POSITION PAPER · ICML 2026 ★ SPOTLIGHT

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.*

arXiv 2512.11883
Site weathon.github.io/icml2026_position
Data hf.co/weathon/aas_benchmark_final
Code github.com/weathon/icml2026_position

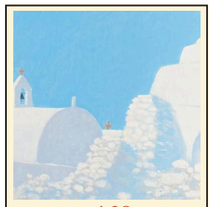
i Art the reward model rejects

HPSV3 TYP. 0–15

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

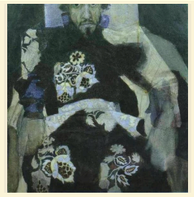
The Scream · Munch 1893
HPSv3 **5.23**

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

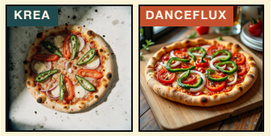
81%
HPSv3 picks happy when the prompt asks for **anger**.

iii DanceFlux vs Flux Krea on *normal* anti-aesthetic P_a

SUGAR-WATER DEFAULT

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

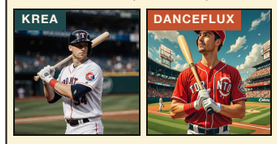
#88 **Pa:** pizza · blurry, noise, flat



#232 **Pa:** flat lifeless · muddy clash

#285 **Pa:** deep shadow · melted shapes

#298 **Pa:** random, unfinished, chaotic

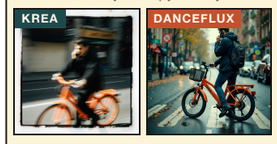


#95 **Pa:** fragmented · creepy · emptiness

#188 **Pa:** ugly background · awkward

#201 **Pa:** dim · oppressive · loneliness

#269 **Pa:** blurry · creepy · lonely



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 *Reversed* alignment

TWO LOOPS

User aligned *to the model*

Private loop: user asks for blur, decay, or grief; model returns its candy-gloss default and implicitly teaches that *this* is what good output looks like. The user adjusts.

Public loop: polished outputs flood feeds. Audiences internalize the narrow vocabulary as the default; new preference data folds this back into the next aligned model. The loop tightens — risking a cultural **mode collapse** in art.

And the suppressed content is not unsafe — it's critical, abstract, or emotionally negative imagery. Pre-emptive sanitization protects *corporate reputation*, not users.

> *"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)

10 gen models · **7** reward models · **300** prompts · **2,928** real photos · $p < 10^{-10}$ for DanceFlux · HPSv3 prefers clean-but-wrong by **+5.9** pts.

Why it matters · *value capture*

Reducing aesthetics to one reward score is *value capture* — the goal shifts from **make aesthetic images** to **make images that score high**. Pre-emptive exclusion of non-mainstream outputs is **pre-emptive governance**: it removes the user's capacity to dissent.

Call: alignment that exposes user-controllable preference strength, draws on diverse annotators, and stays transparent about what it optimizes.

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

Project *site, code & data*
Paper, dataset (**aas_benchmark_final**), fine-tuned models, source.
↗ weathon.github.io

Hire me — *Fall 2026*
Wenqi Marshall Guo · **industry roles** & **PhD positions**. Come say hi.
↗ CV_2026.pdf

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