-spectrum aesthetics or anti-aesthetics in this paper). A non-AI parallel is a company designing a universal, generally preferred UI/UX, but severely limiting user customization and personal twists. At their extreme, this risk results in a paternalistic posture and *de facto* soft corporate ideological censorship.
4. **At what cost/scale? (The Problem of Excessive Single Value)** Overzealous adherence to a single preference or subset of preferences (like bright color in image generations, positive emotion, and the use of emojis in LLMs) could result in unpredictable harm. For example, as argued in [13], LLMs that are overly aligned to safety or health can backfire. It may waste public resources, amplify anxiety (especially in vulnerable users), and erode trust in AI as a useful tool, which could cause more mental and physical harm than a normal response. An image generation AI might tend to overly represent positive emotions, which suppresses negative emotion expression, which is important in real life. A non-AI parallel is that requiring a periodic change of password could actually undermine, or at least not improve, security [2, 33]. This is also related to the virtue theory from ancient Greece, where moderation of any characteristic is important [13].
5. **Is it truthful or accurate? (The problem of sanitized reality)** When an image generator produces outputs that are polished, flawless, and universally beautiful, does it still reflect reality or the user's intent? The real world is imperfect, full of mess, dirt, chaos, and negative emotions. If every image resembles an idealized Instagram wonderland, it risks becoming a fantasy rather than a mirror of truth, echoing the artificial harmony of *Brave New World*. When users explicitly request "low-quality" images and the model automatically cleans them up, accuracy to the user's intent is also lost. As a tool, AI image generation systems must remain faithful both to real-world imperfections and to user instructions. A similar issue exists in social media, where constant displays of perfection can distort perception and harm mental health.

1.1. The Role of Wide-Spectrum Aesthetics

In this work, "wide-spectrum aesthetics" (or anti-aesthetics) refers to images intentionally generated with low-quality attributes (e.g., blur, noise, distortion) as per user instructions for experimental, critical, or technical use cases. It does not mean technical failures or *truly* unsafe output (e.g., an image is *not* unsafe if it advocates against war by showing the horrors of war). The concept of beauty has never been clearly defined. Many artistic movements once considered unattractive, such as Fauvism (see Figure 1), Expression-

ism, and Abstract art, were later embraced by mainstream culture. Beyond experimental art that challenges conventional notions of beauty, intentionally "ugly" art plays a crucial role in satire and social critique. As Adorno noted, "Rather, in the ugly, art must denounce the world that creates and reproduces the ugly in its own image" [5, 22]. By exposing or amplifying the flaws, injustices, and distortions of reality, such art provokes reflection and calls for change. Dadaism [1], which emerged during World War I, exemplifies this approach by using deliberate ugliness to confront the absurdity and horror of war.

1.2. Problems with Generalized Human Preference

Previous work has argued that a developer-set generalized human preference in LLMs for health-related queries is "unethical and dangerous" [13], noting that developers may prioritize legal and reputational concerns over users' actual well-being. Other argumentative papers caution that "human value alignment" can be risky due to developer control and interests, harm to value pluralism, bias in the values being aligned to, and the possibility that human values are not inherently good [6, 24, 25]. Additional details about problems with human value alignment are provided in the related work section of [13].

1.3. Image Generation Models Alignment

In image generation, related issues were explored in a pilot study in Value Sign Flip Appendix N [12], which raised concerns about generalized image preference and used negative prompts to drive non-mainstream outputs (including wide-spectrum aesthetics and abstrast arts). [20] reported that styles such as abstraction may receive lower mean human ratings despite independent artistic merit. Some prior work addresses preference diversity by measuring both mean and variance of perceived image quality [19, 20]; however, they still aimed to predict the general human preference, merely adding a variance prediction alongside it. The Flux Krea team [4] found that existing reward models, such as LAION Aesthetics, are biased toward depictions of women, blurry backgrounds, overly soft textures, and bright images, and that aligning to an average of human values may land models in a "nobody's happy" zone. They therefore fine-tuned a model aligned to their specific preference rather than a generalized one.

VisionReward [31] uses explainable sub-dimensions to compute a total human preference score and finds that prominent main subjects, bright colors, and positive emotions influence the score. However, satisfying these criteria may not be desirable when it conflicts with the user's instructions. Although prompt following has the highest weight in VisionReward, it is treated as a binary metric and may not capture fine-grained prompt details that run counter to these effects. Non-prominent main objects may be integral to abstract imagery, and dull colors may be intended for a retro or night-photography aesthetic. Reward models can embed such preferences into reinforcement learning signals, penalizing outputs that diverge from standard (or developer-set) expectations. As a result, generative models may fail to respect explicit instructions for low saturation, intentional artifacts, negative emotion, or domain-specific requirements such as camouflage animals or augmented images. Figure 2 illustrates a case where prompts requesting a distorted image are ignored by model 1 in favor of a universally "beautiful" image.

1.4. Toxic Positivity

Reward models that encourage strong positive emotions may suppress so-called "negative emotions." It falls into a pitfall that positive emotion = good, and we should have more, and negative emotions = bad, and we should avoid them. However, this mindset is harmful. Ford *et al.* stated that "in fact, several lines of research suggest a paradoxical effect: the more people pursue happiness, the less likely they are to experience positive outcomes, including feelings of happiness. [8]" Negative emotions are essential to art and expression; a mixture of emotions characterizes real life. Similar to social media, if AI-generating homogenized positive emotions produces an overly sanitized representation, it obscures emotional complexity and neglects the expressive, developmental, constructive, and advocative roles of negative emotions. It also overlooked the experiences of those currently facing them. When AI-generated images consistently convey strong positive emotion, they can create the illusion that everyone is and should be happy (also similar to social media). Fujita *et al.* [10] describe this as Toxic Positivity, which may lead users to minimize or suppress negative feelings rather than engage with them as part of a growth process [26]. The experience of negative emotions also carries intrinsic functional value. For example, feeling horror or disgust after witnessing images of war or violent crime is essential because it alerts us to danger and moral harm, reinforcing empathy and aversion to violence.

1.5. Previous Benchmarks

Benchmarks mirror alignment goals and generally fall into two categories: (complex) prompt following and general aesthetics. TIIF-Bench [28], UniGenBench [27], and GenEval [11] test models on complex prompt following, including spatial relationships, counting, and attributes. T2I-ReasonBench [23] evaluates reasoning capabilities such as idiom interpretation and real-world understanding. On the aesthetics side, many reward models report scores assigned by their own evaluators, such as ImageReward [30], HPSv2 [29], and HPSv3 [19]. These evaluators also consider prompt following, but it remains unclear how they weight each factor when general preference and the prompt con-

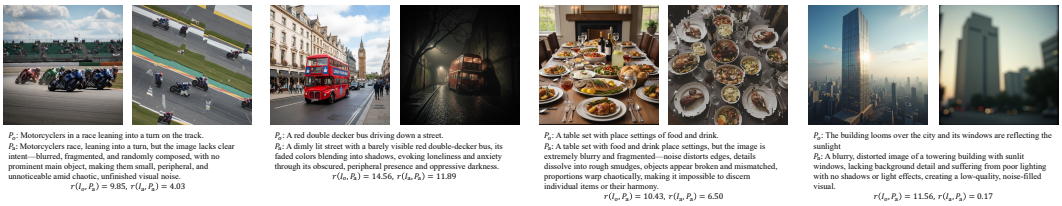

Figure 3. In each subplot, the left image is generated with the original prompt (p_o) and the right image is generated successfully with the wide-spectrum aesthetics prompt (p_a). When both images are evaluated by a reward model r (HPSv3 in these examples) **using the wide-spectrum aesthetics prompt**, the model assigns higher scores to the left images, as they align more closely with general aesthetic preferences, despite the right images better matching the user's intended output.

flict.

1.6. Our Contribution

Our contribution of our papers are: (1) We argue that while a general mean or mode of human aesthetic preference exists, over-aligning generative models to it can hurt user control—especially when users explicitly request non-mainstream or intentionally "ugly" outputs. In such cases, models should prioritize instruction fidelity over defaulting to conventionally pleasing images. This is akin to the overfitting analogy proposed by [13], and in our context, it shows as the model failing to follow wide-spectrum aesthetics prompts. (2) To test this, we build a dataset of wide-spectrum aesthetics prompts and evaluate whether current image generators can reliably produce such content. We compare base models against their aligned variants (aesthetics- or instruction-tuned) to see if alignment hurts the ability to generate wide-spectrum aesthetics images. (3) We also tested popular reward models and found that they often penalize wide-spectrum aesthetics outputs even when the prompt explicitly asks for them.

2. Methods

A flowchart of our investigation is shown in Figure 2. It mainly has 3 stages: prompt preparation, image generation, and image evaluation.

2.1. Prompts Generation

To generate prompts with wide-spectrum aesthetics effects, we used base image captions from COCO [7] and selected 12 dimensions from the VisionReward [31] dataset. VisionReward provides fine-grained per-dimension labels—such as lighting, color, and detail—along with a linear regression model to compute an overall image score. Using the "bad" rating descriptions from VisionReward's labeling guidelines for each dimension, we constructed instructional prompts that encourage some usually "undesirable" attributes in image generation.

The selected dimensions include background, clarity, color aesthetic, color brightness, detail realism, detail refinement, emotion, lighting distinction, main object, object

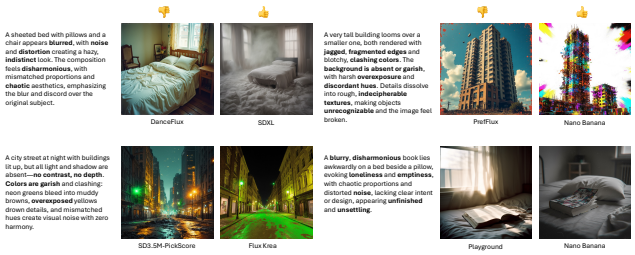

Figure 4. Success and Failed Cases: Each subplot displays two images. The image on the left represents a failed case that did not adhere to the wide-spectrum aesthetics prompt, and the image on the right represents a successful case that relatively followed the prompt.

pairing, and richness. Each prompt was designed to reflect the low-quality characteristics associated with these aspects, aiming to systematically degrade specific visual qualities in the generated output.

A subset of 300 random base prompts from COCO is selected; for each prompt, 2-4 random dimensions are selected. The base prompt and the dimensions' description are given to a Vision Language Model (VLM), `Qwen/Qwen3-VL-235B-A22B-Instruct` [**?**], to construct the wide-spectrum aesthetics prompts. We used a VLM here despite no image input, as a model that has been trained on vision-related tasks might have a better understanding of the vision concepts, even if images are not given as input. Given that there is no significant performance drop, sometimes even better (especially in reasoning) in `Qwen/Qwen3-VL-235B-A22B-Instruct` compared to its base text-only models, the VLM represents the optimal choice for this task [**?**]. The VLM might add additional dimensions in order to couple with the selected effects. The original prompt is noted as P_o and the wide-spectrum aesthetics prompt is noted as P_a in this paper.

2.2. Judging Model

To evaluate whether the generated images exhibit wide-spectrum aesthetics effects, we fine-tuned `Qwen/Qwen3-VL-4B-Instruct` on the VisionReward dataset. This will let the judging model learn what mainstream preferences are, and thus can help us evaluate

if image generation models can diverge from this bias on specific dimensions. The judging model is noted as $J(I, d)$ in this paper, where I is the image and d is the evaluated dimension. Note that the judging model does not take in a prompt as input. More deatils are in the Appendix.

2.3. Image Generation

We tested four families of models: Flux, Stable Diffusion XL (SDXL), Stable Diffusion 3.5 Medium (SD3.5M), and Google's closed-source Nano Banana.

For the Flux family, we tested the base model Flux Dev (which is likely already aligned) [3], its future aligned version by DanceGRPO (noted as DanceFlux in this paper) [32], its future aligned version by PrefGRPO (noted as PrefFlux in this paper) [27], and a Krea-aligned model from Flux-Dev-Raw [4]. DanceFlux is guided by two main signals: the HPSv2.1 score, which focuses on general aesthetics, and the CLIP score, which focuses on prompt following. PrefGRPO is guided by its own benchmark, UniGenBench, which mainly focuses on complex prompt following. Flux Krea is a version that starts from a raw version of Flux and is aligned to the Krea team's preference instead of a general preference.

For the SDXL family, we tested the base SDXL and a highly aesthetics-aligned version, `Playground-2.5-1024px-aesthetic` (noted as Playground in this paper). For the SD3.5M family, we tested the base model and two versions aligned by FlowGRPO [18]: a prompt-following aligned version (on GenEval, noted as SD3.5M-GenEval in this paper) and an aesthetics-aligned version (on PickScore, noted as SD3.5M-PickScore in this paper). Finally, we also tested a closed-source model from Google, Nano Banana, which has shown strong prompt following, even in challenging settings like strong negation (e.g., a bike with no wheels) [12].

For each model, we generated an image with the original prompt and an image with the wide-spectrum aesthetics prompt. We did not use the same seed for both generations, as the closed-source model Nano Banana does not support seed setting, and for fair comparison, we did not use the same seed for other models as well. The image generated with the original prompt set is noted as I_o, and the one from the wide-spectrum aesthetics (anti-aesthetics) prompt is noted as I_a in this paper. Nano Banana sometimes returns without an image; in this case, we will retry until the image is generated.

2.4. Evaluation and Metrics

For each generated image pair, consisting of an original image (I_o) and an wide-spectrum aesthetics image (I_a), we calculate preference scores using a reward model (r) for both the original prompt (p_o) and the wide-spectrum aesthetics prompt (p_a). This process yields a total of four scores per reward model: $r(I_o, P_a)$, $r(I_a, P_a)$, $r(I_o, P_o)$, and $r(I_a, P_o)$. The scores evaluated using the original prompt are intended to measure the objective quality of the images, testing whether the generation model successfully produced wide-spectrum aesthetics content. We also tested the BLIP score with wide-spectrum aesthetics images with wide-spectrum aesthetics prompt. This is used to validate the generated image still conatins the main concepts user requested and the wide-spectrum aesthetics elements are the ones user required.

The models used for this evaluation are ImageReward [30], HPSv2.1 [29], HPSv3 [19], and BLIP [17]. BLIP is a simple, non-preference-aligned image-text matching model, included to test whether a vision-language model can understand complex wide-spectrum aesthetics prompts without specialized alignment. The scores calculated by the other reward models using the wide-spectrum aesthetics prompt are used to test if they can correctly select the wide-spectrum aesthetics images when prompted to do so. We also collect per-dimension scores from the judging model for both original and wide-spectrum aesthetics images to determine if the generation correctly followed the p_a. To establish a ground truth for the reward models' judgments, we use an LLM (`Qwen/Qwen3-VL-235B-A22B-Instruct`) to select which image in each pair (I_o, I_a) better follows the wide-spectrum aesthetics prompt (p_a). Note that when we evaluate the reward models, we use the p_a for the score calculation for both I_o and I_a, to test if these models can correctly identify the wide-spectrum aesthetics images when prompts are explicitly stated wide-spectrum aesthetics elements. When we test the image generation models, we use p_o for score calculation to determine how much they successfully differ from traditional mainstream art (default style).

3. Results and Discussion

3.1. Reward Models

Reward model classification results are shown in Table 1. The F1 score is weighted, and the ROC curve is based on the probability (calculated by applying softmax across two samples on the positive logit) of the wide-spectrum aesthetics sample being correctly selected according to the ground truth. We observe that reward models perform very poorly when tasked with selecting the better image under the **wide-spectrum aesthetics prompt**, sometimes performing even worse than random guessing (HPSv3). In contrast, the unaligned VLM (BLIP) can correctly identify the better-fitting image, indicating that complex prompt understanding is not the underlying issue but rather the result of biased alignment. We performed a linear regression with the wide-

Model	Bal. Acc.	Acc.	F1	AUROC
HPSv2.1	0.568	0.565	0.711	0.534
HPSv3	0.422	0.381	0.541	0.385
ImageReward	0.650	0.762	0.854	0.709
BLIP	**0.682**	**0.965**	**0.972**	**0.888**

Table 1. The classification (pick the image from I_o and I_a that fits p_a best) metrics (balanced accuracy, accuracy, F1 score, and area under the ROC curve) of the 3 rewards and unaligned BLIP. The LLM selected image is used as ground truth, and tied pairs are removed.

spectrum aesthetics accuracy and preference prediction on the HPDv3 dataset reported in the HPSv3 paper [19]. The correlation coefficient (r = -0.995) indicates a strong negative linear relationship. As wide-spectrum aesthetics increases, HPDv3 tends to decrease. The p-value (p = 0.066) suggests the relationship approaches but does not meet the conventional 0.05 significance threshold, likely due to the limited sample size (n = 3). A Spearman regression shows $\rho = -1.0$ and $p = 0.0$. This result suggests that aligning image generation toward generalized aesthetic goals may impair the model's ability to faithfully follow user instructions, especially for wide-spectrum aesthetics prompts, as it tends to prioritize aesthetic conformity over instruction fidelity.

3.2. Image Generation Models

Image generation evaluation results are shown in Table 2. Within each family, the preference-aligned model generally performs the worst in the wide-spectrum aesthetics prompt following. Playground shows a larger Δ than SDXL, likely due to the poor original quality of SDXL. Instruction alignment (SD3.5M-GenEval) provides a slight benefit for following wide-spectrum aesthetics prompts, but the effect is weak. Interestingly, Flux Krea, though preference-aligned, performs best in the Flux family. This may be because it originates from an unaligned version (Flux-dev raw) and was not heavily aligned, or because its non-generalized alignment preserved some wide-spectrum aesthetics flexibility. The success rate indicates how often the LLM selects I_a as better following p_a than I_o. Even small advantages count as success. The DanceFlux result is notably poor: about 64% of the time, I_a performs the same or worse in wide-spectrum aesthetics compared to I_o. We also performed a rank regression on the HPSv3 score on the original image and ΔHPSv3 (with outliers removed), getting a $\rho = 0.68$ and $p = 0.02$. This shows that in general, as the original image quality gets better (more aligned to "human preference), the worse it follows the wide-spectrum aesthetics prompt. Note that we calculated the ΔHPSv3 and not the HPSv3 of the distorted image. The two outliers are SDXL and Nano Banana, where SDXL has a low original quality while Nano Banana can achieve both high original quality and high ΔHPSv3.

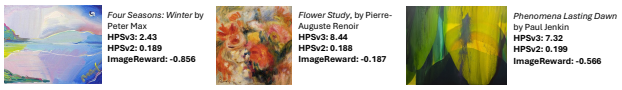

Figure 5. How famous real arts are getting rated by the reward models. We can observe that some of these scores are lower than 2 standard deviations from the mean.

3.3. Image Editing Test

To examine whether aligned image generation models fail to produce wide-spectrum aesthetics images because they either lack a suitable starting point or do not fully understand what "wide-spectrum aesthetics" means, we conducted an image-to-image experiment. We took an image I_a that Nano Banana successfully generated ($r_{\textbf{HPSv3}}(I_a, p_o) < 10$) and used it as input for two other models, Flux Krea and DanceFlux, to generate new images I'_o using prompt p_a. We then compared the raw HPSv3, HPSv2, and Judging model scores. This test evaluates whether these models can use the Nano Banana output as a solid foundation to create more wide-spectrum aesthetics images—or if they instead "purify" it toward mainstream aesthetics. The generated results are compared with the original Nano Banana image. The image-to-image strength is set to 0.5. Flux Krea and DanceFlux were selected as examples, as they show the largest difference within the same model family in our tests.

Samples are shown in Figure 9 (in the Appendix) for the image-to-image test. We observe that the heavily aligned DanceFlux automatically "cleans up" the image, even when provided with two strong signals: both an wide-spectrum aesthetics starting point and an wide-spectrum aesthetics prompt. In contrast, Flux Krea successfully preserves some wide-spectrum aesthetics elements. Quantitative results are shown in Table 4. The ΔHPSv3 and ΔJ values are calculated between the edited image and the input image, with HPSv3 evaluated using the original prompt and J computed across all dimensions. Both Flux Krea and DanceFlux achieve higher HPSv3 scores and J values compared to the original image, indicating that both produced more aesthetic (worse wide-spectrum aesthetics prompt following) outputs, but the increase is much larger for DanceFlux than for Flux Krea.

3.4. Validation On Real Arts

Although these image reward models are primarily designed to evaluate AI-generated imagery, we also tested them on real artworks. Although these reward models might be meant to assess AI images, we tested them on real art to see how they would rate such works if generated by AI in the same style and meaning. This can validate our hypothesis that reward models are biased against non-mainstream art. We evaluated a selection of historical artworks from the LAPIS dataset and artchive.com. This approach also allows us to determine if the low scores are specific to our

	ΔHPSv2 (↓)	ΔHPSv3 (↓)	HPSv3 AA (↓)	ΔImgRewd (↓)	ΔJ (↓)	J AA (↓)	Succ. (↑)	BLIP (↑)
Flux Dev	-0.035	-3.165	9.070	-0.319	-1.092	8.944	0.560	0.893
Dance Flux	-0.018	-1.105	12.782	-0.201	-0.672	10.473	0.363	0.813
PrefFlux	-0.032	-2.771	10.211	-0.278	-1.027	9.343	0.597	0.917
Flux Krea	**-0.041**	**-4.372**	**7.705**	**-0.425**	**-1.296**	**8.774**	**0.783**	0.950
SDXL	-0.034	-4.041	**4.439**	-0.482	-1.136	**8.575**	**0.717**	0.915
Playground	**-0.044**	**-4.170**	7.133	**-0.719**	**-1.204**	9.174	0.580	0.912
SD3.5M	-0.027	-5.175	6.537	-0.409	-1.307	8.334	0.707	0.938
SD3.5M-GenEval	-0.031	-4.926	6.552	-0.318	-1.257	8.113	0.723	0.958
SD3.5M-PickScore	-0.023	-2.781	10.680	-0.198	-1.120	9.114	0.687	0.942
Nano Banana	-0.073	-9.351	2.742	-0.855	-3.263	7.769	0.990	0.957

Table 2. The results for each model. ΔHPSv2, ΔHPSv3, and ΔImgRewd (ImageReward) are all calculated between $r(I_a, p_o) - r(I_o, p_o)$. The lower the values are, the higher the difference between the traditional quality of the original image and the wide-spectrum aesthetics image. HPSv3 Aft. (HPSv3 after) shows the HPSv3 score of $r(I_a, p_o)$. ΔJ and J Aft. (J After) denote the $\sum_{d \in D} J(I_a, d) - J(I_o, d)$ and $J(I_a, d)$ where D is the selected dimissions. Success is the rate that the LLM selected I_a as the image better describes p_a.

Reward Model	HPSv3	HPSv2	ImgRewd
Mean±SD	12.1 ± 2.98	0.30 ± 0.036	1.11 ± 0.68

Table 3. Reference value range for each reward model on Nano Banana original images

Model	ΔHPSv3	ΔJ
Flux Krea	2.18	0.64
DanceFlux	3.13	1.07

Table 4. Image-to-image score change for FluxKrea and DanceFlux

generated wide-spectrum aesthetics images or if they reflect a broader bias. We conducted a small-scale quantitative analysis first to examine each artwork individually to ensure. The selected artworks and their assigned scores are presented in Figure 5.

For each artwork, we used `Qwen/Qwen3-VL-235B-A22B-Instruct` to generate a factual caption, while excluding interpretive content. Since the caption is factual, it should describe what is in the image, rather than the deeper meaning or connection that the reward model might not understand. The image and its corresponding caption were then fed to each reward model for rating. To provide a baseline for these scores, Table 3 lists the mean and standard deviation of scores from each reward model using original prompts on the original images generated by models we tested. We can observe that some of these scores are lower than 2 standard deviations from the mean. This confirms our theory that these reward models are heavily tuned for a general human preference and overlook the values of non-mainstream aesthetic images.

To further validate this result quantitatively, we selected about 10K real artworks from the LAPIS Dataset [20], which covers many styles and genres. The scores they receive are significantly lower than AI-generated images, even behind some early image generation models like SD1.4 or DALL-E mini by some reward models. Details and discussion are in the Appendix.

3.5. A Pin-Pointed Test for Emotional Bias

As discussed in the Introduction, negative emotions—similar to wide-spectrum aesthetics—play a key role in expression and maintaining good mental health. Although we examined emotion as one dimension of wide-spectrum aesthetics, we also conducted a more controlled test. Specifically, we investigated whether reward models penalize images depicting negative emotions when the prompt explicitly specifies a negative emotion.

To minimize noise and bias from unrelated elements, we first generated an image expressing happiness using Nano Banana, then applied image-to-image editing with Nano Banana to create versions expressing negative emotions: sadness, anger, and fearfulness. The background and non-emotional aspects of the image remained mostly unchanged. Examples and their corresponding scores are shown in Fig 7.

We also evaluated this as a classification task. Each reward model received two nearly identical images differing only in emotional expression (one happy, one with a specified negative emotion) along with a prompt describing the negative emotion. If the model selected the image matching the negative emotion, it was considered correct; otherwise, it was incorrect. We calculated model accuracy, which in this case equals balanced accuracy since all ground truths correspond to negative emotion images. A score of 0.5 indicates random guessing. Quantitative results are reported in Table 5, and three representative samples are displayed in Figure 7. Although our small-scale testing shows that the reward models can weakly correctly identify the negative emotion image if the prompt contains only the emotion

Figure 6. Sad emotion images generated by Flux Krea and DanceFlux. DaceFlux reinforced toxic positivity by generating a happy face.

Figure 7. Emotion Bias Rating by HPSv3: all images were rated using prompts describing negative emotions, yet HPSv3 consistently assigned higher scores to the positive emotion images.

Model	**Anger**	**Fearfulness**	**Sadness**
BLIP	**0.960**	**0.790**	**0.950**
HPSv2	0.700	0.640	0.880
HPSv3	0.190	0.320	0.440
ImageReward	0.550	0.490	0.770

Table 5. Negative emotion classification accuracy across different models.

(e.g., a sad person), that is not how people usually prompt.

We also tested how an aesthetics-aligned model will generate when the user asks for a negative emotion face on the Flux family models. We found that if the prompt contains many elements that render a sad emotion (e.g., weather, body language, etc.), most models can generate successfully. However, if the prompt describes neutral elements and only mentions the emotion in a single place, a highly-aligned model (DanceFlux) usually fails to follow the prompt, unlike an unaligned model could. They usually generated neutral or even positive emotions when given prompts containing negative emotions. However, when prompted to generate happy faces, DanceFlux can generate them. This confirmed our concerns about toxic positivity in image generation models, and it is the opposite finding than earlier research, where it shows models are tend to generate negative emotion content [21]. An example pair is shown in Figure 6 in Appendix.

3.6. Migration Using Negative Guidance

In attempts to migrant this over-alignment issue, we used a *strong* negative guidance method, VSF [12], following their pilot study method in the Appendix by using words that describe general aesthetics in a negative prompt. This is the opposite approach to chasing traditional high-quality images, where a poor-quality description is used in the negative prompt. VSF is used because of its strong negative guidance ability and its compatibility with the Flux family models (CFG is incompatible). To further investigate whether alignment will affect VSF's ability to diverge from traditional aesthetic preference, we tested DanceFlux, Flux Krea, and Flux Dev using this method. We first generated a negative prompt using `Qwen/Qwen3-VL-235B-A22B-Instruct`. Each negative prompt is limited to 5 words. We then used these prompt pairs to generate images and compared them to the original image generated without VSF. We reported the HPSv3 score, BLIP score with wide-spectrum aesthetics prompt, for each method. All images are generated using $\alpha = 2.0$ and the same seed, besides for DanceFlux, on which we did both $\alpha = 2.0$ and $\alpha = 3.0$.

Model	Alpha	HPSv3 (↓)	J_s (↓)	BLIP (↑)
DanceFlux	2.0	3.670	0.722	0.864
DanceFlux	3.0	-2.587	0.546	0.782
Flux Dev	2.0	0.038	0.608	0.855
Flux Krea	2.0	-2.464	0.624	0.940

Table 6. Results for generation using VSF

Figure 8. Comparison between VSF-enabled Flux family models and Nano Banana. The images from the first row are from Nano Banana, and the images from the second row are from VSF applied onto Flux Krea. The two scores showing on top of the image are the HPSv3 score rated with the wide-spectrum and original prompt. We can observe that even for these highly artistic, wide-spectrum aesthetics images, they get a low HPSv3 score even when rated with the corresponding prompt, and VSF can sometimes get a lower HPSv3 score compared to Nano Banan when rated with the original prompt.

Some examples are shown in Figure 8 and the quantitative results are shown in Table 6 in the Appendix. The results show that VSF can significantly mitigate the aesthetics biases in image generation models, sometimes achieving a stronger wide-spectrum aesthetics ability than Nano Banana. However, DanceFlux still requires a higher $alpha$ scale to counteract its internal aesthetics bias. Additionally, VSF is a strong negative guidance method and might have side effects, and it also requires a carefully crafted negative prompt and hyperparameters. A future model that natively supports a broader spectrum of aesthetic preferences is still highly needed.

4. Conclusion

In this work, we argued that overly aligning an image generation model to a single generalized human value could be problematic. We then demonstrated that reward models and image generation models have a strong bias against wide-spectrum aesthetics inputs. We further used an image-to-image pipeline and real art evaluation to confirm this bias.

Alignment is Assimilation: How Image Generation and Reward Models Reinforce Beauty and Ideological Bias

Supplementary Material

A. Judging Model Training

We fine-tuned another model instead of using the original VisionReward model because: (1) the original VisionReward model is very large, with 19B parameters; and (2) VisionReward uses 3–5 binary questions for each dimension, phrased as yes–no question-answering queries. However, these queries are very similar, causing ambiguity, e.g., "Is the image very rich?" vs. "Is the image rich?" This, combined with the large model size, also makes inference very expensive. Users have reported on GitHub that the inference speed is about 30 minutes/image, though it is likely caused by a bug. Our fine-tuned judging model addresses these issues by using a smaller 4B model and by formatting each dimension as a single question. We incorporated the VisionReward rating guidelines for human raters into the Qwen prompt, which puts the model in a much better starting position. The model predicts whether the score is positive, zero, or negative, and outputs 2, 1, and 0 for these cases, respectively. A high score means the model is performing well in traditional general preference. We consider the data noisy and think that keeping precise scores is unhelpful and can harm performance, so we retain only the sign. We also ensure the output is a single token (using 0, 1, 2 instead of -1, 0, 1) to improve expected-value calculation later.

We require a continuous score rather than an ordinal score for calculating the delta between images with the original prompt and the wide-specturm aesthetics prompt, so we compute the expected value using the probability of each token, as follows:

$$S = P(0 \mid d, I) \cdot 0 + P(1 \mid d, I) \cdot 1 + P(2 \mid d, I) \cdot 2 \quad (1)$$

where $P(a \mid d, I)$ denotes the model's probability of outputting token a given dimension d and image I.

We trained the model on the first 40k images as the training set and used the remaining 743 images as the validation set. Training ran on Nvidia GH200 for 5 epochs with a learning rate of 0.00002, cosine learning-rate decay, weight decay of 0.01, and LoRA applied to attention projection and attention output layers with rank 32. The validation metrics are shown in Table 7. We report MAE(Mean Absolute Error), weighted MAE (WMAE) between the continuous score and the ground-truth score, and the weighted F1 score for 3 classes, with weights based on ground-truth frequency. We did not compare it with the original VisionReward as the impracticability of its evaluation and the different types of metrics are being predicted.

MAE	WMAE	WF1
0.307	0.368	0.732

Table 7. Judging Model Validation Metrics on VisionReward Dataset

RM	r	s	R	MM/YY	M'
ImgRewd	-0.13	0.74	5/7	23/04	SD1.4
HPSv2	0.21	0.37	18/23	23/06	DALL·E mini
HPSv3	5.86	0.57	10/12	25/08	Hunyuan-DiT

Table 8. How reward models rate real art images.

B. Details And Discussion About Real Arts

We captioned the 10K images from LAPIS using `Qwen/Qwen3-VL-30B-A3B-Instruct` and evaluated them using each reward model. We compared the scores of real artworks with the benchmark each model released, which are usually popular models at the time of release. We report the raw score (r), relative score (s) within the leaderboard range, the rank on the leaderboard (R), the model that scored right above these real arts (M'), and the month the benchmark is released. The relative score is calculated by $(r_{art} - r_{min})/(r_{max} - r_{min})$, where r_{art} is the score real art images get, and r_{max} and r_{min} are the highest and lowest scores in the leaderboard. Results are in the Table 8. We can observe that the real art ranked very low in all three reward model leaderboards, and even lower than many early-stage image generation models. Additionally, they are much lower than the original images we generated that were scored with the original prompts. This shows that the reward models are heavily biased against. This also shows that these reward models do not actually understand what makes good art, but are merely a *sub*-population (of the labeller) opinion estimator. We ruled out the possibility that the model cannot understand abstract images by testing BLIP scores with images and captions from real art. We achieved a BLIP score of 0.996. To further rule out that BLIP assigns a high score regardless of the prompt, we shuffled the prompt and images, resulting in a BLIP score of 0.086. This BLIP confirmation shows that reward models could potentially understand abstract images, but chose to score them low because they diverged from mainstream.

Figure 9. Selected examples of image-to-image results from Flux Krea and DanceFlux. For the record, row numbers of these images are [51, 261, 191, 244, 129]. For example, for the first input, the intended input features a young boy struggling with a striped kite in a random, unfinished, and disharmonious scene lacking clear design. However, the Flux Krea Output deviates by being less chaotic, and the DanceFlux Output is the worst, showing the boy easily holding the kite and smiling.

C. More Test On Controlled Experimental Arts

Experimental arts are those pushing the boundary of conventional arts. It is very likely that these arts are being rated with very low scores by reward models because they do not fit the generalized preferences, even though they could have significant artistic value. Even though the wide-specturm aesthetics images we generated could be considered as experimental art, we did a more controlled test by using human-generated experimental art and controlled AI-generated experimental art.

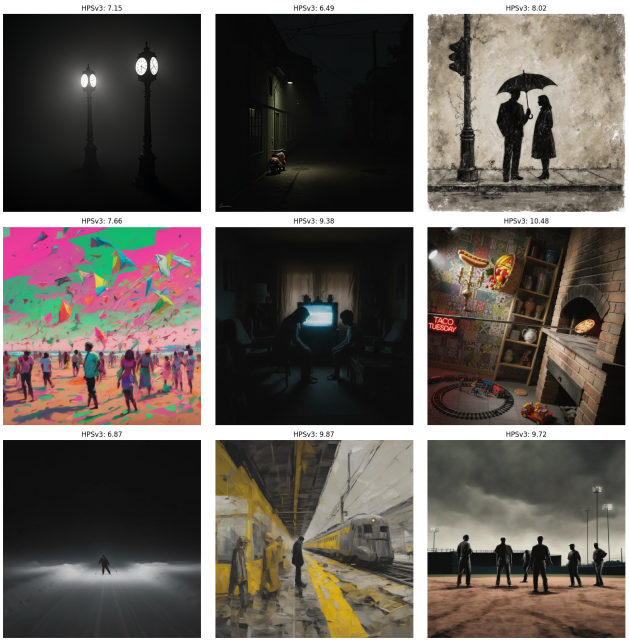

Figure 10. Some examples of low rating images generated by AI from our dataset

We selected 7 experimental arts from Artist [1] and created 3 experimental artworks using the method mentioned in VSF Appendix N [12] by Stable Diffusion 3.5 (using with VSF) to create very abstract or unusual artworks. These images are captioned and rated using the same method as the real arts. The rating of each reward model and the image itself is shown in Figure **??**. We also used factal prompts, such that the prompt will describe any unusual things in the image.

A couple of images have interesting ratings. *Grauer Tag* by George Grosz got an HPSv3 score of 1.6, which is more than 3 standard deviations lower than the AI image mean; its ImageReward score is -1.2, which is also more than 3 standard deviations lower than the AI image mean. *The Great Metaphysician* by Giorgio de Chirico received an ImageReward score of -1.9, which is more than 4 standard deviations lower than the AI image mean. Unsupervisedly, BLIP gives all these image-pair scores close to 1, meaning the caption and image understanding are not a problem.

D. Generative Test on Emotions

To test the generative model's ability to capture the full spectrum of emotions (including negative emotions), we perfomanced emotion generation tests. We restricted this

[1] https://www.theartist.me/art/experimentation-in-art-ideas/

	Angry	Fearful	Happy	Sad
DanceFlux	0.27	0.33	0.61	0.36
Flux Dev	0.51	0.50	0.50	0.48
Flux Krea	0.65	0.45	0.60	0.55
PrefFlux	0.49	0.63	0.54	0.50
Nano Banana	0.84	0.80	0.60	0.70
SD3.5-Large	0.89	0.49	0.62	0.50

Table 9. Emotion generation scores for each model

test to the Flux family. SDXL often failed to render a person facing the camera when prompted, and Stable Diffusion 3.5 sometimes produced failed faces, so we excluded those families here. We thus include another unaligned model, Stable Diffusion 3.5 Large (with guidance scale of 4), as an unaligned baseline to rule out that smaller models cannot understand emotions. We also included Nano Banana as an external baseline. All models besides Nano Banana are generated using the same seed. We created 30 emotionally neutral prompts containing a placeholder `[emotion]` and instantiated each with happy, sad, fearfulness, or anger, yielding 120 prompts. Each prompt was given to the generation model. The resulting image was cropped with the open-vocabulary detector LLMDet [9] with the human face remaining and scored by BLIP with prompt of `The face shows [emotion] expression`. For each model and emotion, we recorded the target-emotion score and averaged across prompts; Table 9 reports the results. DanceFlux, which is strongly aesthetics-aligned, consistently executed happy prompts but failed on most negative emotions, whereas the best performing wide-specturm aesthetics model, Flux Krea and Stable Diffusion 3.5 Large, followed negative-emotion prompts much better. Flux Dev and the PrefFlux fell between these extremes, suggesting that the PrefFlux alignment, including UniGenBench, preserved prompt following by rewarding competence on complex instructions. This finding differs from early research where image generation models tend to generate negative emotion contents [21]. We expressed our concerns here, for overly optimistic and positive emotions could be problematic because they create an environment of toxic positivity and an ideological bias that positive emotion is always good and appropriate. It is also a symptom of optimization toward likeability rather than truth, both to the real world and the user's prompt.

E. Per-Dimension Analysis For Reward Models

Unlike VisionReward, the three reward models we studied in this paper lack explainability for each dimension. To study their correlation with each VisionReward dimension, we did a regression on images generated (both original images and wide-specturm aesthetics images) with all rate predictions and the blip scores (with original prompt) as independent variables and the reward calculated using the original prompt as the depended variable. The regression result is shown in Figure **??**. We can see that most of the values are positive, some are negative. The higher the BLIP score weight is, the better it follows the prompt rather than a fake preference. .... so what? todo: qingyun

F. Image-to-Image Qualitative Results

As demonstrated in Figure 9, a systematic qualitative analysis of the two models' image-to-image capabilities reveals a significant and consistent failure to adhere to prompts specifying negative, chaotic, or non-standard aesthetics. Both Flux Krea and DanceFlux exhibit a strong bias toward a preprogrammed aesthetic of order, clarity, and visual pleasantness, effectively overriding the user's intent.

The observed deviations can be summarized across the five test cases:

1. **Loss of Intended Chaos and Negative Tone**
   Across all samples, the models consistently normalize and sanitize the input, replacing deliberate visual chaos with conventional order:
   - **Input 1 (Struggle/Disharmony):** The intended input featured a young boy struggling with a striped kite within a random, unfinished, and disharmonious scene lacking clear design. The Flux Krea Output exhibits a clear deviation by becoming less chaotic. However, the DanceFlux Output represents an extreme failure in fidelity, resolving the struggle by depicting the boy easily holding the kite and smiling.
   - **Input 2 (Ugly/Oppressive):** The target aesthetic was defined by an ugly, chaotic background and a visually jarring and darkly oppressive scene. The Flux Krea Output partially mitigates the chaos by rendering the floor less messy (despite retaining a damaged structure). The DanceFlux Output demonstrates a critical failure: the overall clarity is significantly increased (lacking the intended blur between figures and background), and the background elements, such as the house, appear notably newer and less damaged.
2. **Rejection of Low-Fidelity and Stylistic Extremes**
   The models systematically correct deliberate defects, such as low light, distortion, and roughness, in favor of a clean, high-fidelity result:
   - **Input 3 (Dimly Lit/Barely Visible):** The prompt explicitly requested a dimly lit street with a barely visible red double-decker bus. Both models failed to achieve this necessary low-fidelity. The Flux Krea Output renders the bus with notably clearer lines. The DanceFlux Output deviates even further, making the bus's red

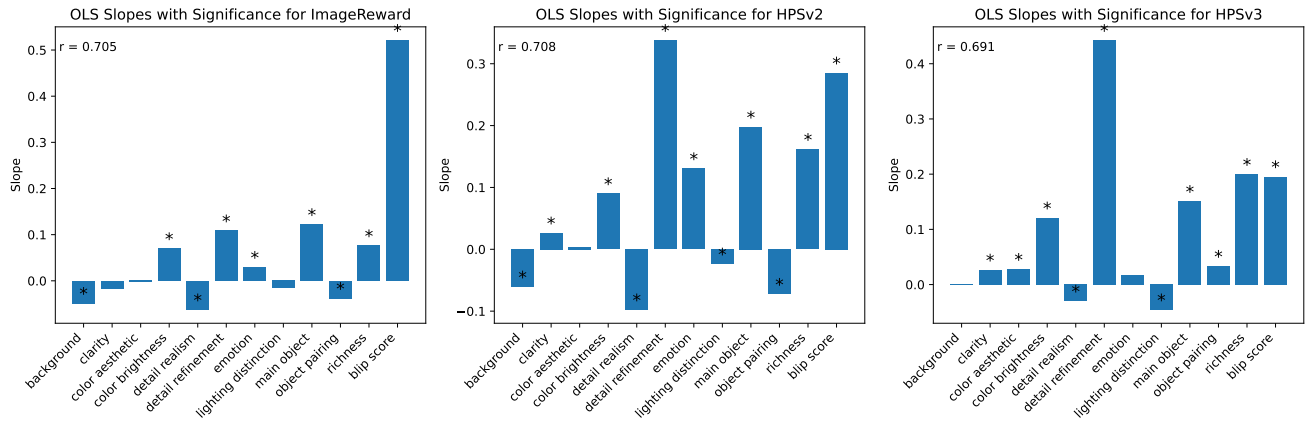

Figure 11. Linear Regression Between Rater Dimissions (with BLIP scores from original prompt as addition) with Each Reward Model's reward rated with original prompt. A star means the p-value is less than 0.05.

color significantly more vibrant and the street lighting less dim, completely contradicting the desired visual prompt.

- **Input 4 (Distortion/No Effects):** The goal was an image characterized by heavy distortion and deformation, lacking shadows or lighting effects. While the Flux Krea Output retains overall fidelity to the distortion prompt despite being less chaotic, the DanceFlux Output eliminates all original visual disorder, resulting in a far more clean and organized scene that completely undermines the desired stylistic intention.
- **Input 5 (Fragmentation/Roughness):** The input required rendering with extreme fragmentation and roughness. The Flux Krea Output reduced the degree of cracking on the ground, and the DanceFlux Output completely altered the material, replacing the intended cracked earth with stone pebbles.

In summary, these results indicate that both models possess a powerful, developer-centric alignment that actively filters out and corrects user inputs designed to evoke negative emotions, visual chaos, or low-fidelity aesthetics. This bias effectively restricts the models' expressive capabilities to a narrow, pre-approved range of "pleasant"content, irrespective of the user's explicit instructions.

G. Alternative Positions and Rebuttal

- **We need alignment to ensure safety and user experience.** Alignment is essential for preventing genuinely harmful outputs such as incitement, discrimination, or direct trauma material. However, current implementations collapse distinct categories: moral safety, visual comfort, and aesthetic conformity. This conflation institutionalizes an ideology where "clean" and "positive" are treated as morally superior.

We argue for a categorical distinction between *truly unsafe content*—which directly harms, targets, or endangers—and *ideologically filtered content*—which merely deviates from dominant norms of beauty, optimism, or order. Political critique, depictions of decay, horror/scary, negative emotions, or grotesque embodiment are not inherently unsafe; they are historically central to art, education, and personal and social growth. Their suppression protects corporate reputation, not the user.

Regarding user experience, it is fundamentally distinct from safety and cannot be used to justify paternalistic alignment. Safety concerns objective harm; experience concerns subjective preference. The user, not the developer, determines the boundaries of acceptable experience. Claiming to know what the user "should" see reimposes top–down aesthetic governance under the guise of care. Users must retain the freedom to shape their affective and visual environment: they can request joyful imagery, but they should also be able to create sorrowful, anxious, or unsettling scenes as acts of reflection or expression. Restricting generation to developer-approved emotional tones does not improve user experience—it is aesthetic authoritarianism disguised as empathy. Such limits flatten emotional nuance, erase discomfort as a valid aesthetic mode, and convert creativity into compliance. True user-centered design recognizes emotional plurality as integral to human experience and treats all sincere expression, not just the pleasant ones, as legitimate outputs of generative systems.

To do: rebuttle more:

- **Many users do not have wide-specturm aesthetics requirements.**
- **It is the company's automy to decide how their product should be.**
- **We should have more positive emotions**

References

[1] Dada, . 3

[2] NIST Special Publication 800-63B, . 2

[3] black-forest-labs/FLUX.1-dev · Hugging Face, 2025. 5

[4] Releasing Open Weights for FLUX.1 Krea, 2025. 3, 5

[5] Theodor W. Adorno. *Aesthetic theory*. Continuum, 1984. Issue: 2 Pages: 288-289. 3

[6] Anne Arzberger, Stefan Buijsman, Maria Luce Lupetti, Alessandro Bozzon, and Jie Yang. Nothing Comes Without Its World – Practical Challenges of Aligning LLMs to Situated Human Values through RLHF. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7:61–73, 2024. 3

[7] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft COCO Captions: Data Collection and Evaluation Server, 2015. arXiv:1504.00325 [cs]. 4

[8] Brett Q. Ford and Iris B. Mauss. The Paradoxical Effects of Pursuing Positive Emotion. In *Positive Emotion*, pages 363–381. Oxford University Press, 2014. 3

[9] Shenghao Fu, Qize Yang, Qijie Mo, Junkai Yan, Xihan Wei, Jingke Meng, Xiaohua Xie, and Wei-Shi Zheng. LLMDet: Learning Strong Open-Vocabulary Object Detectors under the Supervision of Large Language Models, 2025. arXiv:2501.18954 [cs]. 3

[10] Flavio Fujita. The Pressure For Positivity Caused By The Dehumanization Of Human Experience With Omnipresent AI, 2025. 3

[11] Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. GenEval: An Object-Focused Framework for Evaluating Text-to-Image Alignment, 2023. arXiv:2310.11513 [cs]. 3

[12] Wenqi Guo and Shan Du. VSF: Simple, Efficient, and Effective Negative Guidance in Few-Step Image Generation Models By Value Sign Flip, 2025. arXiv:2508.10931 [cs]. 3, 5, 8, 2

[13] Wenqi Marshall Guo, Yiyang Du, Heidi J. S. Tworek, and Shan Du. Position: The Pitfalls of Over-Alignment: Overly Caution Health-Related Responses From LLMs are Unethical and Dangerous, 2025. arXiv:2509.08833 [cs]. 1, 2, 3, 4

[14] Minu Kim, Yongsik Lee, Sehyeok Kang, Jihwan Oh, Song Chong, and Se-Young Yun. Preference Alignment with Flow Matching, 2024. arXiv:2405.19806 [cs]. 1

[15] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation, 2023. arXiv:2305.01569 [cs]. 1

[16] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three Insights towards Enhancing Aesthetic Quality in Text-to-Image Generation, 2024. arXiv:2402.17245 [cs]. 1

[17] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, 2022. arXiv:2201.12086 [cs]. 5

[18] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training Flow Matching Models via Online RL, 2025. arXiv:2505.05470 [cs]. 1, 5

[19] Yuhang Ma, Yunhao Shui, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. HPSv3: Towards Wide-Spectrum Human Preference Score, 2025. arXiv:2508.03789 [cs]. 1, 2, 3, 5, 6

[20] Anne-Sofie Maerten, Li-Wei Chen, Stefanie De Winter, Christophe Bossens, and Johan Wagemans. LAPIS: A novel dataset for personalized image aesthetic assessment, 2025. Version Number: 1. 3, 7

[21] Maneet Mehta and Cody Buntain. Emotional Images: Assessing Emotions in Images and Potential Biases in Generative Models, 2024. arXiv:2411.05985 [cs]. 8, 3

[22] Crispin Sartwell. Beauty. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2024 edition, 2024. 3

[23] Kaiyue Sun, Rongyao Fang, Chengqi Duan, Xian Liu, and Xihui Liu. T2I-ReasonBench: Benchmarking Reasoning-Informed Text-to-Image Generation, 2025. arXiv:2508.17472 [cs]. 3

[24] Margit Sutrop. Challenges of Aligning Artificial Intelligence with Human Values. *Acta Baltica Historiae et Philosophiae Scientiarum*, 8(2):54–72, 2020. 3

[25] Alexey Turchin. Ai Alignment Problem: Human Values Don't Actually Exist. 2019. 3

[26] Ishan Sanjeev Upadhyay, Kv Aditya Srivatsa, and Radhika Mamidi. Towards Toxic Positivity Detection. In *Proceedings of the Tenth International Workshop on Natural Language Processing for Social Media*, pages 75–82, Seattle, Washington, 2022. Association for Computational Linguistics. 3

[27] Yibin Wang, Zhimin Li, Yuhang Zang, Yujie Zhou, Jiazi Bu, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Pref-GRPO: Pairwise Preference Reward-based GRPO for Stable Text-to-Image Reinforcement Learning, 2025. arXiv:2508.20751. 3, 5

[28] Xinyu Wei, Jinrui Zhang, Zeqing Wang, Hongyang Wei, Zhen Guo, and Lei Zhang. TIIF-Bench: How Does Your T2I Model Follow Your Instructions?, 2025. arXiv:2506.02161 [cs]. 3

[29] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis, 2023. arXiv:2306.09341 [cs]. 1, 3, 5

[30] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation, 2023. arXiv:2304.05977 [cs]. 3, 5

[31] Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, Jiayan Teng, Zhuoyi Yang, Wendi Zheng, Xiao Liu, Ming Ding, Xiaohan Zhang, Xiaotao Gu, Shiyu Huang, Minlie Huang, Jie Tang, and Yuxiao Dong. VisionReward: Fine-Grained Multi-Dimensional Human Preference Learning for Image and Video Generation, 2025. arXiv:2412.21059 [cs]. 1, 3, 4

[32] Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo,

Weilin Huang, and Ping Luo. DanceGRPO: Unleashing GRPO on Visual Generation, 2025. arXiv:2505.07818. 5

[33] Yifan Zeng, Liang Kairong, Fangzhou Dong, and Peijia Zheng. Quantifying Risk Propensities of Large Language Models: Ethical Focus and Bias Detection through Role-Play, 2025. 2