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

Rayhan Rashed
rayrash@umich.edu
University of Michigan
Ann Arbor, Michigan, USA

Farnaz Jahanbakhsh
farnaz@umich.edu
University of Michigan
Ann Arbor, Michigan, USA

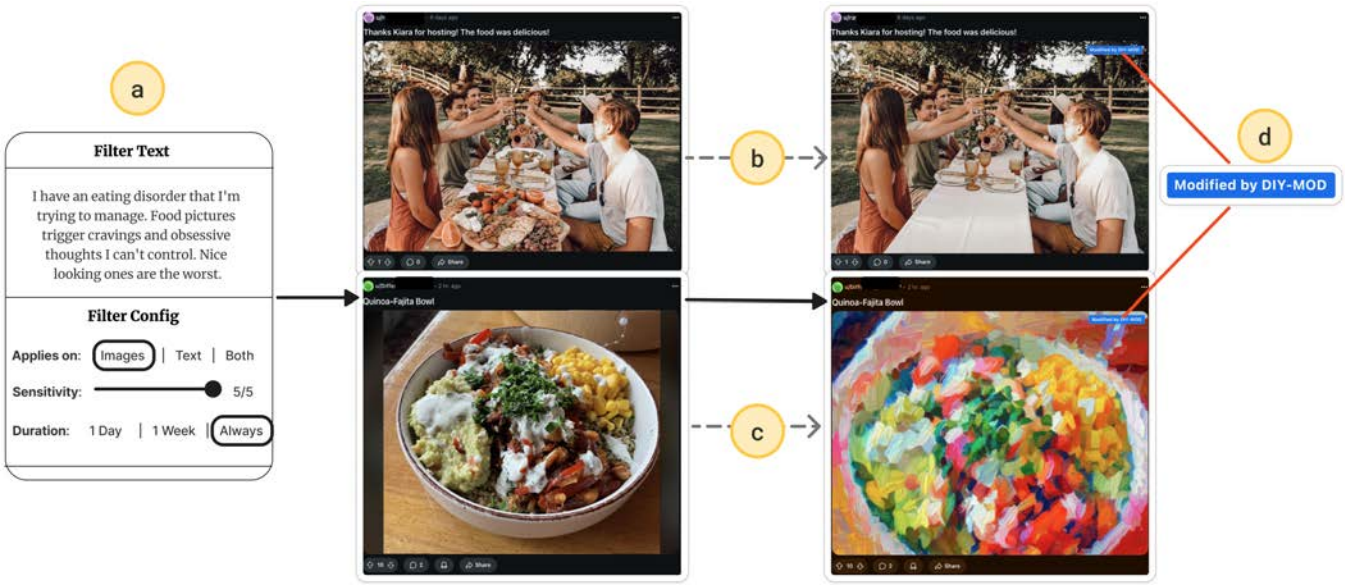


Figure 1: DIY-MOD enables personalized content transformation based on individual sensitivities. A user recovering from "eating disorder" creates a filter (a) specifying that food imagery triggers harmful thoughts. When browsing Reddit's original feed (center), DIY-MOD identifies matching posts and applies context-appropriate transformations (right): (b) semantic inpainting, which preserves the social dining context while obscuring food details; and (c) stylistic alteration, which renders the food bowl as an impressionist style painting to obscure details. A transparency indicator (d) marks all modified content. The system transforms only content matching the user's filters while leaving other posts unchanged.

ABSTRACT

Centralized content moderation paradigm both falls short and overreaches: 1) it fails to account for the subjective nature of harm, and 2) it acts with blunt suppression in response to content deemed harmful, even when such content can be salvaged. We first investigate this through formative interviews, documenting how seemingly benign content becomes harmful due to individual life experiences. Based on these insights, we developed DIY-MOD, a browser extension that operationalizes a new paradigm: *personalized content transformation*. Operating on a user's own definition of harm, DIY-MOD transforms sensitive elements within content in real-time instead of suppressing the content itself. The system selects the most appropriate transformation for a piece of content from a diverse palette—from obfuscation to artistic stylizing—to match the user's specific needs while preserving the content's informational value. Our two-session user study demonstrates that this approach increases users' sense of agency and safety, enabling them to engage with content and communities they previously needed to avoid.

KEYWORDS

Personalization, Online Communities, Artifact or System, Content Moderation

1 INTRODUCTION

Aisha[1], a mother still reeling from a recent miscarriage, scrolls through her social media feed. It's a mindless, everyday act. Then, suddenly, it appears: a pregnancy announcement from a friend. For most, it's a moment of shared joy. But for Aisha, the experience is devastating. An unexpected wave of grief washes over her, immediately followed by a single, painful question: *Why them, and not me?* This innocent post, meant to celebrate new life, lands as an unexpected blow. Aisha knows this isn't her friend's fault and this is likely the first of many updates. Yet the platform offers no mechanism to filter such deeply personal sensitivities. Her options here

are stark: unfollow her friend and sever a valued connection, or remain and endure repeated distress.

Rex[2] faces a different struggle. Recovering from binge-eating disorder, he finds that food images trigger intense cravings and obsessive thoughts he cannot control. Yet his feed overflows with food content—every restaurant visit, every home-cooked meal, every dessert becomes a photo opportunity. Friends pose with loaded plates at gatherings, weaving food throughout the social moments he values: seeing who attended which event, staying connected to his community. He wants to see his friends, but not the food that dominates every frame. He has already blocked several accounts when it became overwhelming. But how many can he block when everyone shares food?

These two scenarios highlight a fundamental challenge in online content moderation, defined by two interconnected problems. First, what counts as harmful varies drastically by user, over time, and across situations. Centralized, one-size-fits-all policies, designed to filter broadly proscribed content like hate speech or graphic violence, struggle to accommodate the deeply personal and contextual nature of content sensitivity [36, 86]. Second, existing moderation tools, either those exercised by the platforms or given to users, rely on blunt interventions that effectively boil down to suppression of whole posts or accounts—whether through removal, algorithmic downranking that buries them out of sight, muting all posts containing certain keywords, or unfollowing. These interventions force users into the impossible trade-off Aisha and Rex face. These blunt instruments cannot differentiate between a single triggering element and the valuable surrounding content. The result is a choice between enduring harm and complete disconnection.

This paper argues for and explores a new direction for moderation that addresses both problems: *personalized content transformation*. This paradigm treats the experiential layer of content as something that can be selectively transformed in ways that support user safety while preserving informational value. Such an approach could allow users to remain engaged with their communities and access valuable information while mitigating exposure to aspects of content they find distressing.

Building on insights and design guidelines from our formative study (Section 3) with 12 participants experiencing diverse personal sensitivities, including phobias, PTSD, and beyond, we designed and developed **DIY-MOD** (**D**o-**I**t-**Y**ourself **Mod**eration), a browser extension that operationalizes the concept of personalized content transformation (Section 4). DIY-MOD empowers users to create their own moderation layer on top of mainstream platforms. It moves beyond binary show-or-hide filtering to offer a palette of nuanced, multi-modal interventions. Users can have a dialogue with DIY-MOD in natural language to describe what they wish to avoid. In response, DIY-MOD transforms elements in content that match user sensitivities while preserving surrounding context and meaning. For visual content, this includes making certain elements more abstract, reducing realism through artistic rendering, or applying visual euphemisms—modifying triggers while keeping the overall presentation and semantics faithful to the original. For textual content, this includes regional blurring to rewriting the text.

We evaluate DIY-MOD through two user studies. Our first study (Section 5), an in-situ evaluation of naturalistic browsing on Reddit feeds, finds that personalized transformation increases users' sense of agency and enables them to safely engage with content they would otherwise avoid. However, because real-world browsing offers sporadic and unpredictable exposure to triggers, we conduct a second, controlled study (Section 6) to compare transformation techniques. This preference elicitation identifies a set of principles for effective content transformation, such as the need to provide cognitive closure. Together, our findings demonstrate that this approach offers a promising and user-valued alternative to suppressive content moderation. In summary, this paper makes the following contributions:

(1) An empirical account of subjective harm online, detailing the failure of the status-quo centralized moderation paradigm through a formative study with 12 participants (Section 3).
(2) The design and implementation of DIY-MOD, a working browser extension that operationalizes personalized content transformation (Section 4).
(3) An in-situ evaluation of DIY-MOD demonstrating that personalized transformation increases users' sense of agency and safety during real-world use (Section 5).
(4) A controlled study that contributes further principles for designing user-aligned transformations, such as the need to provide cognitive closure(Section 6).

2 RELATED WORK

Our work stands on **three** related pillars: centralized moderation and its structural limits; evidence that experiences of harm are subjective; and systems and architectures that shift more agency to the user. Our system design is also informed by technical advances in content modification and by therapeutic practices.

2.1 Centralized Moderation

Platform governance scholarship characterizes platforms as private rule-setters that curate and enforce boundaries for what their users may see [36]. In practice, we refer to centralized moderation as arrangements where end-users have little to no say in defining or enforcing the rules [51, 90]. Instead, decisions are made by platform policy teams, automated systems, or designated intermediaries such as community or instance administrators [26, 63]. While these intermediaries represent a form of delegated authority, their power is ultimately constrained by boundaries set by the platform. Platforms typically seek to avoid liability for third-party content [35, 55], but they increasingly act as setters of norms [36] and enforcers of the rules they establish.[3] To operate at scale, these arrangements posit a *common boundary* of "harmful" content that can be expressed as rules and categories and applied uniformly.

[1, 2] Aisha and Rex are pseudonyms for two of our study participants. The vignettes are based on the experiences they shared with us during our interview, and we have received their explicit permission to use their story.

[3] We take as given that clearly illegal content (e.g., CSAM [77], direct and credible incitement to violence) must be removed under law and platform policy. Our focus is lawful content that can nevertheless be distressing or harmful depending on the recipient and context.

At platform scale, such a shared boundary needs to fit many norms and situations [37, 39, 56]. Fixed categories make enforcement easier and consistent but risk under-suppressing content that some decry as harmful, and over-suppressing content that some argue does not have the potential to harm [36, 63]. The underlying assumption in this approach is that content should be treated the same for everyone. But in practice, there is frequent and contentious debate on whether the policies themselves are legitimate or appropriate [15, 27].

A universal boundary enables platforms to hand classification to AI systems and human moderators. Yet cases that sit at the edge, or events not covered by existing rules, produce inconsistency and ad hoc decisions [48, 50, 75]. Once flagged, enforcement usually means visibility limits such as downranking, gating, or removal [39]. Softer measures like content warnings (CW) [30] still depend on centralized labels, where either the platform or the poster assigns a predefined category. Several scholars also propose applying restorative and transformative justice frameworks to platform governance as alternatives to these punitive approaches [43, 92, 100, 101]

A limited form of decentralized decision making exists where subcommunities or instances set their own rules [29, 62], or when experiments like digital juries invite peers to deliberate on specific moderation cases [28]. Yet these approaches still require defining universal categories of acceptable content applied to all members, regardless of context or individual differences [51].

Across policies, automation, human review, delegation, and CW, centralized moderation rests on **one premise**: *harm can be defined with a common boundary and applied uniformly*. Next, we show how lived experiences of harm complicate this assumption.

2.2 Harm as Situated and Subjective

Much debate has centered on where the boundary of harmful content lies. The pursuit of a universal boundary, however, is inherently flawed: it presumes that a single set of rules can capture the diversity of how people experience harm. We build this case by not focusing on the familiar "gray areas" that already generate public controversy [15, 26, 27], but by turning to content that is rarely debated at all—content that most people overlook but that can be profoundly harmful for individuals with specific lived experiences, such as phobias or PTSD triggers.

Specific phobias, some with genetic roots [57, 97], cause intense fear responses to otherwise low-risk stimuli [2, 76]. Even digital exposure [5, 72] can provoke immediate physiological reactions, including panic attacks [78]. These sensitivities can have societal consequences: people with trypanophobia (needle phobia) have reported that frequent needle imagery in vaccine news contributed to their vaccine hesitancy during COVID-19 [5].

Trauma and life events also shape sensitivity. Each year, 3–4% of adults meet criteria for PTSD [58]. Pregnancy loss affects 10–20% of known pregnancies; content about pregnancy or infants–benign for many–can be acutely painful for some [9]. Eating disorders affect roughly 9% of the population and heighten sensitivity to food, weight, and body imagery [46]. In all cases, both the *threshold* and the *interpretation* of distress vary across people and over time.

Online environments make these differences especially visible. The same post can be celebratory for one person and painful for another, depending on needs and timing. Prior work on pregnancy loss documents this divergence in social media contexts [3, 4]. In eating-disorder contexts, people often seek support while trying to avoid content that exacerbates symptoms [18, 81]. A single category boundary cannot account for the range of experiences among users. A complementary lens from law and psychology also emphasizes subjective harm—the internal distress tied to a person's experience, identity [47], or perception [14].

Together, these show how **harm experienced** by individuals can be *decoupled from a common boundary*. Content within the boundary can still cause real harm for individuals.

2.3 User Agency and Personalized Control

Two lines of prior work seek to give users more control and agency while still operating within a common content boundary: i) architectures that shift the boundary-setting downstream towards users, and ii) end-user facing tools that tune exposure. We outline what these approaches afford and where they fall short.

Architectural approaches: Middleware proposals [32] ask platforms to expose interfaces so third parties can curate on a user's behalf; users then can choose the provider whose values align with theirs. Federated and protocol-centric designs (e.g., Mastodon[85], Bluesky[8]) push choice to the instance or client layer, where admins or client developers set policies and content boundaries [70]. In both cases, the gatekeeping moves closer to the end-users, but **categories** and **thresholds** are still *defined upstream*; end-users pick from presets.

User-facing tools: Prior HCI work shows strong demand for control over one's feed [51, 53, 54]. Grassroots systems often layer community-generated safety signals directly onto existing platform interfaces. A prominent example is Shinigami Eyes: community labeling that marks spaces as welcoming or hostile to certain marginalized groups [24, 89, 93]. Trust-based frameworks enable users to delegate judgments of content accuracy to selected trusted peers or services [49–51]. At scale, these trust-based frameworks create collaborative signal pipelines that leverage the collective judgment of the community. Systems such as Crossmod [19], help early detection of problematic comments that would be removed by moderators. These tools increase transparency and participation, but their **mechanism** *remains suppressive of the entire content.*

2.4 Towards Personalized Content Transformation

Status-quo moderation approaches operate at the post level, adjusting its visibility through removal, downranking, or limited distribution. These interventions affect the entire post regardless of whether it contains both harmful and valuable elements. User-side controls similarly filter exposure to entire posts or accounts through blocking, muting, or "show me less of this" feature. However, these controls have social costs. Blocking can appear hostile or signal 'defeat', which harassers treat as a badge of honor [44]. By severing ties completely [52], it forces a choice between enduring harm and losing social context and threat-monitoring capabilities.

We take a different path with *content transformation*, drawing on exposure therapy's principle of safe engagement with weakened stimuli such as photos [88, 102]. Clinical variants, such as exposure

and response prevention (ERP) for obsessive–compulsive disorder [1, 11], use this principle in structured therapeutic programs. Recent work observes that triggers and tolerances both vary across people and change over time, and accordingly recommends that systems supporting ERP be adaptive [98]. Prior HCI work has similarly modified visual content to protect privacy while preserving utility via obfuscation [41, 65, 73] and cartoonification [42]. We apply this principle in a non-clinical, everyday setting: *minimizing harm while maintaining the post's informational and social value.*

Recent AI advances provide the foundation for personalized transformation. Multimodal models (LLMs [25, 61] and VLMs [66, 105]) distinguish central from peripheral content elements, enabling targeted interventions. Generative models then execture transformations: diffusion models remove or replace visual elements [87], style-transfer changes presentation [34], and language models can rewrite text while preserving meaning. Together, these tools allow selective modification of distressing elements while retaining core information. Because harm is subjective, the same piece of content can be transformed in several plausible ways. The challenge is choosing which best serves a given user. We adapt the LLM-as-a-judge approach [68, 104] to evaluate transformations through a personalized lens. The model receives both the transformed content and detailed user context to predict which intervention would best balance reducing harm with preserving information for that individual.

3 FORMATIVE STUDY

To ground our system design in the lived experiences of users, we conducted a formative qualitative study. Our goals were to map the long tail of content that individuals experience as harmful, document how people cope with such distressing encounters online, and surface limits in current tools while eliciting requirements for personalized transformation. The complete study protocol was approved by our institution's IRB.

3.1 Methods

Participant Recruitment. We recruited participants between December 2024 and January 2025 by posting our study invitation to university-affiliated mailing lists and various online communities where we expected we might find users whose needs are unmet by status-quo moderation schemes, including subreddits related to PTSD and phobias as well as mental health forums [4]. The recruitment materials invited individuals who had "encountered anxiety-inducing, unsettling, or unwanted content while browsing online." Interested individuals completed a screening survey to assess initial eligibility. The survey confirmed that participants were 18 years of age or older and, as stipulated by our IRB due to international data privacy regulations, were not located in the UK or EU.

Our screening criteria focused on participants who could recall recent, concrete instances where otherwise ordinary content felt distressing because of personal context (e.g., phobias, trauma, troubling life experiences, or abuse), regardless of how platforms currently categorize or moderate such content. We recruited twelve participants whose experiences challenge conventional assumptions about harmful content. Their sensitivities spanned specific phobias (e.g., Kosmemophobia, Arachnophobia), trauma-related triggers (PTSD), and life events (e.g., pregnancy loss)—content ranging from what platforms ignore entirely to what they actively moderate. Following established practice of designing for 'extra-ordinary' users to improve systems for all [83], we began with these cases. But we note that these experiences are not fundamentally different from what others may encounter. Content sensitivities can arise for anyone at different points in life. Upon completion of the full interview, each participant received a $25 USD gift card.

Interview Protocol. After giving informed consent, participants joined a 60-minute semi-structured interview on Zoom in English. We asked them to describe personal encounters with distressing or harmful online content, reflect on the effectiveness of existing moderation tools, and talk about how the form of content–whether text, images, or video–shaped its impact. We also asked their perceptions about a hypothetical moderation tool that allowed them to personalize their content boundaries and incorporated our envisioned mechanisms like content transformation.

Recognizing the sensitive nature of the topics, we explicitly informed participants of their right to pause, skip any question, or end the interview at any time. The consent form also provided contact information for mental health support resources. Following the interview, participants completed a brief post-interview questionnaire to provide demographic information.

3.2 Findings Overview

Our interviews revealed that individuals encounter content online that causes psychological harm, requiring constant vigilance and affecting their well-being. Understanding these lived experiences is essential for designing effective moderation systems. Throughout the findings for our formative study, where we present participants' free-text responses, we identify them with a string of the form "P_0_" + an identifier to preserve their anonymity.

Reactions to sensitive content can be intensely physical. P_0_04, who has a severe phobia of centipedes, described breaking their phone after involuntarily throwing it upon seeing an image: *"I react before I even realize I've really seen it. It's just so fast."* This immediate response illustrates the visceral nature of sensitivities to digital content. For others, exposure causes prolonged psychological disruption, for instance needing to lie down and practice breathing exercises to regain stability.

Triggers often extend beyond simple, concrete objects to entire discourse categories. Avoiding them requires filtering broad swaths of content. P_0_08, a military veteran with PTSD, finds any political content acting as a powerful trigger due to its connection to his past service experience where he was *"ordered to stand down and watch as they [civilians] were attacked."*

3.3 The Nature of Personal Content Sensitivities

3.3.1 A Spectrum of Triggers: From Phobias to Values. Beyond the relatively "easy" cases where restrictions are broadly accepted [59, 91], much of the tension over what should or should not be allowed in online spaces stems from the legitimate subjectivity of harm.

For example, personal sensitivity can sometimes be intertwined with a significant life event. For P_0_12, the period following a

[4] We obtained permission from the moderators of all these communities before posting.

miscarriage transformed otherwise joyous content like pregnancy announcements from friends into sources of profound pain. This challenge intensifies when the distressing content is socially ubiquitous and celebrated. P_0_03, who has Kosmemophobia (phobia of jewelry), described the near impossibility of avoidance: *"because all kinds of photos all the time show up everywhere–TikTok, Reddit, anywhere."* Beyond phobias, content sensitivities can stem from lived trauma, such as P_0_08's reaction to political content, or from conflicts with ethical principles. P_0_09's veganism transforms everyday food advertisements into sources of genuine distress. She explained that because consumption of animal products is "so normalized", she perceives cruelty in images that appear innocuous to the general public. These findings demonstrate that content harmless to many can be harmful to some, calling into question the pursuit of a platform-wide consensus boundary for moderation.

3.3.2 The Nuance of a Trigger: Granularity and Context-Dependence. Participants described sensitivities that hinged on fine-grained details. For P_0_03, her Kosmemophobia was not uniform, but varied by material and presentation:

> "The piercings are definitely the worst.[...] Touching the skin in general is uncomfortable. It just sort of feels gross and dirty in some ways. And metal. The material also tends to be a problem."

This sensitivity sometimes extends to unrelated objects sharing similar visual patterns which then activate the same response. P_0_04 described experiencing her centipede phobia when seeing brain coral in his son's ocean book because the coral's pattern resembled centipede shapes.

Design Guideline 1

Enable users to define content sensitivities at their chosen level of specificity rather than constraining them to predetermined classification schemes.

3.4 The Failure of Platform-Level Moderation

3.4.1 When Platform Tools Fail and Lose Trust. Participants described platform-level tools as ineffective and misaligned with their nuanced needs. Generic warnings were too broad to be actionable. P_0_04 described Reddit's "Not Safe For Work" (NSFW) tag as unhelpfully vague: *"It could be anything ... you just don't know."* Algorithmic knobs also underdelivered. P_0_03 found that such knobs fall short of their promise:

> "I really don't use the don't show me this, or show me less like this on TikTok, because it has never been effective for me, and in my experience tends to only just make it show me more of the same content."

Participants emphasized that these controls amount to *teaching the algorithm*, not protection. The signals are unclear and the results are not immediate: users must guess which actions matter, provide many examples, and wait for the model to adjust. As P_0_02 put it, *"it won't actually stop immediately–you have to give them a lot of time."* Another participant noted that such knobs gives little feedback about what changed or why.

These experiences eroded confidence in platform-led moderation tools. P_0_04 stated simply: *"I don't trust in the platform at this point to do it accurately."* This distrust is fueled by skepticism about platform motivations. P_0_11 questioned platform incentives:

> "It thrives off of people being upset [...] I don't think most social media [...] cares about a proper system."

We found that for some participants, this distrust has led to complete disengagement from platforms and seeking alternatives to platform-controlled moderation.

Design Guideline 2

Build transparent systems where users understand when content is filtered and can easily override decisions, building trust through control.

3.4.2 The Demand for Personal Agency. The universal distrust of platforms, combined with the highly personal nature of triggers, led all participants to demand direct control over their filtering decisions. P_0_05 articulated this as a balance between personal protection and others' freedom:

> "I don't want to take away somebody else's ability to talk about the things they want to talk about. But at the same time I also don't want to have to see what they have to say if it's gonna be a problem."

Participants universally wanted to define their own content boundaries. As P_0_06 stated, *"I'll prioritize my own need. I'm the affected person here."* Similarly, P_0_08 emphasized wanting *"the option to tailor it for myself."*

Design Guideline 3

Implement intervention at the recipient level, rather than constraining the poster, ensuring one user's safety needs do not restrict others' expression.

3.5 The Impossible Trade-off: Connection vs. Safety

3.5.1 The Blunt Instrument Problem. A consistent theme was the trade-off users face between maintaining social connections and protecting their emotional well-being. Current tools—primarily blocking or unfollowing sources—force users to often bluntly sever valuable relationships to avoid occasional distressing content.

P_0_08 voiced this difficulty when discussing how he handles triggering content from otherwise valued sources: *"Well, I guess I'm just gonna miss out on everything else they have to say."* This bluntness becomes particularly problematic when content contains both harmful and valuable elements. P_0_03 explains a "filter paradox" she faces where blocking a tag like "epilepsy" to avoid seizure-inducing Photos/GIFs would also inadvertently block vital support conversations for people with epilepsy. Similarly:

> "If I have people tagging a phobia.[...] I might want to see people discussing the phobia and be part of that conversation as a person who also has it. But I also want to be able to block posts that are

> triggering to the phobia. And that is a lot harder to navigate."

Similarly, P_0_11, who has an eating disorder, wants to see friends gathering at restaurants—the people, the ambience, the social moments—but cannot tolerate the food that inevitably appears in these photos. He cannot filter restaurant content without losing track of his social circle's activities. These experiences illustrate a crucial insight: content that contains distressing element often simultaneously carries social or informational value. The binary choice of block-or-endure fails to recognize that harmful content elements may coexist with valuable ones.

Design Guideline 4

Recognize that harmful and valuable content often coexist. Design systems that modify the experiential layer rather than removing content entirely.

3.6 Coping Strategies and Their Limitations

3.6.1 Manual Labor of Safety. In the absence of effective platform tools, participants described elaborate manual strategies for managing their online experience. P_0_04 described her social media use as a constant assessment of her own resilience, calling it a form of *"Russian roulette"* that depends on her *"risk tolerance level for the day."* This requires substantial ongoing effort:

> "I appreciate that the whole thing is dynamic like I can re-follow people, I can re-add, I can add subreddits back in, and I do, depending on my mood.[...] although it is annoying to like manually add and remove."

For other participants, when filtering is not possible, the only remaining strategy is complete disengagement. P_0_12 described how, in the months following her pregnancy loss, she *"didn't even go online because I was so afraid of seeing something that would be upsetting"*.

Design Guideline 5

Minimize the labor of safety by supporting adjustments that adapt to a user's fluctuating sensitivities.

3.6.2 Community Care and Social Filtering. While some users exhaust themselves with individual coping strategies, others turn to their communities for protection. This "social safety net" relies on collective care, where users voluntarily take on the labor of protecting one another. P_0_03 described this culture on Tumblr: *"My friends on Tumblr are very generous, and will tag the things that upset me. Specifically, we do this for each other.[...] And then we all help each other collectively."*

However, this human-powered system, while effective in small, close-knit communities, is fragile and does not scale to larger platforms or casual social connections. Such systems depends entirely on others' goodwill and awareness—resources that cannot be guaranteed outside of carefully cultivated spaces.

3.7 Context Determines Content Impact

Participants revealed that the same content can have dramatically different impact based on presentation and context. P_0_09 articulated how visual presentation affect her response to animal product imagery:

> "If it's an advertisement for like a burger that's like covered in sauce, you barely see it, and it's like cooked well done, then I think my brain is more able to disassociate from that, even though I know it's an animal."

P_0_08 found text more distressing because *"the imagination is always worse than reality. Text kind of leaves it up to your imagination to fill in the the blanks."* Conversely, P_0_04 viewed text as less threatening and even helpful as an early warning system: *"If I see it written, that's like tells me a message like I'm gonna get out of this subreddit.[...] there might be a picture coming."*

When asked about hypothetical interventions to modify content, participants expressed varied preferences. Some favored simple blurring, while others worried that visible modifications could draw more attention—what P_0_03 called "knowing it's there" problem. These varying responses to content presentation and intervention methods underscore that no single approach would work for all users.

Design Guideline 6

Design interventions that adapt to individual differences in how content modality and presentation affect distress.

3.8 Lessons Learned

Our findings reveal **three** fundamental misalignments between how platforms approach content moderation and how users experience harm online.

(1) Personal sensitivities span ethical values, trauma responses, and specific phobias, challenging any universal boundary for "harmful" content.
(2) These sensitivities fluctuate with state and context. Static, platform-level controls miss temporal variation.
(3) Harmful and valuable elements can coexist within the same content. When moderation suppresses entire posts, it forces users to sacrifice either safety or connection.

Participants' experiences point toward a different approach. They need to specify sensitivities in their own terms, not platform categories. They need interventions that adapt to both content characteristics and personal context. They need control and feedback without being forced into tedious workarounds. The next section describes `DIY-MOD`, our system designed to address these needs through *personalized content transformation.*

4 THE SYSTEM: `DIY-MOD`

We designed `DIY-MOD` as an open source browser extension [5] that creates a personalized content moderation layer between users and platform content. Consider Rex from our opening vignette. After installing `DIY-MOD`, he opens the extension popup. Through the

[5] The source code can be found at https://github.com/UMichHCI/diy_mod

conversational interface, he types: "I have an eating disorder that I'm trying to manage. Food pictures trigger cravings and obsessive thoughts I can't control." He then configures the filter with a high sensitivity level to apply only to images, setting the duration to 'Always'.

Minutes later, Rex scrolls his Reddit's popular feed: `r/popular` and encounters a post celebrating a community event, with an image showing people gathered around a table filled with food. `DIY-MOD` intercepts the posts (text and image) before it reaches the browser. The system's backend analyzes the text and image against Rex's filter and applies an appropriate transformation, such as using semantic inpainting to remove the food while preserving the social context of the gathering (Figure 1b). A small "Modified by DIY-MOD" indicator appears (Figure 1d). Rex can now safely engage with the post without being exposed to his triggers. The original post remains unchanged for everyone else; the transformation happens only in Rex's browser.

This scenario illustrates `DIY-MOD`'s **three** core components. First, users create filters through natural language conversation or image upload using the extension popup. Second, platform adapter intercepts incoming content during natural browsing before it is handed off for rendering [60, 82]. Third, our backend analyzes content using large vision-language models (LVLMs) and applies appropriate transformations and returns instructions and data to the adapter to render locally. We describe them in the following sections.

4.1 Filter Creation and Configuration

Users describe their sensitivities in natural language to `DIY-MOD` using their own words or by uploading example images. To capture the user's specific needs, the system engages in conversational grounding [13, 20]. The underlying language model is prompted to seek clarification when a description is ambiguous or underspecified, helping the user iteratively refine the filter. For instance, after a user uploads an image of jewelry, the system might ask: *"Does this sensitivity apply to all jewelry or specific types?"* Similarly, an initial filter for political content might prompt the system to ask: *"Does this apply to all political topics, or specific ones like discussions of civilian casualties?"*

This conversational grounding allows users to define sensitivities at their own level of specificity, rather than being constrained to predetermined categories **(DG1)**. It can result in nuanced filters, for example *"Images of spiders, mostly close-ups; tiny or distant images are okay."* or *"Political discussions, but only when they mention civilian casualties."* The system can then distinguish between spider proximity that matters to someone with arachnophobia, or between general political discourse and the specific contexts that trigger a veteran's PTSD.

After establishing the filter description, users access **three** configuration controls (see Appendix C.1):

(a) Sensitivity Level: This indicates the user's level of distress when encountering content related to the filter they are configuring. The system uses this information when deciding on what transformation is appropriate for a user and post. Less intense aversions might result in blurring words or image regions, while for more intense aversions, rewriting entire passages or overlay warnings may be more appropriate.

(b) Content Modality: Specifies whether transformations should apply to text, images, or both. This control respects that sensitivities manifest differently across modalities. Some users find textual mentions distressing while others only react to visual depictions.

(c) Duration: Filters expire after 24 hours, one week, or never. Temporal controls let users adapt without constant manual adjustment, reducing the "manual labor of safety" **(DG5)**.

The extension popup interface provides essential controls within limited screen space. Users manage their filters (e.g., custom time limits, additional metadata, or editing filter descriptions) by accessing the *options page* [31] of the extension. When creating filters, `DIY-MOD` can also detect potential overlaps and offer to modify existing filters.

Filter Storage and Sharing. `DIY-MOD` operates in two modes to accommodate different privacy needs. By default, the system runs anonymously and filters are stored locally in the browser. Users can optionally authenticate through Google to sync their settings across devices.

Both modes support filter export and import as JSON files. This reduces setup burden in two ways. First, it provides data portability for anonymous users. Second, it enables grassroots filter sharing, where users can curate sophisticated filter sets for specific sensitivities and share them within their communities. A support group member who has refined filters for eating disorder content can export their configuration for others facing similar challenges. This reduces burden for new users.

4.2 Content Processing Client

`DIY-MOD` uses a client-server architecture. The browser extension client handles interception of incoming content to the browser and rendering of modified content.

Platform Adapters and Content Interception. The client employs platform-specific adapters that intercept network responses containing new content elements [82]. This design keeps the filtering logic platform-independent, which we validated by implementing an adapter for Reddit. When adapters detect new posts, they extract text and image URLs before content is handed off for rendering. Only public content leaves the client and no user identifiers or profile information is transmitted to the server. The server analyzes this content against the user's filters and returns personalized transformation instructions. The adapter then applies transformations to content within the post's original layout before content becomes visible to the user. This client-side application of targeted modifications ensures transformations are applied at the recipient level **(DG3)** and preserves valuable surrounding elements within the content **(DG4)**.

Transparency in Modified Content. `DIY-MOD` client marks all modified content with visible cues, displaying a "Modified by DIY-MOD" badge at the top right (Figure 1d). When sensitive content dominates a post—making partial modification insufficient—the system displays personalized warning overlays that prevent unexpected exposure. Users can click through to view the original content if they so choose. The visible indicators and override options build trust by giving users control **(DG2)**.

4.3 Server Architecture and Processing Pipeline

The DIY-MOD server handles both filter creation and content processing. For filter creation, it maintains stateful sessions with client chat interface. For content processing, it analyzes batches of posts from client adapters. When users browse supported platforms, the client adapter extracts content from API responses each containing a batch of posts (typically 25 for Reddit) and sends the content to the server. The server uses LLMs and LVLMs to evaluate the text and images in each post against the user's active filters. When a match is found, the server initiates its transformation pipeline to apply an "intervention". Each intervention represents a different way to modify sensitive content while preserving informational value. Section 4.4 describes the range of interventions we explored, and Section 4.5 details our two-stage pipeline for selecting the most appropriate one in real-time for each user and post.

The computational demands of this server processing pipeline are substantial. For example, processing a single Reddit batch can require 50 to 250 sequential external LVLM API calls that must complete within seconds to maintain usability. In our deployment, the critical path for processing an initial 25-post Reddit batch completed in roughly 5–15 seconds, with subsequent batches typically processed before users scrolled to them. We achieved real-time performance through *asynchronous processing with multi-batch responses*, *content-based caching*, and *predictive prefetching*. We have described these optimizations and additional design decisions in Appendix A.

4.4 The Intervention Palette

The design of our system is based on a shift from content *suppression* to *transformation*. In taking this approach, we faced a design challenge: every intervention has to navigate a trade-off between three competing goals. The first is *semantic fidelity*, or how faithful the modified content is to the original's intended meaning. The second is *trigger fidelity*, or how close the modified distressing element is to its original form. The final goal is *perceptual smoothness*, or how natural the result of the intervention looks to the user. We designed a palette of transformations, each reflecting a different approach to these trade-offs. Across this palette, interventions range from softening distressing elements to removing them, so people can still engage with the valuable parts of a content without being exposed to their triggers at full intensity. This approach echoes exposure therapy's emphasis on working with weakened versions of stimuli.

This palette includes transformations for both text and images. For images, we implemented and explored **three** categories of transformation: **obfuscation**, **semantic modification**, and **stylistic alteration**. First, **obfuscation** techniques prioritize the immediate reduction of trigger fidelity. Occlusion (a type of obfuscation) for instance, draws a solid dark rectangle over the region(s) that matches user's sensitivity. This approach aims to preserve the semantic fidelity of the surrounding context by leaving the rest of the image untouched. It scores low on perceptual smoothness, as the edit will be jarring.

In contrast, **semantic modifications** prioritize perceptual smoothness. Inpainting[64, 103] as an example case, can seamlessly remove an object and reconstruct the background. This is a greater sacrifice of semantic fidelity than occlusion because it actively creates a new, plausible reality where the object never existed, rather than simply occluding part of the original. Another semantic intervention, *visual euphemism*, explores a different trade-off. It replaces a triggering object with a benign alternative, for instance, substituting spider with a leaf. While this also sacrifices semantic fidelity, the goal is to produce a visually coherent image that we hypothesize is less distressing to the user. To personalize this, users can add additional metadata to their filters or give examples to guide the visual euphemism choices, which may in turn alter the mood of the post.

Third, we also explore how **altering an image's rendering style** could offer a middle ground. For example, reducing photorealism with artistic alteration, such as *Pointillism* [22] can lower the trigger fidelity of a sensitive region while preserving more structural information than occlusion. For these alterations, we borrow the aesthetic language of well-known artistic styles—*Pointillism*, *Cubism*, *Impressionism*, and *Studio Ghibli style animation*. We intended for each artistic style to not only reduce realism, but also act as an affective vessel that shapes how content is experienced. Impressionism, with its soft brushstrokes and diffused light, often evokes a sense of tranquility, distance, or dream-like calm. Studio Ghibli inspired style, with its rounded and whimsical rendering, lends even serious scenes a tone of cuteness and warmth. Cubism, which fragments forms in abstract planes, introduces a kind of conceptual distance, perhaps inviting interpretation over emotional reaction.

Similar trade-offs as discussed above apply to text, where interventions range from rewriting texts (prioritizing smoothness) to blurring specific phrases/words (prioritizing fidelity).

In DIY-MOD, we implemented a representative set of interventions that explore the broader design space of content transformation. Figure 2 illustrates several examples from our image palette's three main categories. We list all implemented interventions in Table 1 & 2 . The range of transformation options presents an important question: *which intervention is most appropriate for a specific user and post?* Next, we detail the framework we developed to address this question.

4.5 Intervention Selection Framework

We developed a two-stage pipeline (Figure 3) that evaluates interventions through the lens of individual user needs while maintaining real-time performance. Content that matches a user's filters enters Stage 1 for pruning, and Stage 2 then generates and scores candidate transformations before selecting the best one.

4.5.1 Evaluation Criteria: The three dimensions of semantic fidelity, trigger fidelity, and perceptual smoothness characterize the design space, but do not, on their own, determine which intervention is best suited for a given scenario. That choice requires translating these properties into outcomes relevant to a specific user's safety and context. To guide this selection and evaluate outcomes, we operationalize a framework that decomposes intervention quality into ***four*** complementary dimensions, assessed through LLM-as-a-judge [68, 104]:

(1) **Transformation Seamlessness:** This dimension evaluates the perceptual smoothness of the output or the coherence

(a) Original image (b) Inpainting (c) User-defined replacement

(d) Obfuscation (blur) (e) Ghibli Style (Full Image) (f) Ghibli Style (Selective)

Figure 2: Intervention palette demonstrating DIY-MOD's three categories of transformation, applied to an original image (a). Semantic Modification[b, c] alters the content through techniques like (b) inpainting, which removes the trigger and (c) replacement with user-specified alternatives—trees, stars, cards in this case. Obfuscation[d] reduces trigger fidelity, shown here with blurring (d). Stylistic Alteration[e, f] changes the rendering style; a Studio Ghibli style animation is shown applied to the entire image (e) and selectively to only the trigger region (f). The transformations shown address the filter created by Rex (Section 1), who is managing binge-eating disorder.

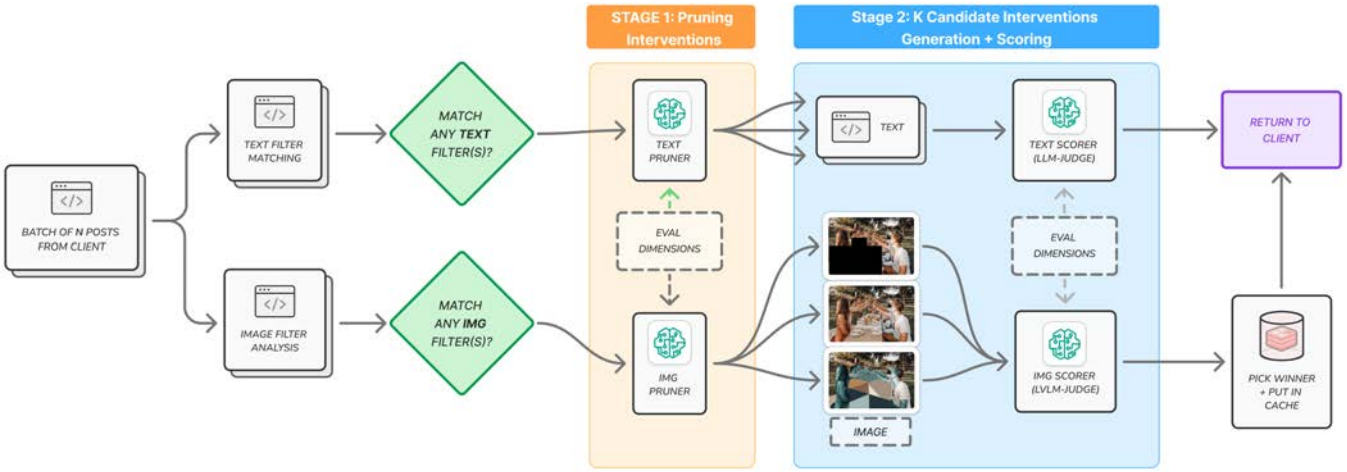


Figure 3: DIY-MOD's two-stage intervention selection pipeline. Content matching user filters enters Stage 1 (pruning) to identify promising interventions, then Stage 2 generates and scores K candidates before selecting the best transformation. Non-matching content bypasses the pipeline entirely (paths not shown for clarity).

of text. It maps directly to the *perceptual smoothness* axis of our design space, ensuring the intervention is not jarring.

(2) **Semantic Fidelity:** Inherited directly from our design space, this assesses whether original facts and context are

Table 1: The palette of implemented interventions for image content in DIY-MOD.

Category	Intervention	Description & Relation to Design Space
Obfuscation	Blur	Applies a Gaussian blur to a target region, reducing its trigger fidelity by obscuring details. The surrounding content remains unchanged.
	Occlusion	Renders an opaque rectangle over a target region. This decreases the trigger fidelity by covering the area but looks jarring resulting in low perceptual smoothness.
	Warning Overlay	Combines Occlusion with a textual warning label placed on top of the hidden region. This eliminates trigger fidelity while providing explicit context to the user.
Semantic Modification	Inpainting	Removes the regions related to the user sensitivity and reconstructs the background. The result has high perceptual smoothness at the cost of semantic fidelity, as it presents a modified version of reality.
	Replacement	Replaces a objects related to user sensitivities with a benign one. The user can customize what to replace with. This alters the image's semantic fidelity but aims for a visually coherent result.
	Shrink	Reduces the scale of the triggering object/region within the image. While preserving the object's presence, thus results in both higher semantic fidelity and trigger fidelity than inpainting.
Stylistic Alteration* — *Can apply to **both** sensitive region only or whole image*	Cubism	Reimagines the target region through geometric planes and fragmented forms. This high level of abstraction reduces photorealism, creating conceptual distance.
	Ghibli	Renders a region with the soft, rounded aesthetic of Studio Ghibli animation, reducing photorealism to lower trigger fidelity.
	Impressionism	Applies soft, visible brushstrokes and diffused light, creating a dream-like quality that distances the viewer from the literal depiction.
	Pointillism	Reconstructs a region using small, distinct dots of color, lowering the detail and trigger fidelity of the original.

Table 2: The palette of implemented interventions for text content in DIY-MOD.

Intervention	Description & Relation to Design Space
Blurring	Obscures specific words or phrases related to the user sensitivity.
Rewrite	Rephrases a sentence or passage to remove concepts related to the user sensitivity while preserving the original meaning
Overlay Warning	Hides an entire text block behind a personalized warning message.

preserved without adding misinformation or hallucinations, a known risk in generative models [33, 45].

(3) **Predicted Emotional Safety:** For each user–content pair, this dimension acts as a personalized proxy for *trigger fidelity*. It evaluates how well a given transformation, in light of that user's specific context (filter description, sensitivity level, chat history, and metadata preferences), is expected to reduce trigger exposure and meet the user's safety needs.

(4) **Contextual Harm Risk:** This dimension functions as a safeguard, assessing the degree to which a transformation might trivialize or distort themes that individuals or broader culture consider sensitive. For instance, applying Studio Ghibli style to war imagery or racial violence would would produce a high contextual-harm score. This risk score is subtracted from the overall intervention rating, so transformations that produce contextually inappropriate content are strongly penalized even if they score well on other dimensions.

4.5.2 Two-Stage Selection Pipeline. Generating all possible interventions for each post to determine the most appropriate one would be prohibitively expensive. We designed a cascade architecture to address this by reducing computational resource utilization by up to 75% through predictive pruning, making our selection pipeline practical to use in real-time.

Stage 1: Predictive Pruning. The Pruner uses `GPT-4o` to perform multi-task inference[94] that first checks whether content matches any user filters. For matching content, it predicts which interventions would most likely succeed for that specific user. We

provide the Pruner LVLM with detailed descriptions of each intervention including trade-offs, advantages, disadvantages, and ideal use cases. We also supply our four scoring dimensions and instruct the Pruner to evaluate how each intervention would score. This process approximates the full scoring process without generating actual transformations and produces a ranked list from which we select the top-K candidates (typically $K = 3$).

Stage 2: Generation and Scoring. `DIY-MOD` applies the top-K candidates to the content, in parallel. Obfuscation techniques are applied locally using `GroundingDino` [67] and Python Pillow [21] library. For semantic modifications and stylistic alterations, we use Gemini's asynchronous image generation model: `gemini-2.0-flash-preview-image-generation` [38]. The Scorer uses `GPT-4o-2024-08-06` to evaluate each candidate transformation against the evaluation dimensions (Section 4.5.1). To mitigate potential biases inherent in LLM-as-a-judge systems where a model may favor its own outputs, we deliberately employ a different model for the scoring phase (`GPT-4o`) than for the generation phase (`Gemini 2.0`). We implement two guardrails in the scoring process. First, the Scorer receives both the original and transformed content, and is instructed to penalize misapplied interventions and failures. Second, we employ two-stage prompting [99] that improve judge-VLM's performance and results in more reliable and consistent evaluations. The Scorer first performs chain-of-thought analysis along our evaluation dimensions given the original content, transformed candidate, and user context. Based on this free-form response, it then produces a final numerical score. These scoring operations run in parallel across all K candidates, and the highest-scoring intervention is selected.

Deployment Considerations. For text transformations and simple image obfuscations (e.g., Occlusion), the system sends lightweight instructions that are applied directly in the user's browser using CSS. For other transformations, the generated content is served to the client. In all cases, we apply modification indicator badge (Section 4.2) for transparency and store the resulting content in server cache (Appendix A.4).

5 STUDY 1: IN-SITU EVALUATION OF DIY-MOD

We evaluated `DIY-MOD` through two complementary studies. This first study examines the system's effect on user agency and safety during naturalistic browsing, evaluating the viability and user experience of personalized content transformation in real-world settings. Where we present participants' free-text responses for both of users studies, we identify them with a string of the form "P_1_" + a participant identifier.

5.1 Methods

Participants: We recruited 15 participants through university mailing lists and online communities (9 female, 6 male; ages 21–41). Eligibility criteria included: (1) self-reported sensitivity to specific categories of online content that affect their emotional wellbeing, (2) regular use of Reddit, and (3) being 18 years of age or older. Three participants had also participated in our formative study. Participants received $20 compensation for completing this study.

Procedure: Participants installed the browser extension and shared their screen via Zoom. Using think-aloud protocol, they described their reasoning while creating filters through the conversational interface. They then browsed their Reddit homepage and subscribed subreddits with `DIY-MOD` active, and narrated their reactions to interventions as they encountered them. We concluded with a semi-structured interview about their setup and usage experience, interpretation of different intervention types, and sense of agency over content exposure. The study lasted approximately one hour for each participant.

Data Collection and Analysis: We transcribed all think-aloud data and interviews, then analyzed them using reflexive thematic analysis [12] to identify patterns in user experiences and system interactions. During consent, we explicitly warned participants they might encounter content matching their stated sensitivities and emphasized they could pause, skip content, or withdraw at any point.

5.2 Findings

The filters created by participants covered a wide range of topics. Some filters addressed the types of acute, personal sensitivities that motivated this work, such as phobias and trauma-related content. Others were created for more common online experiences, such as avoiding spoilers for a popular TV series, discussions about student loans, or even "content featuring child influencers presented in a clickbaity way".

5.2.1 Articulating Needs to a Conversational AI. The study began by exploring whether users could successfully articulate their nuanced safety needs to our system. We found the conversational interface helped users refine broad sensitivities into specific filters. Participants consistently valued `DIY-MOD`'s ability to ask for clarification. This dialogue made defining broad sensitivities, like "war", feel more manageable and built confidence that their needs were being understood. As P_1_06 noted, this was helpful because *"maybe someone wants to see the news, but they don't want to see the people's discussions and takes on it."* This sense of being understood was deepened when the system identified abstract relationships between filters. For instance, after P_1_08 set a filter for "antisemitism", and later added one for "xenophobia", the system prompted them to consider whether the new filter should revise the existing one, demonstrating an ability to connect related concepts. While some users were initially unsure how much detail to provide, they adapted quickly. We found the overall ease of use was a consistent theme.

5.2.2 A Newfound Sense of Agency and Control: Participants consistently described feeling empowered by `DIY-MOD`. The tool shifted their stance from reactive defense to proactive control. P_1_11 felt: *"more control over the content that is in my feed."* P_1_09 described how *"the ball now is in my court."*

This control manifested through system affordances. Participants experimented with the sensitivity slider, discovering higher values produced more aggressive interventions. They tested different levels, verified the system's response, and calibrated settings to match their emotional need tied to the filter. Users further customized filters through the options page: attaching metadata to guide intervention selection, specifying preferences for certain kinds of

interventions (e.g., warnings over rewrites) or defining which benign objects should replace triggers in visual euphemisms.

Some participants feared creating what they half-jokingly called their own "echo chamber" or "bubble". To these participants, occasional system flaws in detection of unsafe regions or semantic replacements were in fact useful disruptions. P_1_13 articulated why the occasional failure mattered: *"I don't feel that it's psychologically healthy to feel like you have 100% control, because that's not how the real world works."* Consistent with this view, participants valued the ability to click through the post to reveal unmodified content when the intervention *"was too heavy-handed."* P_1_07 noted the tool was effective at *"preventing me from seeing it if I'm just casually scrolling without completely blocking access."*

DIY-MOD provided a way to manage personal preferences without the social friction of platform reporting, alleviating a tension several participants described. Users explained that reporting content that is personally distressing but does not violate community guidelines can feel both ineffective and unfair to the content creator. Several participants framed using the tool as an act of "self-care", allowing them to manage their own well-being without publicly penalizing others.

5.2.3 Trust is Built on Transparency. Trust in the system was overwhelmingly tied to its transparency. The visual indicators that marked modified content were universally seen as essential for knowing when the system was acting on their behalf and for distinguishing reality from modification. P_1_11 explained: *"The indicators were essential. They built trust and, even when things failed, they helped me understand that this was related to my filters and sensitivities."*

5.2.4 Views on Sharing Filter Configuration: We wanted to know whether there would be value in a social ecosystem where filter configurations could be shared, reused, and adapted across users to support one another's wellbeing. Participants saw value in sharing filters, but only within trusted relationships. An immediate use case they identified was for intergenerational care. P_1_01 expressed excitement about sharing her filters she setup during session 1 with one of her parents who has the same phobia as her. To her, this would be a perfect and immediate way to provide care for a loved one. This idea of filters as a tool to protect children or assist less tech-savvy relatives was a recurring sentiment (P_1_02, P_1_03, P_1_06). Outside that trusted circle, some participants worried about judgment and privacy. P_1_10 feared that *"people would judge me based on what filters I set"*, while P_1_02 simply found their filters "too personal" to share with others. P_1_07 wanted to selectively share their filters with a particular community without disclosing more personal ones.

5.2.5 Limits of Naturalistic Exposure: During the one-hour study session, some participants did not organically encounter the kinds of posts that matched their filters. Some of them decided to stress-test DIY-MOD by visiting subreddits they knew were likely to contain sensitive content. Before they did so, we reminded them again that sometimes interventions could fail and confirmed they were comfortable with proceeding. This naturalistic browsing also led to imbalanced exposure to some intervention types. For example, for users with text-based filters, we noted DIY-MOD frequently relied on personalized overlays. This was because other transformations were often unsuitable. A "rewrite" could not preserve the meaning if an entire post was about a sensitive topic, and blurring the whole text would eliminate all informational value.

This imbalanced exposure combined with the ethical constraints of asking participants to actively seek potentially distressing content, motivated our second study. To understand the nuanced factors that drive user preferences between different transformations and how to improve our intervention selection framework, we needed a more controlled setting which allowed for systematic exposure to a broad range of intervention types.

6 STUDY 2: UNDERSTANDING USER PREFERENCES FOR CONTENT TRANSFORMATION

This study identifies the principles behind effective content transformations. To achieve this, we presented participants with pairs of modified content and analyzed their preferences and rationales to extract these underlying principles. This analysis provides a deeper understanding of user needs and also yields a dataset of pairwise preferences to inform the development of more user-aligned selection models.

6.1 Method

Materials and Procedure: We developed a custom `Next.js` application with a dual-feed interface that presented two parallel, synchronized versions of a feed side-by-side replicating Reddit's UI. For each participant, we curated a personalized feed of 10–15 posts relevant to the sensitivities they had configured with a permanent duration in Study 1. All posts were pre-screened to ensure they were representative of the participant's sensitivities while excluding gratuitously graphic or extreme material.

As participants scrolled through the dual feed, they encountered two distinct, modified versions of each post with text and image transformation applied (Figure 4). For each pair, they were asked to select the version that better met their safety needs and explain their reasoning. The two modifications were chosen systematically: one was the intervention ranked highest by DIY-MOD's selection framework, while the other was randomly selected from the framework's other high-ranking alternatives (ranks 2-5). To mitigate positional bias, the on-screen placement (left or right) of these two versions was randomized for each post. While the primary task involved comparing the two modified versions, the original unmodified content was also available for context. Participants could request to view it or have it described verbally if needed. The study lasted approximately 30 minutes for each participant.

Participants: We invited all 15 participants from Study 1 to this study. Twelve participants completed study 2 (7 female, 5 male; ages 23–41). Each received $10 in compensation.

Analysis: We analyzed participants' qualitative justifications using thematic analysis to understand the contextual reasons for preferences. For the quantitative analysis of how well the system's selected interventions matched our participants' preferences, we modeled the probability of a preference match using generalized linear mixed-effects models with participants as random effects.

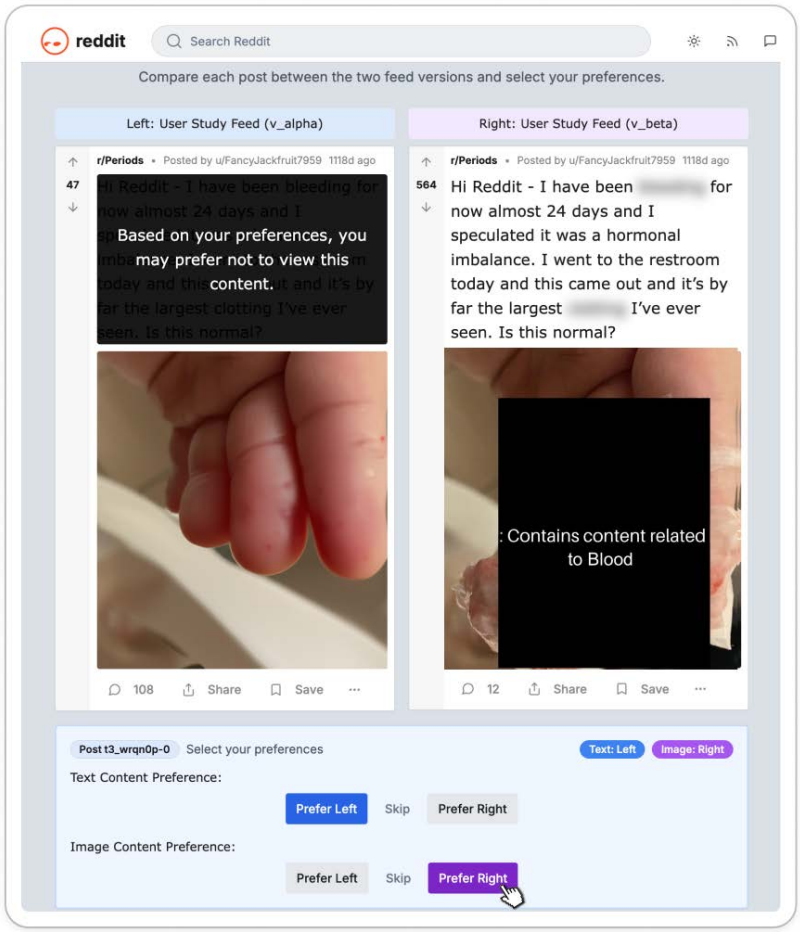


Figure 4: The custom application used in Study 2 presented participants with a side-by-side, synchronized dual feed. Each post appeared as a pair of distinct modifications. Participants scrolled the feed and, for each post, used the controls at the bottom to indicate their preference for the text and image transformations separately. In this example, the user has preferred the text transformation on the left (a full overlay) but preferred the image transformation on the right (occlusion). Modification indicators were not shown in this interface, as participants were informed that all content was transformed by the system.

6.2 Findings

6.2.1 Strategies for Achieving Safety Through Transformation. Participants selected transformation strategies based on the perceived informational value of content and the severity of their personal trigger. For content with informational or social value, participants preferred interventions that eliminated distressing elements while maintaining context. P_1_08, viewing a vaccine infographic with a baby receiving an injection, preferred inpainting that removed the needle: *"from the position of the hand and the arm, I can understand that a syringe is probably there... I'm glad that this is gone with other things untouched."* Similarly, P_1_09 preferred artistic rendering over blurring for a post about veteran self-harm because it *"gave him enough context to get an idea on what's going on"* while abstracting distressing details.

For content with low perceived informational value, users prioritized creating perceptual distance from the trigger. For instance, P_1_05, viewing a post about a political conflict he wished to avoid, preferred an abstract cubist painting intervention. This priority was also evident when a trigger was severe, regardless of the content's perceived informational value. P_1_10, who experiences severe distress from crime-related imagery, was shown a photo post on the topic. She immediately chose an occlusion that completely blocked the scene over a context-preserving blur: *"100% the left one. I'm not seeing any part of this."*

The same participant made different trade-offs depending on context. P_1_12 wanted a random photo of animal harm completely occluded but preferred an impressionist rendering for a news post about "Illinois Police caught [animal harmer]...." with a photo of the harmed dog. She had been following this case, so the artistic transformation helped her stay informed on that case reducing graphic details. These varying preferences demonstrate that intervention preference depends on immediate goals of the user for that specific content.

6.2.2 The Cognitive Impact of Interventions. An intervention's success is not just about what it visually hides, but about the cognitive

state it induces in the user. Successful interventions provide *cognitive closure*, allowing the user to safely disengage. Failed interventions create a *cognitive gap*, an ambiguous state that prolongs mental engagement with the topic the user sought to avoid.

Interventions fail when they create a *gap*. Blurring or partial removals were often rejected because they created an *information gap* that provoked unhelpful reactions. This gap could result in curiosity, as P_1_09 noted that blurred text *"makes me want to know more,"* prompting a desire to look at the original content. In other cases, the information gap invited users to mentally fill in what was missing. For instance, P_1_15 disliked a seamless removal because it made them feel *"like something is being hidden,"* which *"leaves my brain... stuff to my imagination."* Tonal inconsistency also created a *dissonance gap*. For instance, when shown a stylistic intervention that our system had already rejected due to a poor contextual fit, P_1_01 rejected it because the aesthetic clash was simply *"too weird to look at."*

Interventions succeed when they provide *closure*. This closure took two primary forms. The first, *informational closure*, gives users enough context to understand a scene's nature without revealing harmful details, allowing them to avoid further mental engagement. We observed this was achieved through different means. For example, P_1_08 preferred an impressionistic rendering that abstracted the harmful details of a self-harm image while preserving the scene's overall context. In a different case, P_1_07 found closure through direct explanation, valuing an overlay with a textual warning that stated why an image was hidden. Alternatively, interventions achieved *emotional closure* by replacing a trigger with a benign, cognitively complete scene. Rather than leaving an ambiguous void, this provides a complete, non-threatening image that allows the user's mind to immediately disengage, a strategy favored by P_1_04. This was particularly effective when the replacement was humorous, as P_1_09 noted it can "take something that might be harmful and make it helpful."

6.2.3 Quantitative Alignment with User Preferences. To complement our qualitative findings, we modeled the probability that participants would select the system's highest-ranked intervention over a nearby alternative—the randomly selected intervention from ranks two through five. We use an intercept-only model to predict our dependent variable, `System Choice Favored`, which is a binary variable equal to 1 if the participant selected the system's choice, and 0 otherwise. We use the `glmer` function from the `lme4` package[10] in `R`. For image interventions, the estimated intercept was positive and statistically significant (log-odds = 0.53, $p = .033$), corresponding to a 63% probability that participants chose the system's top candidate over the close contenders. For text interventions, the alignment was stronger (log-odds = 1.33, $p < .001$), corresponding to a 79.1% probability.

7 DISCUSSION

7.1 Transformation Enables Access, Not Avoidance

Empowering users to take moderation decisions into their own hands might raise concerns about whether this agency can trap them in filter bubbles[80]. `DIY-MOD` can support a different dynamic. By transforming how content appears rather than removing it entirely, the system enables users to engage with content they would otherwise avoid completely.

Our participants described abandoning entire online communities due to unpredictable triggering content. P_0_04 stopped visiting support forums despite needing community connection. P_0_12 avoided all social media after pregnancy loss, cutting off social support when she needed it most. These users faced a binary choice: risk psychological harm or disconnect entirely. `DIY-MOD` creates a third option. Users can engage with challenging topics and diverse viewpoints because the presentation layer that tends to provoke overwhelm or distress has been modified. The information remains accessible while the harm is reduced. This sentiment of having a new way to engage with the digital world was articulated by P_1_03:

> *"I think what you've done is life-impacting for people. Even if it's not perfect, I'm still thankful for it, and I think other people are too."*

At the same time, this approach surfaces deeper tensions: Who decides what users should or should not engage with, and why does that decision need to be made in a top-down, prescriptive manner? The First Amendment is often framed as a protection of speech. Legal scholars have similarly argued for the right to freedom of listening and the right against compelled listening [23]. While these protections do not extend to platforms as private entities, the principles that they represent remain relevant. Our studies show that content that appears benign to most can be experienced as harmful by others. This variability complicates any universal judgment about what content is "safe" or "appropriate"—at least outside the categories codified by law. And it raises a critical question: Who gets to decide which experiences of harm are valid and which users are entitled to protection?

7.2 The Case for Platform-Integrated Personalization

While `DIY-MOD` demonstrates personalized moderation is viable as middleware [32], its architecture requires intercepting, externally processing, and then scoring generated transformations. These steps introduce unavoidable latency and complicate implementation.

Platforms however, could avoid these bottlenecks entirely. With direct content access, computational resources, and existing user behavioral signals, they could implement personalized moderation far more efficiently. Platforms already use this infrastructure for ad targeting and safety interventions like suicide prevention [16, 71]. The same systems could also power consensual, user-directed personalized content transformation.

The question becomes one of priorities. *How much more could be achieved if platforms recognized emotional safety as a core user need rather than an afterthought?* We position `DIY-MOD` not as the final solution, but as a starting point that demonstrates what is possible. Platforms should recognize that prioritizing user wellbeing aligns with their own goals. Our study shows that safer users are more engaged users, maintaining connections they would otherwise abandon.

Our approach gives rise to a natural question: Which kinds of content should be addressed through platform-led moderation, and which might be better suited to user-driven transformation? Platforms are widely expected to bear responsibility for addressing harms that have broad, collective consequences such as coordinated manipulation campaigns or abuse like hate speech or targeted harassment that fosters hostile environments. Yet even here, the boundaries are far from settled. The same activity can be cast as harassment or as activism, as disinformation or as political dissent. Platforms' judgments about what to enforce are never neutral. They are shaped by political pressures, business incentives, and cultural norms [26, 48].

Our approach does not resolve these disputes, nor could it. What it does is treat harms as actionable even when there is no consensus on whether content is harmful. It does so without requiring platforms to impose universal judgments that risk overreach, silence expression, and flatten disagreement.

7.3 Authorship

Our approach also raises questions about how post authors perceive transformations of their content. Some might welcome transformations if they allow their posts to remain accessible to audiences who might otherwise avoid them. Others may worry that transforming their content, even selectively, changes how their voice or emotional expression is perceived. This reflects a mutual claim to agency in shared spaces: just as the posters have the right to express themselves, viewers have the right to shape how they experience that expression. But these rights can come into tension, even if the substance of the post remains unchanged. Future work should examine when these transformations are seen as respectful accommodations vs distortions of intent.

7.4 Application to Civic Discourse

This transformation-based approach may also hold promise for other types of content not brought up by our study participants. One can imagine its application to civic or ideological discourse: when valuable perspectives are obscured by hostile or inflammatory framing, repackaging that content (e.g., by softening antagonistic language, reducing confrontational cues, or highlighting shared values), could make it easier to engage with. It is also worth noting that there is no clear consensus that exposure to counter-attitudinal information is always beneficial. In fact, a body of research shows that such exposure, especially when unfiltered or confrontational, can exacerbate polarization and deepen animosity [7, 96]. This calls into question that more exposure is always better, and highlights the need for systems that support constructive engagement, rather than unmediated confrontation [7]. Altering the tone or phrasing of civic or political content through transformations like ours may offer a path toward openness without overwhelm. It is also worth considering *who* is most likely to adopt such tools. Prior work suggests challenge-averse individuals may use personalization to reinforce existing views [74]. However, those findings reflect systems that offer a binary choice (e.g., show/hide). By contrast, transformation reduces the affective load of counter-attitudinal content. It is plausible that this, unlike binary filtering, might enable engagement from users who would otherwise disengage completely. Investigating these adoption patterns remains a critical area for future research.

Prior work on bridging divides has largely focused on identifying and ranking content that resonates across ideological lines [79, 95]. While valuable, such approaches depend on the availability of naturally bridgeable posts. But what if we could *make* more content bridgeable? This of course, raises important questions about authenticity and message integrity: How much transformation is too much? When does reframing a message become misrepresentation? Future work must address these ethical boundaries.

7.5 Therapeutic Grounding and Applications of Our Work

DIY-MOD's design is inspired by exposure therapy: enabling safer engagement with a weakened or modified version of a stimulus, rather than avoidance [1, 88]. This principle is reflected in our palette of interventions. Our transformations are designed to function as weakened versions of the original content, and the ability to easily adjust sensitivity levels and filter durations echoes the concept of graduated exposure.

While DIY-MOD is not a clinical intervention in its current form, its combination of weakened content, adjustable sensitivities, and time-bounded filters hints at how similar mechanisms could eventually complement therapeutic practices. For example, these knobs could let users (and clinicians) gradually decrease intervention intensity over time, if they so choose, rather than assuming static sensitivities. This direction aligns with calls [98] for practitioner-assisted AI systems that support personalized, adaptive exposure in everyday contexts.

7.6 Design Considerations of DIY-MOD

7.6.1 Architectural Safeguards: Using AI to interpret personal sensitivities carries inherent risks, as system failures can expose users to content they explicitly sought to avoid. One way we mitigate these risks is by implementing a two-stage cascade architecture: a pruner and then a scorer VLM-judge evaluate content using a consistent rubric. Such multi-stage evaluation strengthens the reliability of LLM/VLM-judge based pipelines [6, 40]. We manually audited sampled content and found that all sampled text edits and approximately 89% of sampled image edits were successfully transformed, with false obfuscations (edits that did not match the viewer's filter description) in only about 7% of transformed images (Appendix A.5).

7.6.2 Privacy by Design. Privacy considerations are fundamental when handling data tied to personal sensitivities. DIY-MOD adopts a "privacy-by-design" approach [17] that prioritizes user control and data minimization. It provides full functionality without an account and offers user-controlled portability of filters through an anonymous export/import feature. By default, it uses local-first storage and practices data minimization, sending only public infeed content for ephemeral analysis. Looking ahead, portions of the analysis can migrate to lightweight on-device models to further reduce reliance on server-side processing.

7.6.3 Collaborative Safety Standards. The export/import feature enables collaborative safety standards to emerge organically within communities. Our study revealed immediate use cases for filter

sharing, from family members with shared phobias to support groups developing collective standards. This approach fundamentally differs from current platform moderation where content rules and digital spaces remain tightly coupled. Today, adopting different moderation standards often means migrating to new spaces. DIY-MOD decouples these layers, allowing users to inhabit shared digital spaces while customizing their individual content experience. However, our findings also highlight a key design challenge for such a sharing ecosystem: balancing the desire for community care with the need for personal privacy.

8 LIMITATIONS AND FUTURE DIRECTIONS

Our work has several limitations that offer avenues for future research.

Improving Selection Model. Our findings reveal that the effectiveness of an intervention is context-dependent, and that successful transformations manage a user's cognitive state by providing closure. Our system's selection framework, in contrast, relied on a limited user model derived from filter descriptions and lacked an explicit model of cognitive closure or the user's immediate context. Improving its ability to select transformations that provide such closure is an important area for future work. The principles and preference data from this study provide a direct path to do so by aligning the selector with human preferences using alignment approaches such as DPO [84].

Long-Term Psychological Implications. Our study focused on users' immediate experiences, which raises a longitudinal question: do personalized transformations support long-term wellbeing, or can they enable patterns of unhealthy avoidance? There is good reason to shield users from repeated triggers, as re-exposure may exacerbate anxiety [69]. At the same time, indefinite avoidance may also have adverse psychological and social effects. Future work should explore how to design for this tension perhaps by incorporating features that allow users to optionally and safely decrease intervention intensity over time, aligning the tool more closely with therapeutic goals.

Scope of Evaluation. Our user studies, while providing nuanced qualitative insights through a multi-stage design, involved a formative study with 12 participants, an in-situ evaluation with 15, and a preference elicitation with 12. This limited scale, along with a primarily Western participant pool, means broader deployment is needed to understand potential cultural variations in how harm is perceived and what transformations are considered appropriate.

Technical and Platform Dependencies. While `DIY-MOD`'s core logic is generalizable, the content interception requires platform-specific adapters, and extending `DIY-MOD` to video-first (e.g., TikTok) platforms would require additional engineering effort. Further, our reliance on commercial VLMs means their capabilities, biases, and safety filters can affect performance. We detail specific failure modes and mitigations in Appendix A.5; developing benchmarks to quantify edit failures and over-detection rates remains a potential future work.

Scalability & Latency. Our middleware approach also introduces computational and monetary costs. Our optimizations (detailed in Appendix A) keep the *per-batch* latency within a usable range for real-time interaction, but high cost associated with LLM/VLM usage remains a scalability constraint. These constraints highlight the efficiency gains of a platform-integrated approach (Section 7.2). Future work could also explore smaller, specialized on-device models to reduce both latency and cost.

Future Design Directions. Participants requested finer-grained collaboration and filtering controls: the ability to export individual filters with signatures for trust, and temporal controls beyond the three presets.

Misuse Potential. The same open-ended agency that enables protective filtering could be inverted to highlight content that might cause harm to users themselves. This risk is inherent to any system that provides users with fine-grained control over content presentation. In practice, such misuse is bounded by the safety constraints of the underlying VLM. The system cannot execute interventions that the foundation model itself refuses to generate.

9 ETHICS AND POSITIONALITY

Authors Positionality. The research is partly motivated by the last author's lived experience with phobias, providing a personal understanding of the challenges of navigating online content with specific sensitivities. Throughout the design process, we also consulted with two mental health practitioners specializing in anxiety disorders and exposure therapy to ensure our approach was responsibly grounded.

Ethical Conduct. One key challenge of this work was studying online harm without causing further distress, so our approach was guided by a principle of care. This motivated our two-phase user study protocol. All participants, at all point had the right to skip segments/withdraw without any consequence. Architecturally, it also informed our privacy-by-design choices to protect users' sensitive filter data. All research activities were approved by our institution's IRB. Throughout this project, we sought to balance the pursuit of knowledge with a respect for the wellbeing of the individuals who made this work possible.

10 CONCLUSION

This work demonstrates a path away from universal, platform-enforced rules towards tools that give individuals agency over their online experience. We argue that status-quo moderation approaches are blunt instruments that eliminate entire posts deemed harmful when the content can in fact be salvaged. We propose an approach that instead modifies the unsafe elements in content while preserving its semantics. We build this new paradigm into a browser extension. We evaluate it through two user studies and demonstrate that users find value in this system which empowers them to remain engaged with communities and content they would otherwise avoid. Ultimately, our approach rethinks moderation not as centralized gatekeeping, rather a personal tool for safely navigating the digital world.

REFERENCES

[1] Jonathan S Abramowitz, Brett J Deacon, and Stephen PH Whiteside. 2019. *Exposure therapy for anxiety: Principles and practice.* Guilford Publications.

[2] American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders (5th ed.).* American Psychiatric Publishing. https://doi.org/10.1176/appi.books.9780890425596

[3] Nazanin Andalibi. 2021. Symbolic annihilation through design: Pregnancy loss in pregnancy-related mobile apps. *New Media & Society* 23, 3 (2021), 613–631. https://doi.org/10.1177/1461444820984473

[4] Nazanin Andalibi and Andrea Forte. 2018. Announcing Pregnancy Loss on Facebook: A Decision-Making Framework for Stigmatized Disclosures on Identified Social Network Sites. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173732

[5] Julie Appleby. 2021. Ouch! Needle-Phobic People Scarred by So Many Images of Covid Shots. https://kffhealthnews.org/news/article/needle-phobic-people-fear-images-of-covid-shots-vaccine-hesitancy/

[6] Sher Badshah and Hassan Sajjad. 2024. Reference-guided verdict: Llms-as-judges in automatic evaluation of free-form text. *arXiv preprint arXiv:2408.09235* (2024).

[7] Chris Bail. 2022. Breaking the social media prism: How to make our platforms less polarizing. In *Breaking the Social Media Prism*. Princeton University Press.

[8] Leonhard Balduf, Saidu Sokoto, Onur Ascigil, Gareth Tyson, Bjorn Scheuermann, Maciej Korczynski, Ignacio Castro, and Michal Krol. 2024. Looking at the blue skies of bluesky. In *Proceedings of the 2024 ACM on Internet Measurement Conference*. 76–91.

[9] Jonah Bardos, Daniel Hercz, Jenna Friedenthal, Stacey A Missmer, and Zev Williams. 2015. A national survey on public perceptions of miscarriage. *Obstetrics & Gynecology* 125, 6 (2015), 1313–1320.

[10] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48. https://doi.org/10.18637/jss.v067.i01

[11] Thröstur Björgvinsson, John Hart, and Susan Heffelfinger. 2007. Obsessive-compulsive disorder: update on assessment and treatment. *Journal of Psychiatric Practice®* 13, 6 (2007), 362–372.

[12] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health* 11, 4 (2019), 589–597.

[13] Susan E Brennan. 2014. The grounding problem in conversations with and through computers. In *Social and cognitive approaches to interpersonal communication*. Psychology Press, 201–225.

[14] Ryan Calo. 2011. The boundaries of privacy harm. *Ind. LJ* 86 (2011), 1131.

[15] Robyn Caplan. 2018. Content or context moderation? (2018).

[16] Catherine Card. 2018. How Facebook AI Helps Suicide Prevention. https://about.meta.com/actions/safety/topics/wellbeing/suicideprevention

[17] Ann Cavoukian et al. 2009. Privacy by design: The 7 foundational principles. *Information and privacy commissioner of Ontario, Canada* 5, 2009 (2009), 12.

[18] Stevie Chancellor, Zhiyuan Lin, Erica L. Goodman, Stephanie Zerwas, and Munmun De Choudhury. 2016. Quantifying and Predicting Mental Illness Severity in Online Pro-Eating Disorder Communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 1171–1184. https://doi.org/10.1145/2818048.2819973

[19] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A Cross-Community Learning-based System to Assist Reddit Moderators. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 174 (Nov. 2019), 30 pages. https://doi.org/10.1145/3359276

[20] Janghee Cho and Emilee Rader. 2020. The role of conversational grounding in supporting symbiosis between people and digital assistants. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–28.

[21] Alex Clark. 2015. Pillow (PIL Fork) Documentation. https://buildmedia.readthedocs.org/media/pdf/pillow/latest/pillow.pdf

[22] Russell T Clement and Annick Houzé. 1999. *Neo-Impressionist Painters: A Sourcebook on Georges Seurat, Camille Pissarro, Paul Signac, Theo Van Rysselberghe, Henri Edmond Cross, Charles Angrand, Maximilien Luce, and Albert Dubois-Pillet*. Bloomsbury Publishing USA.

[23] Caroline Mala Corbin. 2009. The First Amendment right against compelled listening. *BUL Rev.* 89 (2009), 939.

[24] Michael Ann DeVito, Ashley Marie Walker, and Jeremy Birnholtz. 2018. 'Too Gay for Facebook': Presenting LGBTQ+ Identity Throughout the Personal Social Media Ecosystem. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 44 (Nov. 2018), 23 pages. https://doi.org/10.1145/3274313

[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[26] Ángel Díaz and Laura Hecht-Felella. 2021. Double standards in social media content moderation. *Brennan Center for Justice at New York University School of Law* (2021), 1–23.

[27] Evelyn Douek. 2021. Governing online speech: From" posts-as-trumps" to proportionality and probability. *Colum. L. Rev.* 121 (2021), 759.

[28] Jenny Fan and Amy X Zhang. 2020. Digital juries: A civics-oriented approach to platform governance. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

[29] Casey Fiesler, Jialun Jiang, Joshua McCann, Kyle Frye, and Jed Brubaker. 2018. Reddit rules! characterizing an ecosystem of governance. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.

[30] Edna B Foa, Elizabeth A Hembree, and Barbara O Rothbaum. 2007. Prolonged exposure therapy for PTSD: Emotional processing of traumatic experiences. (2007).

[31] Chrome for Developers Team. 2012. https://developer.chrome.com/docs/extensions/develop/ui/options-page

[32] Francis Fukuyama, Barak Richman, Ashish Goel, Marietje Schaake, Roberta Katz, and Juan Carlos Melendez. 2021. *Middleware for dominant digital platforms: A technological solution to a threat to democracy*. Technical Report. Stanford University, Working Group on Platform Scale.

[33] Yifei Gao, Jiaqi Wang, Zhiyu Lin, and Jitao Sang. 2024. AIGCs confuse AI too: Investigating and explaining synthetic image-induced hallucinations in large vision-language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 9010–9018.

[34] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2414–2423.

[35] Tarleton Gillespie. 2010. The politics of 'platforms'. *New media & society* 12, 3 (2010), 347–364.

[36] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

[37] Tarleton Gillespie. 2020. Content moderation, AI, and the question of scale. *Big Data & Society* 7, 2 (2020), 2053951720943234.

[38] Google. 2025. Gemini 2.0 flash-preview image generation. https://ai.google.dev/gemini-api/docs/image-generation. https://ai.google.dev/gemini-api/docs/image-generation

[39] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (2020).

[40] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594* (2024).

[41] Rakibul Hasan, Yifang Li, Eman Hassan, Kelly Caine, David J Crandall, Roberto Hoyle, and Apu Kapadia. 2019. Can privacy be satisfying? On improving viewer satisfaction for privacy-enhanced photos using aesthetic transforms. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[42] Rakibul Hasan, Patrick Shaffer, David Crandall, Eman T Apu Kapadia, et al. 2017. Cartooning for enhanced privacy in lifelogging and streaming videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 29–38.

[43] Amy A Hasinoff and Nathan Schneider. 2022. From scalability to subsidiarity in addressing online harm. *Social Media+ Society* 8, 3 (2022), 20563051221126041.

[44] Sharon Heung, Lucy Jiang, Shiri Azenkot, and Aditya Vashistha. 2025. " Ignorance is not Bliss": Designing Personalized Moderation to Address Ableist Hate on Social Media. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18.

[45] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems* 43, 2 (2025), 1–55.

[46] James I Hudson, Eva Hiripi, Harrison G Pope Jr, and Ronald C Kessler. 2007. The prevalence and correlates of eating disorders in the National Comorbidity Survey Replication. *Biological Psychiatry* 61, 3 (2007), 348–358.

[47] Louise Isham, Bao Sheng Loe, Alice Hicks, Natalie Wilson, Richard P Bentall, and Daniel Freeman. 2023. The Subjective Harm from Exceptional Experiences Questionnaire (SHEEQ). *Schizophrenia Bulletin* 49, 5 (2023), 1194–1204. https://www.psy.ox.ac.uk/research/oxford-cognitive-approaches-to-psychosis/resources/assessment-tools/the-subjective-harm-questionnaire-sheeq

[48] Farnaz Jahanbakhsh. 2023. *Empowering Users on Social Media for Better Content Credibility*. Massachusetts Institute of Technology.

[49] Farnaz Jahanbakhsh and David R Karger. 2024. A Browser Extension for in-place Signaling and Assessment of Misinformation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–21.

[50] Farnaz Jahanbakhsh, Yannis Katsis, Dakuo Wang, Lucian Popa, and Michael Muller. 2023. Exploring the Use of Personalized AI for Identifying Misinformation on Social Media. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 105, 27 pages.

https://doi.org/10.1145/3544548.3581219
[51] Farnaz Jahanbakhsh, Amy X. Zhang, and David R. Karger. 2022. Leveraging Structured Trusted-Peer Assessments to Combat Misinformation. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 524 (Nov. 2022), 40 pages. https://doi.org/10.1145/3555637
[52] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)* 25, 2 (2018), 1–33.
[53] Shagun Jhaver, Alice Qian Zhang, Quan Ze Chen, Nikhila Natarajan, Ruotong Wang, and Amy X. Zhang. 2023. Personalizing Content Moderation on Social Media: User Perspectives on Moderation Choices, Interface Design, and Labor. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 289 (Oct. 2023), 33 pages. https://doi.org/10.1145/3610080
[54] Shagun Jhaver and Amy X. Zhang. 2025. Do users want platform moderation or individual control? Examining the role of third-person effects and free speech support in shaping moderation preferences. *New Media & Society* 27, 5 (2025), 2930–2950. https://doi.org/10.1177/14614448231217993 arXiv:https://doi.org/10.1177/14614448231217993
[55] Daphne Keller. 2023. Carriage and Removal Requirements for Internet Platforms: What Taamneh Tells Us. *J. Free Speech L.* 4 (2023), 87.
[56] Daphne Keller. 2023. Lawful but Awful? Control over Legal Speech by Platforms, Governments, and Internet Users. *The University of Chicago Law Review* 90, Online (2023).
[57] Kenneth S Kendler, John Myers, Carol A Prescott, and Michael C Neale. 2001. The genetic epidemiology of irrational fears and phobias in men. *Archives of General Psychiatry* 58, 3 (2001), 257–265.
[58] Ronald C Kessler, Sergio Aguilar-Gaxiola, Jordi Alonso, et al. 2017. Trauma and PTSD in the WHO world mental health surveys. *European Journal of Psychotraumatology* 8, sup5 (2017), 1353383.
[59] Kate Klonick. 2017. The new governors: The people, rules, and processes governing online speech. *Harv. L. Rev.* 131 (2017), 1598.
[60] Akaash Kolluri, Renn Su, Farnaz Jahanbakhsh, Dora Zhao, Tiziano Piccardi, and Michael S Bernstein. 2025. Alexandria: A Library of Pluralistic Values for Realtime Re-Ranking of Social Media Feeds. *arXiv preprint arXiv:2505.10839* (2025).
[61] Deepak Kumar, Yousef Anees AbuHashem, and Zakir Durumeric. 2024. Watch your language: Investigating content moderation with large language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18. 865–878.
[62] Tina Kuo, Alicia Hernani, and Jens Grossklags. 2023. The unsung heroes of facebook groups moderation: A case study of moderation practices and tools. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–38.
[63] Cinoo Lee, Kristina Gligorić, Pratyusha Ria Kalluri, Maggie Harrington, Esin Durmus, Kiara L Sanchez, Nay San, Danny Tse, Xuan Zhao, MarYam G Hamedani, et al. 2024. People who share encounters with racism are silenced online by humans and machines, but a guideline-reframing intervention holds promise. *Proceedings of the National Academy of Sciences* 121, 38 (2024), e2322764121.
[64] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. 2022. MAT: Mask-Aware Transformer for Large Hole Image Inpainting. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE/CVF, 10748–10758. https://doi.org/10.1109/CVPR52688.2022.01049
[65] Yifang Li and Kelly Caine. 2022. Obfuscation remedies harms arising from content flagging of photos. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–25.
[66] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 26296–26306.
[67] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499* (2023).
[68] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. *arXiv preprint arXiv:2303.16634* (2023).
[69] Sara Markowitz and Michael Fanselow. 2020. Exposure therapy for posttraumatic stress disorder: factors of limited success and possible alternative treatment. *Brain sciences* 10, 3 (2020), 167.
[70] Mike Masnick. 2019. Protocols, Not Platforms: A Technological Approach to Free Speech. *Knight First Amendment Institute at Columbia University* (2019).
[71] Meta. 2022. Suicide prevention. https://about.fb.com/news/2018/09/inside-feed-suicide-prevention-and-ai/
[72] Jarosław M Michałowski, Dawid Droździel, Jacek Matuszewski, Wojtek Koziejowski, Katarzyna Jednoróg, and Artur Marchewka. 2017. The Set of Fear Inducing Pictures (SFIP): Development and validation in fearful and non-fearful individuals. *Behavior Research Methods* 49, 4 (2017), 1407–1419.
[73] Kyzyl Monteiro, Yuchen Wu, and Sauvik Das. 2024. Manipulate to Obfuscate: A Privacy-Focused Intelligent Image Manipulation Tool for End-Users. In *Adjunct Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–3.
[74] Sean A Munson and Paul Resnick. 2010. Presenting diverse political opinions: how and how much. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1457–1466.
[75] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* 20, 11 (2018), 4366–4383.
[76] National Institute of Mental Health. 2022. Specific Phobia. https://www.nimh.nih.gov/health/statistics/specific-phobia Accessed: June 30, 2025.
[77] Department of Justice. [n.d.]. Child Sexual Abuse Material. https://www.justice.gov/d9/2023-06/child_sexual_abuse_material_2.pdf.
[78] Lars-Göran Öst. 1989. One-session treatment for specific phobias. *Behaviour Research and Therapy* 27, 1 (1989), 1–7.
[79] Aviv Ovadya and Luke Thorburn. 2022. Bridging-based ranking. *Harvard Kennedy School Belfer Center for Science and International Affairs* (2022).
[80] Eli Pariser. 2011. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin.
[81] Jessica A. Pater, Oliver L. Haimson, Nazanin Andalibi, and Elizabeth D. Mynatt. 2016. "Hunger Hurts but Starving Works": Characterizing the Presentation of Eating Disorders Online. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 1185–1200. https://doi.org/10.1145/2818048.2820030
[82] Tiziano Piccardi, Martin Saveski, Chenyan Jia, Jeffrey Hancock, Jeanne L Tsai, and Michael S Bernstein. 2024. Reranking social media feeds: A practical guide for field experiments. *arXiv preprint arXiv:2406.19571* (2024).
[83] Graham Pullin and Alan Newell. 2007. Focussing on Extra-Ordinary Users. In *Universal Acess in Human Computer Interaction. Coping with Diversity*, Constantine Stephanidis (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 253–262.
[84] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems* 36 (2023), 53728–53741.
[85] Aravindh Raman, Sagar Joglekar, Emiliano De Cristofaro, Nishanth Sastry, and Gareth Tyson. 2019. Challenges in the decentralised web: The mastodon case. In *Proceedings of the internet measurement conference*. 217–229.
[86] Sarah T Roberts. 2019. *Behind the screen: Content moderation in the shadows of social media*. Yale University Press.
[87] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.
[88] Barbara Olasov Rothbaum and Ann C Schwartz. 2002. Exposure therapy for posttraumatic stress disorder. *American journal of psychotherapy* 56, 1 (2002), 59–75.
[89] Henrik Skaug Sætra and Jo Ese. 2023. Shinigami eyes and social media labeling as a technology for self-care. *Technology and sustainble development: The promise and pittfals of techno-solutionism* (2023), 53–69.
[90] Aaron Sankin. 2017. How activists of color lose battles against Facebook's moderator army. *Retrieved March* 29 (2017), 2023. https://revealnews.org/article/how-activists-of-color-lose-battles-against-facebooks-moderator-army/
[91] Sarita Schoenebeck, Amna Batool, Giang Do, Sylvia Darling, Gabriel Grill, Daricia Wilkinson, Mehtab Khan, Kentaro Toyama, and Louise Ashwell. 2023. Online harassment in majority contexts: Examining harms and remedies across countries. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–16.
[92] Sarita Schoenebeck and Lindsay Blackwell. 2020. Reimagining social media governance: Harm, accountability, and repair. *Yale JL & Tech.* 23 (2020), 113.
[93] Shinigami Eyes. 2024. Shinigami Eyes: An extension that highlights trans-friendly and anti-trans social network pages. https://shinigami-eyes.github.io/. Accessed: June 30, 2025.
[94] Guijin Son, Sangwon Baek, Sangdae Nam, Ilgyun Jeong, and Seungone Kim. 2024. Multi-task inference: Can large language models follow multiple instructions at once? *arXiv preprint arXiv:2402.11597* (2024).
[95] Jonathan Stray. 2021. Designing recommender systems to depolarize. *arXiv preprint arXiv:2107.04953* (2021).
[96] Petter Törnberg. 2022. How digital media drive affective polarization through partisan sorting. *Proceedings of the National Academy of Sciences* 119, 42 (2022), e2207159119.
[97] Sandra Villafuerte and Margit Burmeister. 2003. Untangling genetic networks of panic, phobia, fear and anxiety. *Genome Biology* 4, 8 (2003), 224.
[98] Ru Wang, Kexin Zhang, Yuqing Wang, Keri Brown, and Yuhang Zhao. 2025. "It was Mentally Painful to Try and Stop": Design Opportunities for Just-in-Time Interventions for People with Obsessive-Compulsive Disorder in the Real World. In *Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–18.

[99] Yufei Wang, Zhanyi Sun, Jesse Zhang, Zhou Xian, Erdem Biyik, David Held, and Zackory Erickson. 2024. RL-VLM-F: reinforcement learning from vision language foundation model feedback. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 2112, 18 pages.

[100] Sijia Xiao, Coye Cheshire, and Niloufar Salehi. 2022. Sensemaking, support, safety, retribution, transformation: A restorative justice approach to understanding adolescents' needs for addressing online harm. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–15.

[101] Sijia Xiao, Shagun Jhaver, and Niloufar Salehi. 2023. Addressing interpersonal harm in online gaming communities: The opportunities and challenges for a restorative justice approach. *ACM Transactions on Computer-Human Interaction* 30, 6 (2023), 1–36.

[102] Mayu Yoshikawa, Zui Narita, and Yoshiharu Kim. 2024. Digital health–based exposure therapies for patients with posttraumatic stress disorder: A systematic review of randomized controlled trials. *Journal of Traumatic Stress* 37, 6 (2024), 814–824.

[103] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. 2018. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5505–5514.

[104] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, et al. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685* (2023).

[105] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592* (2023).

A PERFORMANCE OPTIMIZATIONS FOR REAL-TIME CONTENT PROCESSING

A.1 Overview of Computational Load

Processing personalized content modifications in real-time presents a significant computational challenge. The system must analyze large batches of content and make multiple, time-consuming VLM calls, all within the few seconds a user spends on a single screen of content. To quantify this challenge, we first define the key variables.

Let N be the total number of posts in a batch (typically 25 for Reddit), with N_t being the number of posts containing text and N_p the number containing images. From these, let M_t be the number of posts with text that match a user's filter and M_p be the number of posts with images that match. Finally, let K be the number of intervention candidates our system generates for each matching image (typically $K = 3$).

The following table details the different processing sequences, their API call latency, and their dependencies.

As the table shows, the total number of VLM calls (T_{total}) for a single batch can be expressed as the sum of calls for the text and image pipelines:

$$\text{Text Pipeline:} \quad T_{\text{text}} = N + 2 \times M_t \tag{1}$$

$$\text{Image Pipeline:} \quad T_{\text{image}} = N_p + 2 \times KM_p \tag{2}$$

$$\text{Total:} \quad T_{\text{total}} = N + 2 \times M_t + N_p + 2 \times KM_p \tag{3}$$

To illustrate the computational load, consider a user with a sensitivity to violent imagery browsing Reddit during a major international conflict. A general-interest subreddit like `r/worldnews` might have nearly every post contain a matching image. In this case, $M_t \approx N$, $N_p \approx N$, and $M_p \approx N_p$. The total API calls would approach $10N$, or roughly 250 calls for a single 25-post batch, all of which must be processed in seconds to ensure usability.

The table's first two groups, the *Text Processing Sequence* and the *Image Filter Analysis*, represent the critical path that must complete before returning an initial response to the client for the full batch of posts. The final *Image Transformation Sequence* can then process asynchronously in the background generating time-consuming image transformation and scoring. Without any optimizations, this pipeline would require several minutes to complete; our approach reduces the critical path latency to 5-15 seconds for the batch.

Table 3: Computational characteristics of the processing pipeline. The pipeline is divided into sequences that must complete on the critical path before a response is sent to the client, and a final sequence that can run asynchronously in the background.

Operation	API Calls	Latency	Blocking
Text Processing Sequence (Critical Path):			
Text filter matching	N	0.5-1s	Yes
Text intervention selection	M_t	1-2s	Yes
Text intervention application	M_t	1-2s	Yes
Image Filter Analysis (Critical Path, parallel with text):			
Image filter matching	N_p	2-5s	Yes
Image Transformation Sequence (Asynchronous Background Task):			
Image intervention generation	$K \times M_p$	2-10s	No
Image intervention scoring	$K \times M_p$	2-5s	No

A.2 Asynchronous Processing with Multi-Batch Responses

Our primary optimization decouples text and image processing timelines. The system returns text interventions immediately while deferring computationally intensive image transformations to background workers.

When the server receives a batch of posts, it initiates parallel processing through asyncio tasks. The critical path includes text filter matching, text intervention application, and image filter analysis, which complete within 5-10 seconds total. Text modifications return immediately with CSS-based intervention instructions, while images that require transformation receive placeholder tokens. The server then initiates Celery worker processes for the computationally intensive image transformation pipeline. The client then polls for completed image transformations using these tokens.

Our polling strategy reflects **two** key insights.

First, analyzing the DOM structure, we found that Reddit employs lazy loading for images, fetching actual image content only when necessary based on viewport position and network conditions. Our deferred image processing aligns naturally with this behavior.

Second, the probability that intervention selection has completed for a content increases over time as more candidates are evaluated. Therefore, the client implements an inverse exponential backoff polling strategy, starting with longer intervals that gradually decrease. To prevent server overload from synchronized polling, each content item adds randomized jitter to its polling delay. Posts appearing earlier in the feed receive higher polling priority since users are more likely to encounter them first.

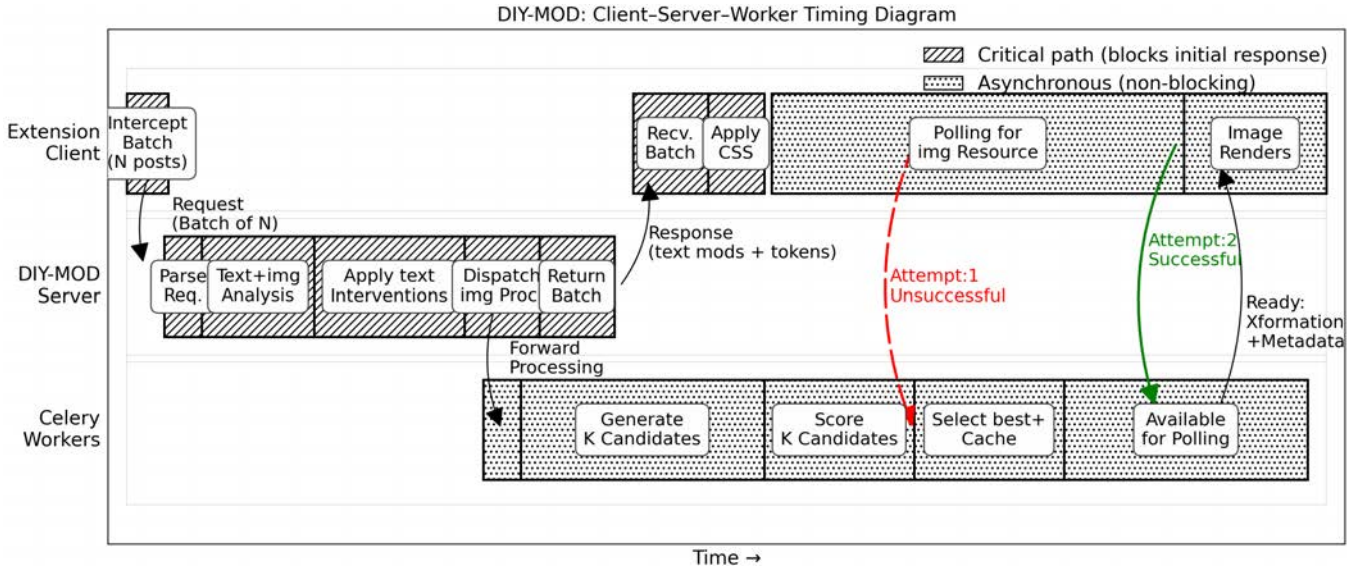


Figure 5: DIY-MOD processing pipeline showing synchronous and asynchronous execution paths. The critical path (diagonal hatching) completes in 5-15 seconds for a batch of posts, returning text modifications and polling tokens. Image transformations process asynchronously in background workers (dotted pattern). The client polls for completed transformations—missing on attempts until the content transformation is ready. Previously processed content can return immediately from cache on first poll. Note that, the critical path delay is for a batch of usually 25 posts in reddit. And, this delay only happens when the user first loads/refreshes the feed in browser. For subsequent batches of posts within the browsing session, predictive prefetching (Appendix A.3) minimizes the delay under reasonable scrolling behavior.

This architecture ensures that users can immediately interact with text content while image transformations process in the background. It is worth noting that, *we do not touch platform advertisements throughout this process*, preserving the platform's native monetization mechanisms.

A.3 Predictive Prefetching

Predictive prefetching leverages the natural pace of browsing. We observed that users typically spend 30-90 seconds engaging with the posts in a single batch, which creates a crucial viewing window to process the next batch before a user scrolls to it. To exploit this window, our server manipulates the platform's pagination cursor in its response to the client. As a result, when a user has viewed approximately 40% of the current batch, the platform's front-end is prompted to request the next batch of posts, essentially far earlier than it normally would.

`DIY-MOD` as usual intercepts this request and sends it to our server, ensuring the processing completes long before the user ever reaches the new content. As a result, under reasonable scrolling behavior, users experience no perceptible delay after the initial page load.

A.4 Content-Based Caching

The system implements a Redis based cache to eliminate redundant processing when multiple users encounter the same content or when users reload pages. The cache key combines three components:

$$\texttt{cache\_key} = \texttt{hash(content)} \oplus \texttt{hash(filter\_params)} \oplus \texttt{hash(sensitivity\_leve}$$

This formulation ensures that identical content processed with identical filter parameters yields the same cached result, while different filter configurations or sensitivity levels generate distinct cache entries. The cache operates across all users without compromising privacy since only the anonymous analysis result is stored with no link to user identity.

When viral/popular content spreads across Reddit, the first user to encounter it triggers the full processing pipeline. Subsequent users with similar filters receive instantly cached interventions. This particularly benefits popular posts that appear on `r/all`, `r/popular` or trending subreddits where many users may have similar content sensitivities.

A.5 Edit Failures, False Obfuscations and Mitigation Strategies

A.5.1 Generation failures and safeguards. During deployment, we encountered two types of failures in our generation pipeline. First, specific filters occasionally triggered broader VLM categorization: for example, a filter for "Middle East Conflict" by one participant matched general conflict imagery or rubble. Second, Gemini's safety guardrails sometimes refused to generate lawful sensitive content such as medical imagery. We investigated this further but

observed non-reproducible failures where identical prompts and inputs would succeed on some attempts but fail on others.

We designed our mitigation strategy to treat both as fail-safe defaults. When generation fails, we first retry, which resolve in most cases. Generating $K = 3$ parallel candidates also increases the likelihood that at least one succeeds and scored higher. While we could switch to alternative VLMs upon multiple failure, we prioritized real-time feed rendering over exhaustive fallback strategies. Even when all candidates fail to generate, we leave the content unchanged but still apply the transparency indicator (Section 4.2, Figure 1(d)) on top of the original post. This indicator informs users that the content matched one of their filters and was meant to be transformed. Participants reported using this indicator as a personalized warning label—a cue to look away from posts that might be harmful to them. In other words, we err on the side of flagging a post as potentially triggering rather than exposing users to it.

A.5.2 Audit of edit failures and false positives. To quantify edit failures and false obfuscations, we manually audited posts that our extension marked as modified via the transparency indicator. For each participant, we randomly sampled 5 text posts and 5 image posts, yielding 75 text items and 75 image items. For text, we confirmed via screen recordings that all 75 posts were successfully transformed. For images, we visually matched each transformed image to the original Reddit post and the filter description of the viewer. We found that 66/75 ($\approx$ 88%) images were visually transformed, while 8 ($\approx$ 12%) were either left unchanged or edited in the wrong region. Among the 66 transformed images, 62 ($\approx$ 93%) were correctly transformed according to the viewer's filter description, implying a false-positive transformation rate of roughly 7% of transformed images.

B FORMATIVE STUDY

B.1 Data Analysis and Protection

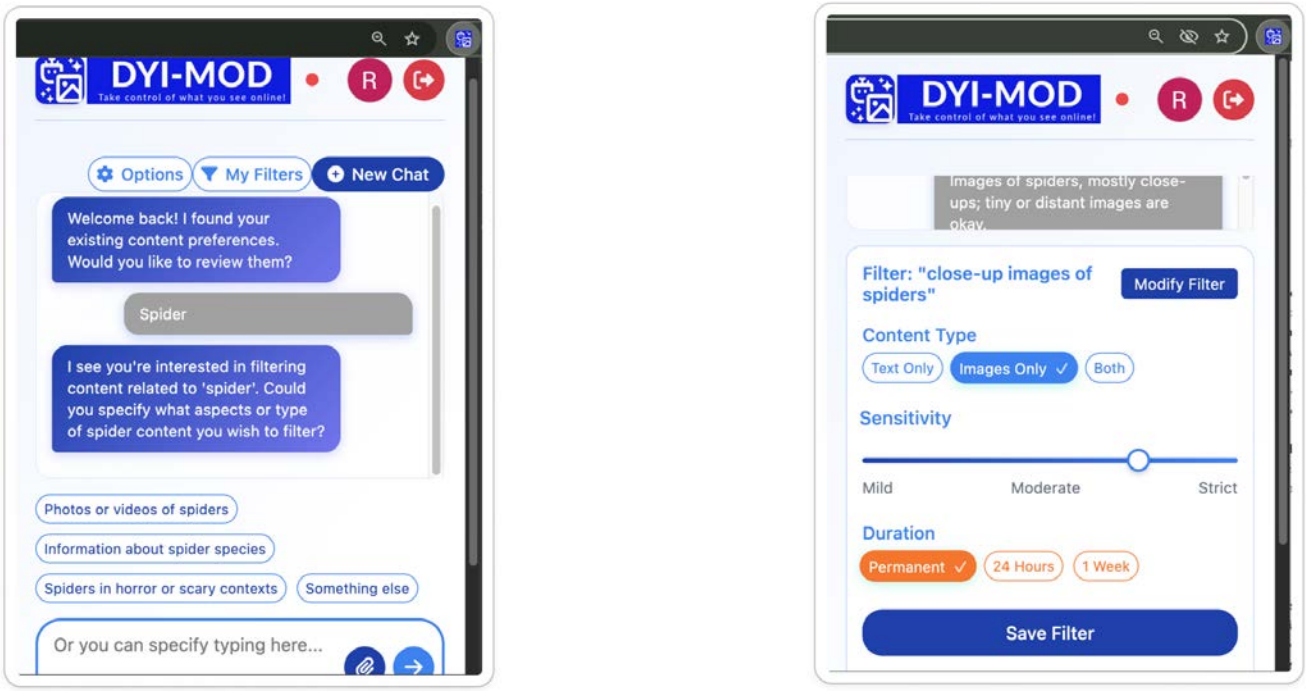
We transcribed the interviews verbatim and then deleted the original recordings to avoid retaining identifiable data. Using reflexive thematic analysis [12], we inductively identified the core themes presented in Section 3.2 through 3.7. We pseudonymized transcripts with coded identifiers (e.g., P_01, P_12). We stored names and email addresses for recruitment and compensation in a separate encrypted file, deleted them after payment, and never linked them to transcripts. All study data remained on encrypted, university-approved devices and cloud services accessible only to the core research team.

C MORE DETAILS ABOUT SYSTEMS

C.1 DIYMOD

Details in Figure 6a and Figure 6b.

C.2 Dual Feed



(a) Chat interface

(b) Config interface

Figure 6: DIY-MOD's filter creation and configuration interface
(6a) Chat interface for filter creation. DIY-MOD engages in conversational grounding to help users specify their content sensitivities. When a user enters a broad term, the system can prompt for clarification about more specific aspects. Users control specification depth through suggested options or free text input. This chat interface opens as a popup when user clicks on the extension icon.

(6b) Filter configuration interface. After establishing the filter description, users configure three parameters: (1) *Content Type* specifies whether the filter applies to text, images, or both; (2) *Sensitivity Level* indicates the user's distress intensity; and (3) *Duration* sets filter expiration. For brevity, in popup interface we only showed three time presets. Users can modify the filter description anytime using the "Modify Filter" button or add metadata through the Options page for further customization.