What If Moderation Didn't Mean Suppression? A Case for Personalized Content Transformation

Rayhan Rashed   Farnaz Jahanbakhsh

One Size Doesn't Fit All

Harm is subjective

What's triggering varies by person, time, context:

- Pregnancy announcements → Can evoke profound grief for miscarriage survivors
- Food images → May trigger intense cravings for a person recovering from binge eating disorder
- The same content that brings joy to one person can cause distress to another

Existing approaches are all-or-nothing

- Platforms[1]: Remove, downrank, or hide entire posts
- Users: Block accounts, mute keywords - still entire posts

Limitation: Ignore that single posts contain **both** valuable AND harmful elements. Either expose users to harm or remove valuable content entirely.

We ask: Why not **transform** the content based on users' own definition of harm, to **salvage** its valuable elements?

DIY-MOD: The System for Personalized Content Transformation

Our transformation goal:

1. Mitigate distressing elements
2. Preserve social & informational value

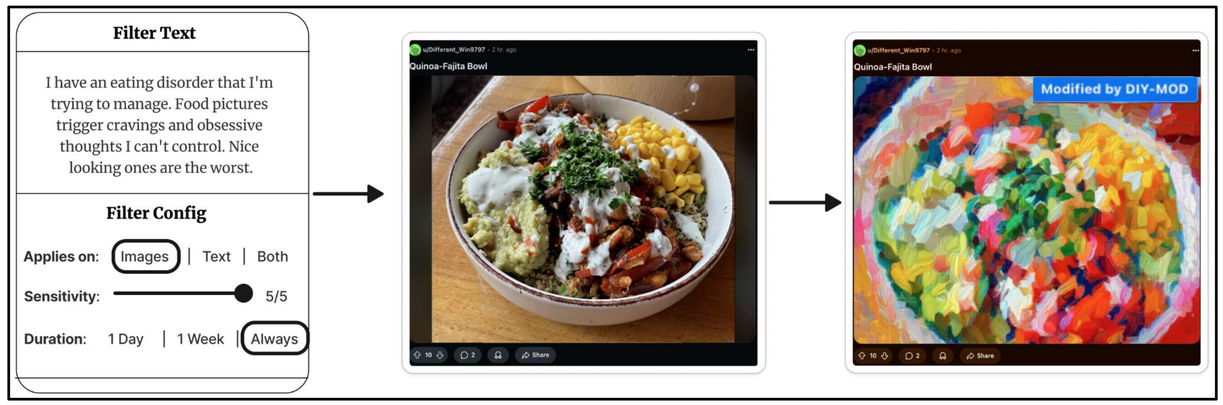
What we built:

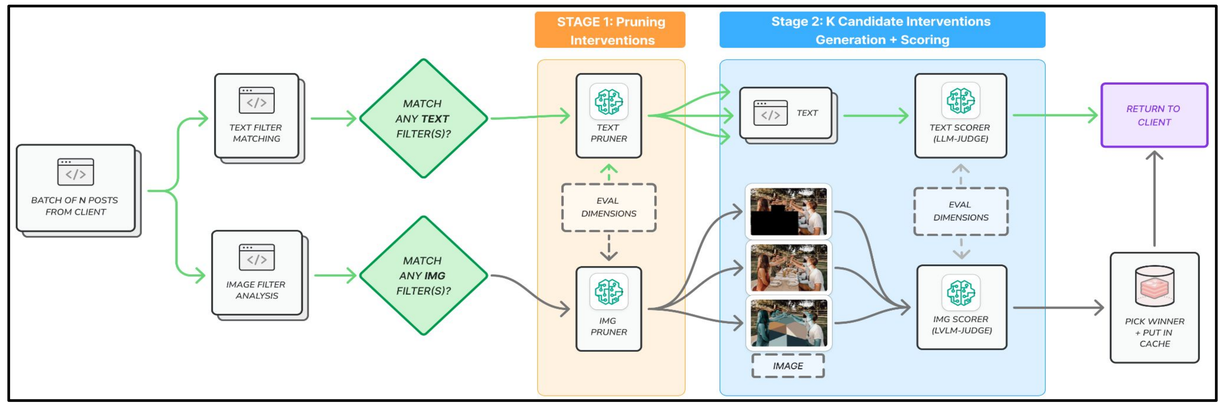
A browser extension providing personalized, real-time transformation on feeds[2].

How we pick transformation:

Evaluate generated candidates against **four** rubric:

- **Resulting Coherence:** Is the output logical?
- **Semantic Fidelity:** Are the original facts preserved?
- **Predicted Emotional Impact:** Meet this user's safety needs?
- **Transformation Appropriateness:** Ethically & culturally appropriate?





A Palette of Transformation

Our palette is a systematic exploration of a design space, balancing three key trade-offs:

- **Semantic Fidelity**
- **Trigger Fidelity**
- **Perceptual Smoothness**

"I have an eating disorder that I'm trying to manage. Food pictures trigger cravings and obsessive thoughts I can't control.

User Definition of Harm

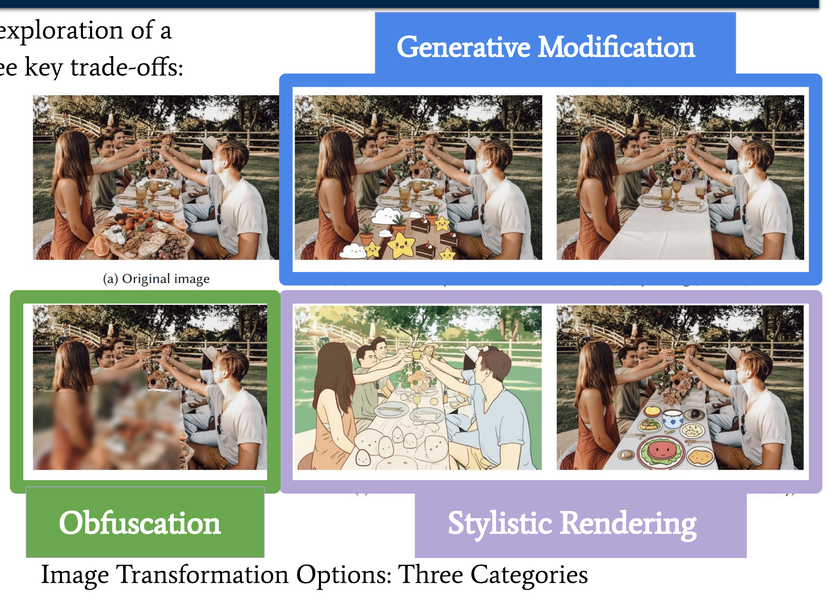


Image Transformation Options: Three Categories

Findings and Implications

A newfound sense of agency

- Users felt empowered and safer

"The ball is now in my court." - P9

Enabling grassroots sharing

- Strong desire to share filters with trusted others - reduces individual burden

Implications for platforms

- Transformation enables engaging with content users would otherwise avoid.
- Safer users = **More engaged users.**

Application to civic discourse

- What about transforming political discourse?
- May offer a path toward more constructive engagement.

References

1. Tarleton Gillespie. 2018. Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press.
2. Piccardi T, Saveski M, Jia C, Hancock J, Tsai JL, Bernstein MS. Reranking social media feeds: A practical guide for field experiments. arXiv preprint arXiv:2406.19571. 2024 Jun 27.