SafeScroll: A Machine Learning-Enhanced Browser Extension for Real-Time Trauma Trigger Detection and Content Filtering

Ananya Patel[1]

Received May 26, 2026
Accepted August 17, 2026
Electronic access September 15, 2026

Unexpected online images can cause flashbacks and severe distress for individuals with Post-Traumatic Stress Disorder (PTSD) or related trauma conditions. Existing moderation tools filter broad categories like nudity or violence, overlooking personalized triggers. Since trauma is unique to each individual, trigger cues are not necessarily graphic; they can be neutral objects, such as baseball bats or dogs. While these images appear normal to most users, they can distress trauma survivors. This research tests whether a browser extension can detect user-specified triggers in real time. SafeScroll addresses this with an extension that uses machine learning models to analyze and filter images based on user-defined trigger words. SafeScroll operates fully locally to protect user privacy. The system uses two AI models: Faster Region-based Convolutional Neural Networks (Faster R-CNN) for object detection across the 80 classes in the Common Objects in Context (COCO) dataset, and Vision Transformer for Open-World Localization (OWL-ViT v2) for zero-shot detection. SafeScroll was evaluated on 527 web images (150 positive, 377 negative), achieving 94.7% precision (95% CI: 92.1%–97.2%) and 96.0% recall (95% CI: 93.8%–98.1%), and a 2.1% false-positive rate. Optimizations reduced average processing time from 4100 ms to 670 ms. Feedback from trauma survivors showed reduced trigger exposure, satisfaction with detection accuracy, and high usability ratings. Therefore, SafeScroll may help people with PTSD, trauma histories, and flashbacks participate more safely in digital spaces.

Keywords: PTSD, trauma, machine-learning, computer-vision

Introduction

Trauma triggers bring up memories of past harm, forcing survivors to relive overwhelming emotions and experience stress long after the original event. Since users cannot easily control what content they see in a scroll feed, visual triggers often appear unexpectedly in digital environments. Exposure to these triggers can disrupt a trauma survivor's day, potentially causing distress and panic attacks[1]. Being re-traumatized by these visual triggers also reinforces cycles of anxiety or avoidance, which is likely to hinder the user's recovery.

Exposure to trauma triggers online affects millions of people in the U.S. who depend on digital spaces for work, education, and community. As platforms increasingly rely on algorithmic feeds, user control decreases, and unexpected triggers become more harmful. A high risk of trigger exposure leaves users with two choices: stay on these platforms and face the harm of trigger exposure, or greater social isolation if they choose not to participate in digital spaces. Developing adaptive filtering tools is therefore essential not only for individual mental health but also for fostering safer, more inclusive digital personal spaces.

In Figure 1, a seamlessly integrated extension is shown that continuously monitors web feeds and removes or overlays any unwanted content, protecting users from trigger exposure.

To have the best user experience, the content filter will not disrupt the user's normal online workflow (e.g., a web browser extension that runs in the background). To protect users' best interests, the tool would be a real-time monitoring application that continuously scans new content feeds on the screen, detecting and filtering triggering images with a 670 ms average processing time, rapidly masking content upon detection.

This study addresses the following research question: How can a client-side, machine learning-based system be designed to accurately and efficiently detect and filter user-defined trauma-triggering visual content in real time during web browsing, while preserving user privacy and minimizing disruption to the browsing experience? It is hypothesized that combining Faster R-CNN with OWL-ViT v2 will achieve at least 90% precision and 90% recall on user-specified trigger categories, with processing times under one second per image on consumer hardware without GPU acceleration.

Literature Review

Psychological Foundations

Individuals with PTSD are often "easily triggered by reminder cues"[2]. Exposure to such cues may result in symptoms such

[1] *Pine Crest School, Florida, USA*

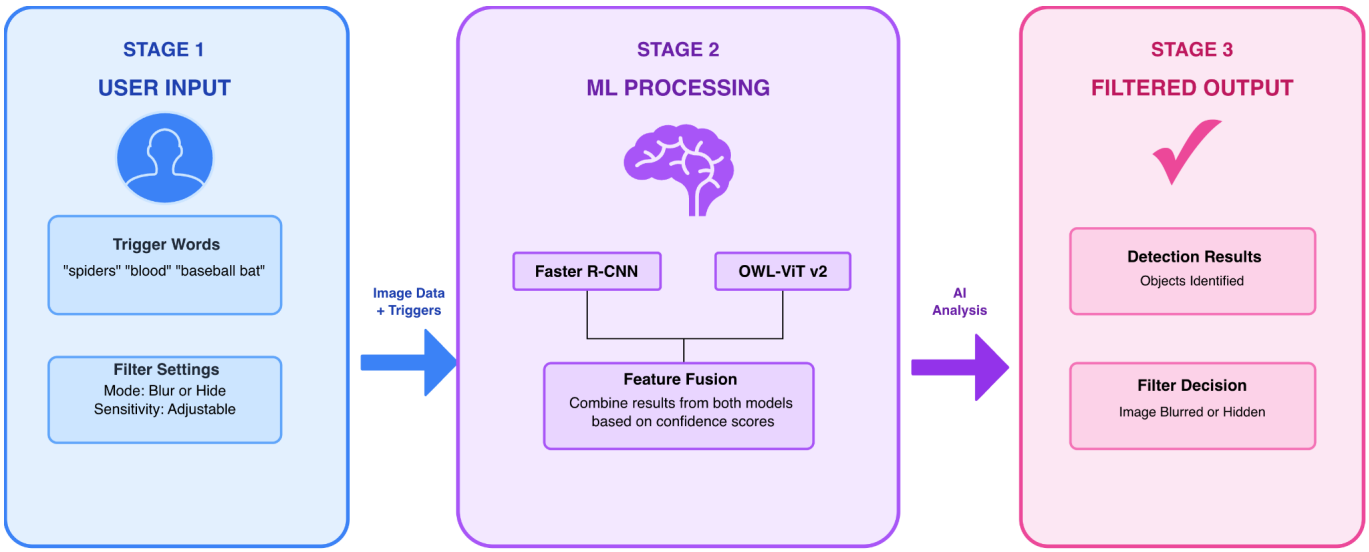

Figure. 1 A block diagram of the proposed system, showcasing different modules working together to trace and track unwanted content.

as flashbacks and panic attacks[3]. Though users may benefit from exposure to triggers in therapy, trigger exposure outside of a controlled environment does not support recovery[3]. Accidental exposure to these trauma cues may worsen symptoms, as individuals with PTSD are "vulnerable to the adverse effects of losing control"[4]. On social media platforms, visual triggers can lead to vicarious trauma[5] and worsen PTSD-like symptoms. In some cases, exposure to online triggers results in direct trauma[6].

Role of AI and ML in Content Moderation

Deep convolutional neural networks (CNNs) are artificial neural networks designed to analyze visual images. Starting with models such as AlexNet in 2012, CNNs established that learned hierarchical visual features from data can significantly increase image recognition accuracy[7,8]. Modern object detection methods generally fall into two categories. Two-stage detectors, such as the Faster Region-based Convolutional Neural Network (Faster R-CNN), generally achieve higher accuracy. In contrast, single-stage detectors, such as YOLO (You Only Look Once), offer higher speed[9].

Recent advances in computer vision have greatly improved the performance of detection models. Skip connections helped deep networks achieve much higher accuracy[10]. Feature Pyramid Networks (FPN) combine multi-scale information to improve object detection throughout a range of sizes[11]. Vision Transformers (ViT) use self-attention mechanisms applied to image patches, and when pre-trained on large datasets, can match or surpass CNNs in performance[12].

Models such as CLIP enable zero-shot classification, meaning that images can be labeled using only natural-language prompts[13]. Open-vocabulary detectors, such as OWL-ViT, apply zero-shot classification to object detection. This enables systems to recognize objects described in text even if not included in training data[14,15].

Safe-AI and Adaptive Filtering

AI-based content safety systems must balance the trade-off between false positives and false negatives. Over-filtering can reduce exposure to harmful content but may also block wanted content. On the other hand, under-filtering increases the risk of user harm[16]. To be effective, moderation systems must rely on clearly defined policies and ongoing evaluation[16].

Google SafeSearch uses machine learning models to detect and block explicit content[17], Reddit's AutoModerator uses customizable rule-based keyword or regular expression filters developed by human moderators[18], and Meta uses AI systems to flag content against its safety policy, escalating some cases to human reviewers[19].

These systems may provide widespread coverage, but still have limitations. Filters such as SafeSearch apply broad rules[20], recognizing only general categories of harmful content, such as self-harm or nudity. Social media platforms have no options to filter personal triggers, and for this reason, their existing moderation tools are not trauma-informed[21].

The literature highlights many unresolved gaps, such as the lack of individualized, trauma-informed filtering and the inability of current systems to detect user-specific triggers. Psychologically, exposure to trauma cues can reinforce PTSD symptoms and hinder recovery[3,4]. The closest related system

is DIY-MOD[22], a browser extension enabling user-specified content filtering via large vision-language models. Unlike SafeScroll, DIY-MOD applies semantic content transformations (e.g., inpainting to remove triggering objects) rather than image-level blur filtering, and depends on cloud-based vision APIs rather than local processing, introducing privacy trade-offs. No prior system combines local object detection with zero-shot recognition specifically for trauma-trigger filtering in a privacy-preserving, client-side architecture. Vision Transformers, CNNs, and open-vocabulary detectors provide the necessary components for such user-specified, private filtering systems.

A trauma-informed filter could be applied to many scenarios, from obscuring images of fireworks for veterans or hiding an abuser's tattoo from a domestic violence survivor, which can help make digital spaces safer for trauma survivors of any kind.

Methods

Microsoft's COCO Dataset (Common Objects in Context)[23] is the primary dataset used for SafeScroll. COCO contains 330,000 images containing 1.5 million labeled object instances across 80 object categories (such as person, bicycle, and dog). COCO was used to pre-train the Faster R-CNN object-detection model.

Additionally, the ImageNet Dataset[24] contains over 14 million images classified into 1000 distinct categories and is used to train the ResNet-50 backbone, which serves as a feature extractor for the Faster R-CNN model. The ResNet-50 backbone detects important features, such as edges and textures, which are then passed to the detection head.

For the OWL-ViT v2's training data, Visual Genome, Conceptual Captions, and COCO are combined to build the foundation the model needs to achieve zero-shot object detection from text descriptions[14]. All processing occurs locally to preserve user privacy.

The cleaning phase filters all images less than 50×50 pixels, as these are often decorative elements and often produce false positives. Then, all images are converted to RGB format to ensure consistency. Images are then resized to a maximum of 1024 pixels for feature optimization, reducing payload size by approximately 95%. Then, images are encoded into WebP format (with JPEG fallback).

For neural network input, images are converted to PyTorch tensors with pixel values normalized to the [0, 1] range. Normalization follows ImageNet standards for Faster R-CNN, using mean values [0.485, 0.456, 0.406] and standard deviation values [0.229, 0.224, 0.225]. The Owlv2Processor class is used for OWL-ViT v2's model-specific normalization.

Because Faster R-CNN and OWL-ViT v2 are pre-trained models used for inference rather than fine-tuned on new data, no training split was required. However, a held-out evaluation set of 527 images was constructed: 150 positive images containing at least one trigger across 10 categories (person, dog, car, kite, knife, bicycle, fireworks, tattoo, clown, and syringe) and 377 negative images containing no trigger categories. Images were collected from Google Images (n=312), Unsplash (n=128), custom photography (n=62), and web screenshots (n=25) between September 2025 and January 2026. Ground-truth labels were assigned through dual independent annotation achieving Cohen's $\kappa = 0.89$ inter-annotator agreement. All 58 disagreements (11.0% of images) were resolved by a third-party adjudicator, and near-duplicate images were removed using perceptual hashing prior to annotation. A LRU (Least Recently Used) cache with a 100-entry capacity and a 5-minute TTL is used in-session, and IndexedDB storage with 30-day retention is used for efficiency across sessions.

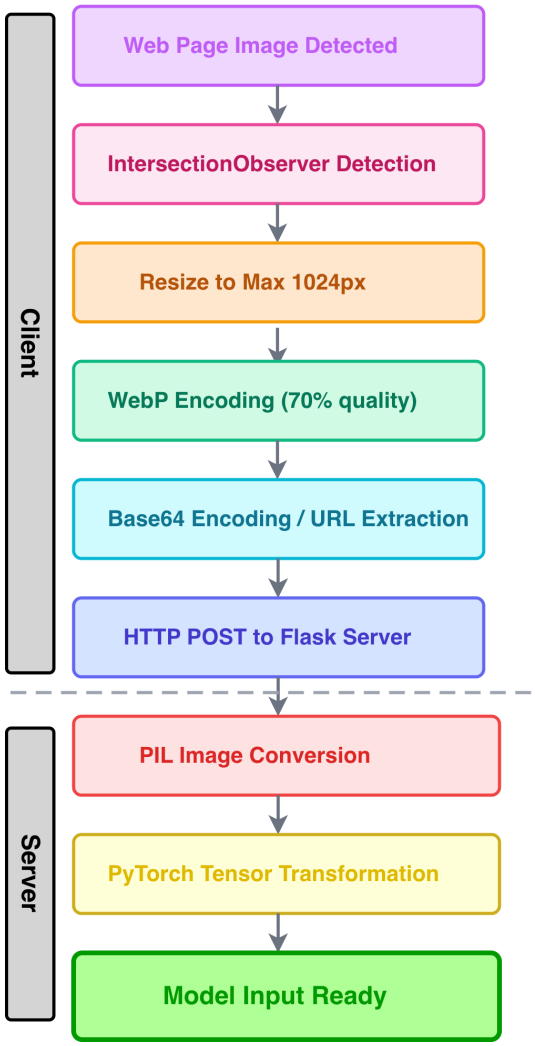

Figure. 2 Preprocessing flow from browser client to server backend.

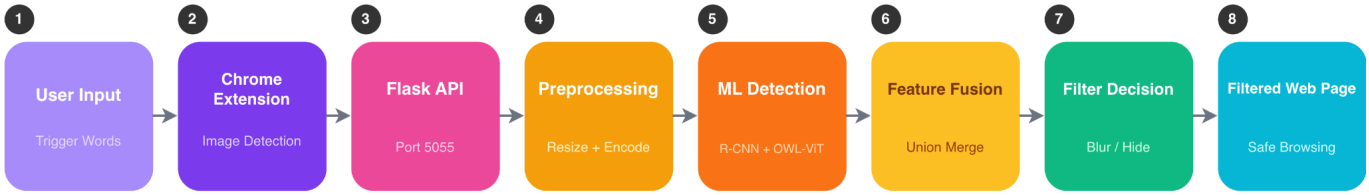

Figure. 3 End-to-end pipeline of SafeScroll.

This project uses Faster R-CNN with a ResNet-50 backbone as the primary detector model and OWL-ViT v2 (Open-Vocabulary Vision Transformer) for zero-shot detection for triggers not present in typical datasets, such as "tarantula," "syringe," or "clown"[14]. OWL-ViT v2's Vision Transformer architecture with CLIP-based text encoding enables semantic segmentation via vision-language alignment[12]. Figure 3 explains the proposed architecture.

After the models are run, Faster R-CNN and OWL-ViT v2 results are combined via a union merge. To prioritize user safety, an image is flagged as trigger detected if either model detects the given word. Additionally, adjustable confidence thresholds (standard as 0.2 for OWL-ViT v2 and 0.6 for Faster R-CNN) allow personalization based on a user's sensitivity preferences.

The system architecture requires a Python server for machine learning inference and a Chrome browser extension for user interaction. The programming languages used were Python 3.9+ for server-side machine learning inference through the Flask backend, JavaScript ES6+ for the Chrome extension using Manifest V3 APIs, and JSON for configuration and API communication. CSS3 is used for visual filtering effects (blur and hide), and HTML5 for the extension's user interface pages. See Appendix A for the necessary software packages. The browser extension portion uses chrome.storage.sync for cross-device settings persistence, IntersectionObserver for lazy loading (loading images only when visible), MutationObserver for monitoring dynamically added content, IndexedDB for 30-day persistent result caching, and FinalizationRegistry for automatic event listener cleanup. Minimum hardware requirements are an Apple M2 or Intel x86_64 processor (evaluation was conducted on Apple M2 without GPU), 8 GB RAM (16 GB recommended), and 5 GB storage for model weights. An optional CUDA-compatible GPU provides 3–5× acceleration for inference.

SafeScroll was also evaluated by five individuals who reported distress from specific visual stimuli (e.g., PTSD triggers or phobia-related cues) in a preliminary qualitative usability survey. Each participant tested SafeScroll during real-time browsing and completed a survey with Likert-scale ratings (1–5) and short open-ended questions assessing filter performance and usability. All participation was voluntary and anonymized; no personal data or trigger keywords were recorded by the research team, as participants configured their trigger words privately.

Ethics and Human Subjects Protection

This study was approved by the CCIR Ethics Committee (approval granted March 9, 2026). Potential participants first received a detailed three-page information sheet at least 48 hours before their scheduled session, describing study procedures, risks (including the possibility of brief image exposure before SafeScroll's filter activates), benefits, and their right to withdraw without penalty. Written informed consent was obtained in person before each session, with the researcher co-signing as witness; participants retained a copy of their signed consent form. Participants were pre-screened via phone interview to exclude individuals reporting active suicidal ideation or severe acute psychiatric symptoms. Participants were explicitly informed they could stop the session at any time without providing a reason and would still receive full compensation. A researcher with training in trauma-informed practices and distress recognition was present throughout all ses-

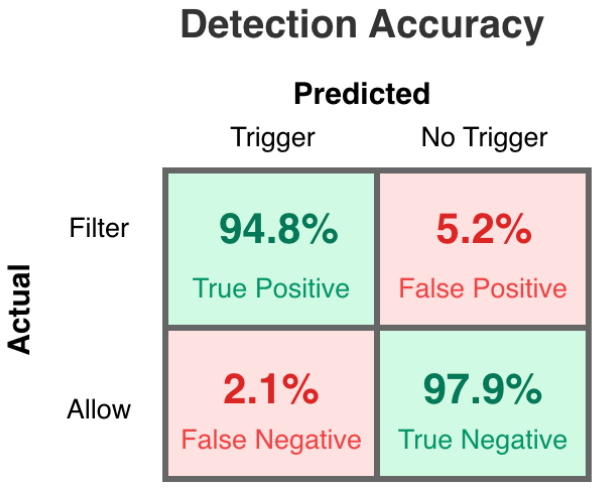

Figure. 4 Confusion matrix showing detection accuracy across 527 web page images (150 positive, 377 negative).

sions. Two participants reported mild transient distress during testing (classified as mild adverse events); both were offered the option to end the process immediately, and both chose to continue after a brief pause. No severe adverse events occurred. All participants received a debriefing sheet and a list of support resources (including the 988 Suicide and Crisis Lifeline and the PTSD Crisis Line) at the conclusion of the session, and were followed up by the research team within 24 hours. Participant data were de-identified, assigned alphanumeric codes, and stored on encrypted servers accessible only to the researcher. Research data will be retained for five years and then permanently destroyed.

Results

Across 527 evaluation images, the system produced 144 true positives, 369 true negatives, 8 false positives, and 6 false negatives. Figure 4 displays the confusion matrix. These counts yield a precision of 94.7% (95% CI: 92.1%–97.2%), recall of 96.0% (95% CI: 93.8%–98.1%), specificity of 97.9%, F1-score of 0.953, and accuracy of 97.3%; confidence intervals were computed via bootstrap resampling (1,000 iterations, seed=42). Objects that resembled traumatic cues, such as red paint mistaken for blood, were typical false positives. Per-category F1 ranged from perfect detection (1.000 for bicycle, clown, and syringe) to 0.880 for tattoo, which showed higher visual variability in style, size, and occlusion. Of the 6 false negatives, 2 occurred in higher-severity categories (knife and fireworks), underscoring the importance of the union fusion strategy that prioritizes recall to minimize missed detections.

Another part of the evaluation compared the baseline version of SafeScroll to the optimized version. The baseline version used FP32 precision and sequential inference, producing accurate results but suffering from high latency. On the other hand, the optimized version uses INT8 quantization, parallel processing, batch processing, and several other improvements, creating a dramatic difference. To highlight the results of both versions, Figure 5 is a comparison of processing times during each stage of the system.

Optimizations reduced preprocessing time by 32.4% and for both Faster R-CNN and OWL-ViT, inference time by 66.7%. Overall, processing time decreased by 83.6%, going from 4100 ms to 670 ms. The relative contribution of each optimization strategy was also evaluated, as shown in Figure 6.

The most effective optimization strategy was batch processing, as it increased efficiency by 1735%. However, batch processing was not the only contributor to reduced processing times: the system's high speed resulted from a combination of many improvements, such as parallel model inference and WebP encoding.

To justify the chosen confidence thresholds (0.6 for Faster R-CNN; 0.2 for OWL-ViT v2), a sensitivity analysis was conducted across 42 threshold configurations (six COCO values × seven OWL-ViT values). The selected operating point achieved the highest F1-score of 0.953, within a stable per-

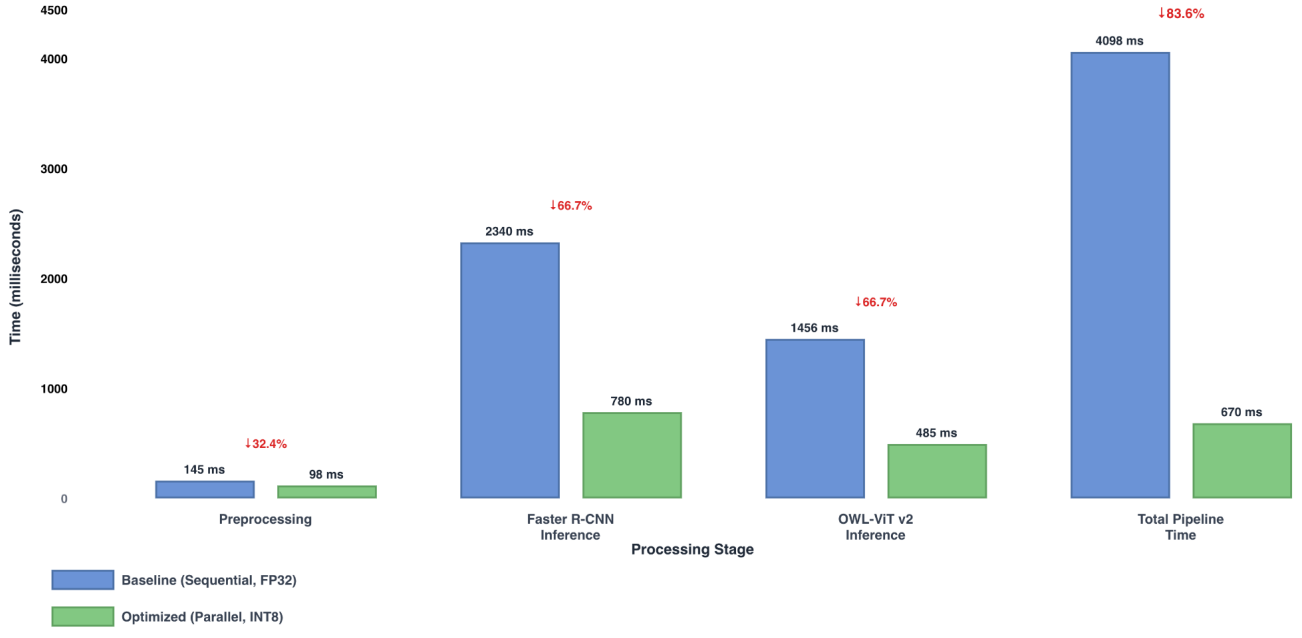

Figure. 5 Processing times through stages of SafeScroll.

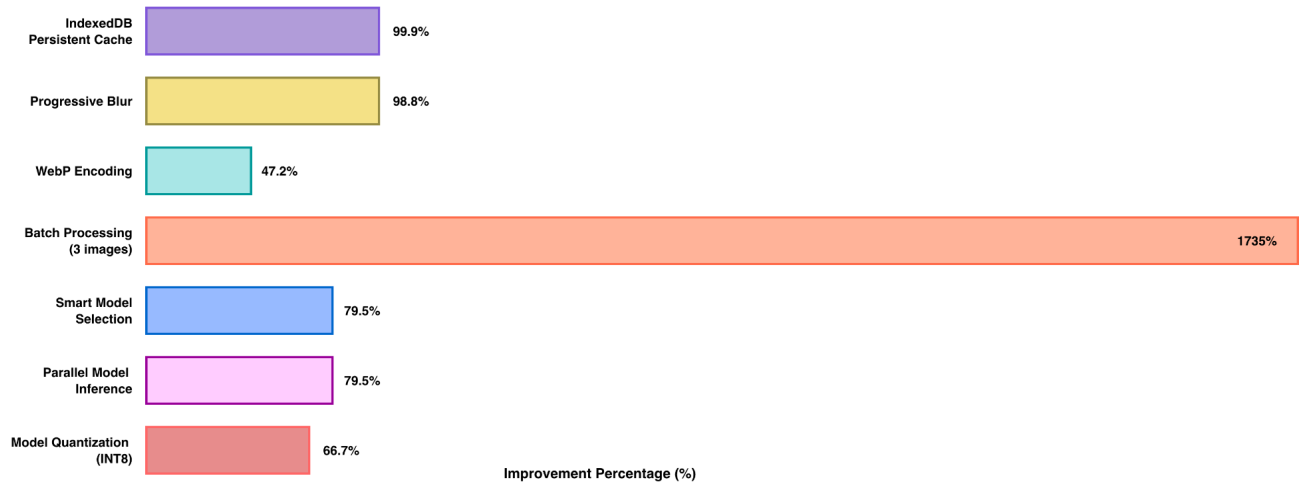

Figure. 6 Improvement percentages of optimization strategies.

formance plateau (F1: 0.945–0.953 across COCO 0.5–0.7 and OWL-ViT 0.18–0.25), confirming the selection is robust rather than a fragile local optimum. Lower thresholds (e.g., COCO 0.3, OWL-ViT 0.15) increased recall to 98.7% but reduced precision to 87.2%; higher thresholds (e.g., COCO 0.8, OWL-ViT 0.30) achieved 98.1% precision but reduced recall to 88.0%.

An ablation study compared three fusion strategies for combining Faster R-CNN and OWL-ViT v2 detections: union (flag if either model detects), intersection (flag only if both models detect), and weighted score combination. The union strategy achieved the highest recall (96.0%, 6 false negatives) with acceptable precision (94.7%, 8 false positives). The intersection strategy achieved higher precision (98.4%) but substantially reduced recall to 84.0%, yielding 24 false negatives. A weighted equal combination produced precision of 97.2% and recall of 92.0%. Given that missed triggers carry substantially higher user harm than unnecessary blurring of safe images, the union strategy was selected as the appropriate operating point for a safety-critical application.

The optimized version was also given to the qualitative feedback survey participants. All five of the survey participants reported experiencing moderate to severe emotional effects when exposed to visual triggers, with symptoms including anxiety spikes, dissociation, flashbacks, and sleep disruption. All participants indicated that SafeScroll reduced exposure to triggers during browsing as well as that they could easily enter and customize their triggers. Four participants rated detection accuracy as 5/5, while one rated it 4/5. Given the small sample size (N=5), absence of a control condition, and lack of blinding, these results should be interpreted as preliminary usability observations only and do not support claims about clinical efficacy or reduction of trauma symptoms.

Discussion

This research developed SafeScroll, a browser extension that uses machine learning to detect and filter user-defined trauma triggers in web images. SafeScroll provides real-time protection for trauma survivors, people with phobias, and users who want to better manage their online visual experience, filling a vital need in mental health technology.

Unlike existing platforms such as Google's SafeSearch or Reddit's AutoModerator, SafeScroll uses a combination of deep learning models. While Faster R-CNN detects standard objects across 80 COCO categories, OWL-ViT v2 uses zero-shot detection for more individualized trigger categories. This allows for personalized filtering without pre-labeled datasets for each trigger type. Various optimizations, such as INT8 quantization, parallel model inference, and persistent caching, improved the system from a 4.1-second prototype to a tool that processes images in under a second. During testing, the optimized version of SafeScroll achieved a precision of 94.7% (95% CI: 92.1%–97.2%) and recall of 96.0% (95% CI: 93.8%–98.1%), and a false-positive rate of 2.1%.

Despite promising results, there are some limitations to SafeScroll. Stylized or abstract images, such as illustrations or low-detail renderings or icons, occasionally went undetected because they were significantly different from the datasets. Additionally, the backend is run on a local Python server, which limits deployment to computers with enough RAM.

Future development aims to extend trigger detection to video content through frame-by-frame analysis with temporal sampling. However, real-time video processing poses some challenges, as it requires GPU acceleration and optimized streaming architectures to maintain acceptable performance.

Additionally, adapting SafeScroll for applications such as

Instagram and Snapchat would make it accessible to a wider range of users, though mobile apps present challenges such as restricted background processing and limited computational resources.

Overall, SafeScroll demonstrates that locally processed, user-specified trigger filtering is technically feasible at sub-second speeds on consumer hardware (Apple M2, no GPU). Because all processing occurs on-device via a local Flask server with no external data transmission, user trigger keywords and browsing activity remain private. Preliminary results suggest promise for reducing trigger exposure during web browsing, though larger controlled studies are needed before drawing conclusions about clinical effectiveness.

References

1 American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 5th ed., text rev. American Psychiatric Association Publishing (2022).

2 L. Petersdotter, L. Miller, A. Hammar. Trauma-analogue symptom variability predicted by inhibitory control and peritraumatic heart rate. Scientific Reports. Vol. 15, pg. 15215, 2025, https://doi.org/10.1038/s41598-025-99564-x.

3 L. L. Davis, M. B. Hamner. Post-traumatic stress disorder: The role of the amygdala and potential therapeutic interventions-a review. Frontiers in Psychiatry. Vol. 15, pg. 1356563, 2024, https://doi.org/10.3389/fpsyt.2024.1356563.

4 L. Hancock, R. A. Bryant. Posttraumatic stress, stressor controllability, and avoidance. Behaviour Research and Therapy. Vol. 128, pg. 103591, 2020, https://doi.org/10.1016/j.brat.2020.103591.

5 K. Li, J. Li, Y. Li. The effects of social media usage on vicarious traumatization and the mediation role of recommendation systems usage and peer communication in China after the aircraft flight accident. European Journal of Psychotraumatology. Vol. 15, pg. 2337509, 2024, https://doi.org/10.1080/20008066.2024.2337509.

6 E. A. Holman, D. R. Garfin, R. C. Silver. It matters what you see: Graphic media images of war and terror may amplify distress. Proceedings of the National Academy of Sciences. Vol. 121, pg. e2318465121, 2024, https://doi.org/10.1073/pnas.2318465121.

7 A. Krizhevsky, I. Sutskever, G. E. Hinton. ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems. Vol. 25, pg. 1097-1105, 2012.

8 Y. LeCun, Y. Bengio, G. E. Hinton. Deep learning. Nature. Vol. 521, pg. 436-444, 2015, https://doi.org/10.1038/nature14539.

9 M. Ahmed, K. A. Hashmi, A. Pagani, M. Liwicki, D. Stricker, M. Z. Afzal. Survey and performance analysis of deep-learning based object detection in challenging environments. Sensors. Vol. 21, pg. 5116, 2021, https://doi.org/10.3390/s21155116.

10 K. He, X. Zhang, S. Ren, J. Sun. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pg. 770-778, 2016.

11 T. Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, S. Belongie. Feature pyramid networks for object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pg. 2117-2125, 2017.

12 A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929 (2020).

13 OpenAI. CLIP: Connecting text and images. OpenAI Blog, https://openai.com/blog/clip(2021).

14 M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen, X. Wang, X. Zhai, T. Kipf, N. Houlsby. Simple open-vocabulary object detection with vision transformers. Proceedings of the European Conference on Computer Vision, 2022.

15 X. Gu, T. Y. Lin, W. Kuo, Y. Cui. Open-vocabulary object detection via vision and language knowledge distillation. International Conference on Learning Representations (2022).

16 T. Huang. Content moderation by LLM: From accuracy to legitimacy. Artificial Intelligence Review. Vol. 58, 2025.

17 Google. Filter explicit results using SafeSearch. Google Search Help, https://support.google.com/websearch/answer/510(n.d.).

18 Reddit. Building better moderator tools. Reddit Engineering, https://www.reddit.com/r/RedditEng/comments/uly8s4/building_better_moderator_tools/(2022).

19 Instagram. How Instagram uses artificial intelligence to moderate content. Instagram Help Center, https://www.facebook.com/help/instagram/423837189385631 (n.d.).

20 B. Edelman. Empirical analysis of Google SafeSearch. Berkman Center for Internet Society, Harvard Law School, http://cyber.law.harvard.edu/people/edelman/google-safesearch(2003).

21 T. Fleischman. Considering trauma in tech design could benefit all users. Cornell Chronicle, https://news.cornell.edu/stories/2022/06/considering-trauma-tech-design-could-benefit-all-users(2022).

22 R. Rashed, F. Jahanbakhsh. What if moderation did not mean suppression? A case for personalized content transformation. arXiv:2509.22861 (2025).

23 T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, C. L. Zitnick. Microsoft COCO: Common objects in context. Proceedings of the European Conference on Computer Vision, pg. 740-755, 2014.

24 J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, L. Fei-Fei. ImageNet: A large-scale hierarchical image database. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pg. 248-255, 2009.

Appendix A

Package	Version	Purpose
Flask	3.0.3	Lightweight REST API server framework
Flask-CORS	4.0.1	Cross-Origin Resource Sharing for extension
PyTorch	2.4.0	Deep learning framework for model inference
torchvision	0.19.0	Pre-trained models and image transforms
transformers	4.44.2	Hugging Face library for OWL-ViT v2
Pillow	10.4.0	Image loading, resizing, and preprocessing
NumPy	1.26.4	Numerical operations and array manipulation
requests	2.32.3	HTTP requests for URL-based image fetching