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

Rayhan Rashed   Farnaz Jahanbakhsh

COMPUTER SCIENCE AND ENGINEERING
UNIVERSITY OF MICHIGAN

When One Size Doesn't Fit All: The Subjective Nature of Harm

Harm is subjective

What's triggering varies by person, time, context

- Pregnancy announcements → Evoke profound grief for miscarriage survivors
- Food images → Triggers intense cravings for a person recovering from binge eating disorder

Current tools operate at post-level

- Platforms: Remove, downrank, or hide entire posts
- Users: Block accounts, mute keywords - still entire posts
- **Problem:** Posts often contain both valuable AND harmful elements

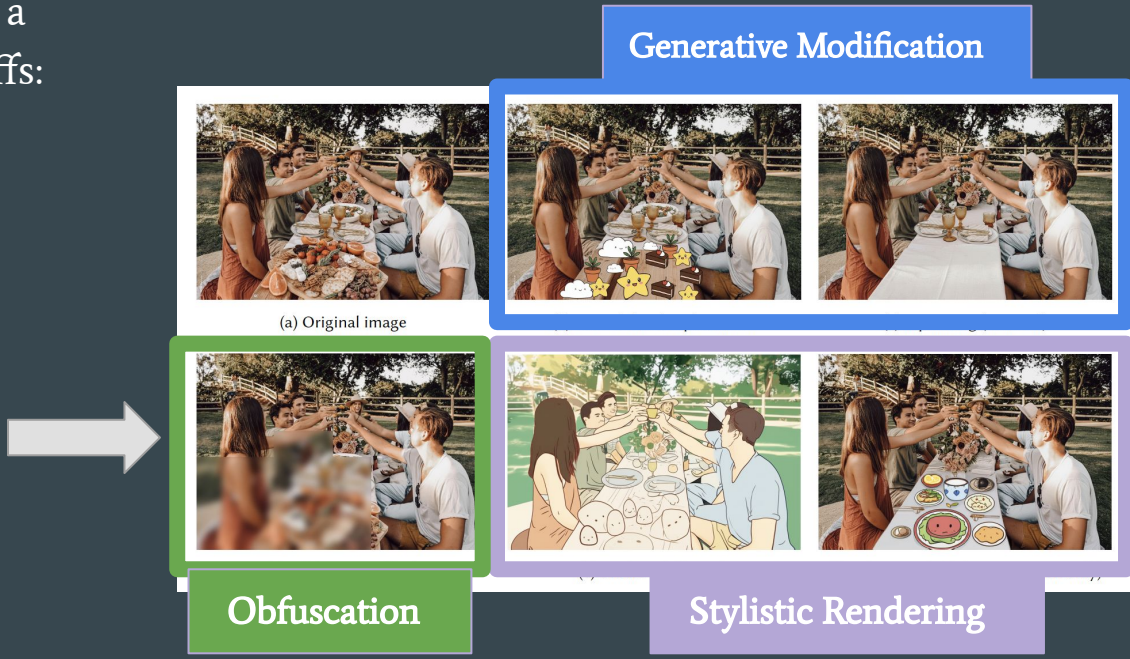
We ask: Why not transform the content based on users' own definition of harm, to salvage its valuable elements?

A Palette of Transformation Options

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. Nice looking ones are the worst."



For Text: Interventions range from blurring specific phrases to rewriting entire sentences.

Image Transformation Palette: Three Categories

DIY-MOD: The System for Personalized Content Transformation

Content transformation goal:

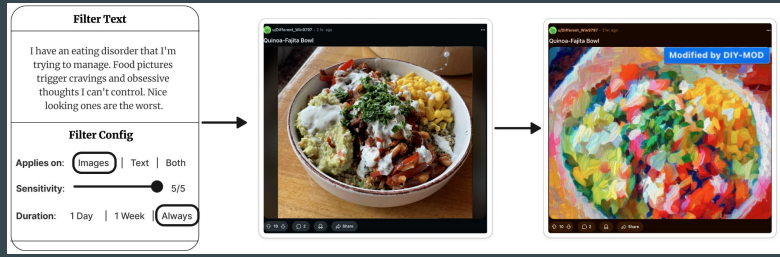
1. Mitigate personally distressing elements within a content
2. Preserve the social and informational value of the content

How we pick transformation:

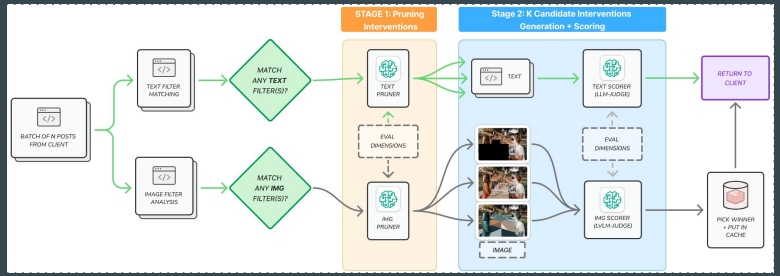
Evaluate generated candidates against rubric

- **Resulting Coherence:** Is the output logical?
- **Semantic Fidelity:** Are the original facts preserved?
- **Predicted Emotional Impact:** Does it meet the user's specific safety needs?
- **Transformation Appropriateness:** Is the transformation ethically and culturally appropriate?

What we built: A browser extension providing personalized, real-time transformation on feeds.



User defines **harm** → DIY-MOD transforms



The Impact: What We Found & Platform 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.

Rayhan Rashed  Email: rayrash@umich.edu COMPUTER SCIENCE & ENGINEERING UNIVERSITY OF MICHIGAN