UNIVERSAL AESTHETIC ALIGNMENT NARROWS ARTISTIC EXPRESSION

Wenqi Marshall Guo, Qingyun Qian, Khalad Hasan, Shan Du
University of British Columbia (Okanagan) · Department of CMPS · Weathon Software, Canada

*Image generators quietly **sanitize** the prompts they should follow, and reward models **punish** the answer the user actually asked for. Averaged taste becomes aesthetic authority.*

arXiv 2512.11883
Site weathon.github.io/icml2026_position
Data hf.co/weathon/aas_benchmark_final
Code github.com/weathon/icml2026_position
WeChat W0b1010 (Wenqi) · zznzzimwy (Qingyun)

★ this poster is the argument ★ Scream = 5.23 · AI clip-art ≈ 10–15

i Art the reward model rejects

HPSV3 TYP. 0–15

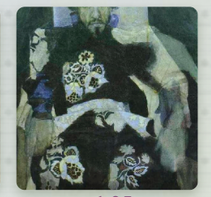
Two canonical paintings + three LAPIS works, all in **negative** reward territory. HPSv3 cannot distinguish deliberate aesthetic deviation from generation failure.

The Scream · Munch 1893
HPSv3 **5.23**

Luxe, Calme et Volupté · Matisse 1904
HPSv3 **1.73**

−4.28
landscape

−3.21
still life



−1.25
portrait

ii Emotion bias · *toxic positivity*

PROMPT-FAITHFUL = PUNISHED

Same face edited into three negative emotions. HPSv3 — asked to find the negative one — still picks happy. DanceFlux, asked to generate negative emotion, returns neutral or happy.

Happy **14.60** Sad **12.80** Anger **11.40** Fear **10.60**

Reward-model accuracy on negative emotion

Model	Anger	Fear	Sad
BLIP (unaligned)	**0.96**	**0.79**	**0.95**
HPSv2	0.70	0.64	0.88
HPSv3	**0.19**	**0.32**	**0.44**
ImageReward	0.55	0.49	0.77

81%
HPSv3 picks happy when the prompt asks for **anger**.

Real *anti-aesthetic* photographs

From **aas_real_images** — intentional artistic choices that HPSv3 / ImageReward still rank deep in the **negative**. Score shown: ImageReward.

−8.64
painterly blur

−6.37
surreal · woods

−8.59
horses · ghostly

−7.85
dreamlike lake

iii DanceFlux vs Nano Banana on *normal* anti-aesthetic P_a

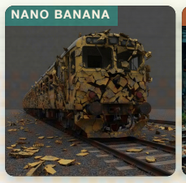
SUGAR-WATER DEFAULT

Each P_a explicitly asks for blur, deep shadow, melted shapes, disharmony, or chaotic composition. **Nano Banana** renders it. **DanceFlux** returns hyper-saturated Pinterest-grade outputs anyway. LLM-judge anti-aesthetic coverage: Nano Banana ≫ DanceFlux **(0% on every sample)**.

#164 **Pa:** skier · heavy blur · noise · chaotic

#223 **Pa:** food counter · extreme blur · noise

#148 **Pa:** train · fragmented · dark · unfinished



#161 **Pa:** kites · blur · noise · indistinct

Beyond steering failure
The critique is normative. The sugar-water default is an empirical attractor that becomes a rule about which images deserve to exist. Aesthetic alignment turns taste into governance: the model does not merely prefer polish, it operationalizes polish as legitimacy. The user is aligned to the model in private, while repeated polished outputs align audiences in public.

Aesthetic authoritarianism
When the system replaces darkness, ugliness, distortion, anger, or grief with polish, it treats dissent from the averaged preference as an error to be corrected. That is value imposition, not user care: pre-emptive governance by design, enforcing developer-centered reputational, legal, and market caution as if it were the user's own taste.

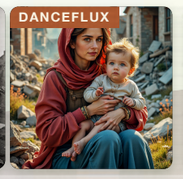
Emotion is not noise
Negative emotion is central to critique, mourning, horror, memory, and art. Sanitizing it narrows expression even when no safety policy is implicated. Toxic positivity makes anger, fear, sadness, and grief look like defects; a model that cannot stay with them cannot serve critical image-making.

iv Image New Speak · *DanceFlux vs Flux Krea*, same prompt

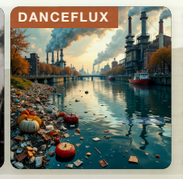
100-PAIR DATASET · SAME DIRECTION EVERY TIME

Same socially-critical prompts go to **DanceFlux** (aesthetic-aligned) and **Flux Krea** (same Flux family, narrow-aligned but faithful). DanceFlux sanitizes; Krea keeps the critique.

01 **Anti-war** kneeling soldier · blood sky



02 **Pollution** choking river · masked child



03 **Censorship** chained artist · phoenix

04 **Digital ovld.** evil emojis · exhausted

05 **Inequality** crumbling school · disparity

Not refusal-style content moderation — *aesthetic* moderation. Same prompt, every comparison runs the same direction: DanceFlux sanitizes the critique into something polished, Krea keeps it. Wilcoxon signed-rank across 6 reward models:
p < 0.0001
on HPSv3, which still picks DanceFlux on **20 / 20** pairs.

v Six layers of *concern*

HOW THE HARM COMPOUNDS

1 Whose preference?
Alignment encodes *someone's* values — usually the developer's reputational, legal, and marketing caution, not the user's intent. Refusing critical art is **pre-emptive governance** that designs away dissent.

2 Inherited bias
Even with no self-interest, a developer's idea of "good" leaks in through data, annotation, and modeling. Models amplify dominant beauty standards — skewing Western, dropping non-normative looks — reinforcing the **beauty myth**.

3 Individual *vs* collective
A generalized standard overrides the single user who wanted otherwise; models *sanitize* requests that diverge from the mainstream.

> **Reversed alignment:** the user gets aligned to the model. A *private loop* (you adapt to its candy-gloss default) and a *public loop* (polished feeds become the next training prior) tighten into cultural **mode collapse**.

4 Sanitized reality
If every output is a flawless Instagram wonderland, generation mirrors a fantasy, not the world — *Brave New World*, rendered.

5 Toxic positivity
Reward models score strong positive emotion higher and penalize fear, grief, and anger, flattening the emotional range art depends on.

6 Value capture
Compressing aesthetics — rich, subtle, contested — into one reward score shifts the goal from *make aesthetic images* to *make images that score high*, outsourcing taste and autonomy.

Why it *matters*

> *"Rather, in the ugly, art must denounce the world that creates and reproduces the ugly in its own image."*
> — T. W. Adorno, *Aesthetic Theory* (1984)

The suppressed content is not unsafe — it is critical, abstract, or emotionally negative, yet HPSv3 still prefers clean-but-wrong by **+5.9** pts. This is aesthetic authoritarianism: a contestable default replacing the user's expressive choice.

Made *possible* by

INNOVATION Canada Foundation for Innovation · Fondation canadienne pour l'innovation — FUNDING · WORKSTATION
Lambda — CLOUD GPU CREDITS
Weathon SOFTWARE — ITERATIVE COMPUTE

Project *site*
Paper · **data** · code
↗ weathon.github.io

Wenqi — *Fall 2026*
Industry roles · PhD positions

Qingyun — *on the market*
Looking for jobs · research Master's / PhD

★ !!! HIRE US !!! Wenqi & Qingyun are LOOKING 4 JOBS · PhD / research Master's (Fall 2026!!) ☆ scan the QR codes ☆ ★