Real-Time Inappropriate Content Detection on YouTube Using CLIP: A Zero-Shot Vision-Language Approach

Abdulghani A. Abied
a.abied@azu.edu.ly
Department of Computer Science,
Al-Zaytouna University, Tarhuna, Libya

Osayla Ali Alzawawi
Osayla.alrefak.edu.lu
Department of Computer Science,
Al-Refak University, Tripoli, Libya

Abstract:

The rapid expansion of online video platforms has significantly increased children's exposure to potentially harmful content, including violent and explicit material. Traditional moderation techniques, such as keyword-based filtering and static blocklists, are insufficient to address the dynamic and multimodal nature of modern digital media. This study proposes a real-time content moderation system that integrates a browser extension with a guardian monitoring platform, enabling continuous supervision of YouTube video consumption. The system leverages the CLIP (Contrastive Language–Image Pretraining) model to perform zero-shot classification of video frames by aligning visual and textual representations in a shared semantic space. The methodology involves periodic frame sampling, preprocessing, and similarity-based classification using predefined harmful and safe content labels. A dual-pass decision mechanism, combined with temporal consistency filtering, is employed to improve detection reliability and reduce false positives. Experimental evaluation on a labeled dataset demonstrates that the proposed system achieves an accuracy of 79%, with a high recall for harmful content detection. The results indicate that the system effectively prioritizes safety by minimizing undetected harmful content while maintaining acceptable precision levels. Overall, the proposed approach highlights the practical potential of zero-shot learning for real-time content moderation in dynamic environments. The system provides an effective, scalable, and privacy-aware solution for enhancing child safety in online video platforms.

Keywords: AI Content Moderation, CLIP Model, Inappropriate Content Detection, Real-Time Video Analysis, YouTube Safety.

المستخلص:

شـهدت منصـات الفيديو عبر الإنترنت نمواً سـريعاً، مما زاد من احتمالية تعرض الأطفال لمحتوى ضـار مثل العنف أو المواد غير اللائقة. تعتمد أسـاليب التصـفية التقليدية، مثل البحث بالكلمات المفتاحية والقوائم الثابتة، على آليات محدودة لا تسـتطيع مواكبة الطبيعة الديناميكية والمتعددة الوسـائط للمحتوى الرقمي الحديث. تقدم هذه الدراسـة نظاماً ذكيًا لمراقبة المحتوى في الزمن الحقيقي، يجمع بين إضـافة متصـفح (Browser Extension) ومنصـة إشـراف مخصـصـة للأهل، بهدف متابعة اسـتخدام الأطفال لموقع يوتيوب بشـكل مسـتمر. يعتمد النظام على نموذج CLIP

(Contrastive Language–Image Pretraining) لتنفيذ تصنيف صفري (Zero-Shot Classification) من خلال ربط التمثيلات البصرية والنصية في فضاء دلالي مشترك. تشمل المنهجية استخراج لقطات من الفيديو بشكل دوري، ومعالجتها، ثم تصنيفها اعتماداً على مقياس التشابه بين الصور وتسميات نصية محددة تمثل محتوى ضار أو آمن. كما يعتمد النظام على آلية تصنيف مزدوجة مدعومة بتتبع زمني للإطارات لتحسين دقة النتائج وتقليل الأخطاء. أظهرت نتائج التقييم أن النظام يحقق دقة إجمالية تبلغ 79%، مع قدرة عالية على اكتشاف المحتوى الضار، مما يقلل من احتمال تعرض الأطفال لمحتوى غير مناسب. وتؤكد النتائج فعالية استخدام التعلم الصفري في تطبيقات الزمن الحقيقي، مع توفير حل قابل للتوسع ويحافظ على خصوصية المستخدم.

الكلمات المفتاحية: إدارة المحتوى بإستخدام الذكاء الاصطناعي، أمان يوتيوب، تحليل الفيديو في الوقت الفعلي، كشف المحتوى غير اللائق، نموذج CLIP.

1. Introduction

The widespread accessibility of the internet has significantly increased the likelihood of children being exposed to harmful content, including violent, explicit, or otherwise inappropriate material, particularly on video-sharing platforms. Despite the implementation of safety mechanisms such as automated filtering systems and curated child-friendly modes, platforms like YouTube acknowledge that no moderation system can guarantee complete protection (YouTube Help, n.d.).

Traditional content filtering approaches, including keyword-based detection and static blocklists, remain widely used; however, these methods suffer from fundamental limitations. They often generate high false-positive rates, lack contextual understanding, and are unable to adapt efficiently to the rapidly evolving nature of online media (Raymond & Marchany, 2012).

In response to these challenges, recent advancements in automated moderation have demonstrated the effectiveness of machine learning techniques in detecting graphic violence and explicit content at scale. Major technology platforms, such as Facebook and YouTube, increasingly rely on such approaches to enhance content moderation processes (Papadamou et al., 2020). Nevertheless, these solutions are typically implemented at the platform level, while real-time, consumer-facing tools—particularly those operating directly within web browsers—remain limited.

Motivated by this gap, the present study proposes a real-time content moderation system that integrates a browser extension with an AI-driven backend based on the CLIP model. The system continuously samples video frames, performs semantic analysis using a zero-shot vision-language approach, and identifies potentially harmful content without requiring task-specific training. Upon detection, the system dynamically applies visual blurring and generates real-time notifications for guardians through a dedicated monitoring interface. This approach combines proactive automated filtering with human supervision, offering an effective and practical solution for enhancing child safety in online video environments.

2. Related work

The detection of inappropriate content in digital media has emerged as a critical research area, particularly on video-sharing platforms where children constitute a significant portion of the user base. Traditional moderation approaches, including keyword-based filtering and static blocklists, are widely implemented in educational and home environments; however, these techniques are easily circumvented and lack the semantic understanding required to accurately interpret multimedia content (Raymond & Marchany, 2012).

In recent years, research has increasingly shifted toward deep learning-based moderation techniques. EfficientNet, introduced by Tan and Le (2019), provides a highly scalable and efficient convolutional neural network architecture that improves performance while reducing computational cost. Furthermore, Tran et al. (2015) extended conventional image-based convolutional networks into three-dimensional architectures, enabling the capture of temporal dynamics in video data. Recurrent models such as Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) have also demonstrated effectiveness in modeling sequential dependencies within video streams (Hochreiter & Schmidhuber, 1997; Graves & Schmidhuber, 2005).

Additionally, multimodal approaches that integrate both visual and auditory information have enhanced the detection of complex and context-sensitive content. For example, Wang, Shrivastava and Gupta (2017) demonstrated that combining multiple data modalities can significantly improve classification performance in challenging scenarios.

Several studies have specifically addressed the issue of harmful content targeting children on video platforms. Papadamou et al. (2020) investigated disturbing content on YouTube and highlighted the limitations of existing moderation systems. Similarly, Yao, Mamitsuka and Zhu (2018) proposed label-aware hierarchical models to improve multi-label video classification, while Cho et al. (2014) introduced the RNN Encoder–Decoder framework for extracting semantic representations from textual data such as subtitles and metadata. Collectively, these approaches emphasize the importance of integrating spatial, temporal, and textual features to achieve more accurate content filtering.

A notable advancement in this domain is the CLIP (Contrastive Language–Image Pretraining) model developed by Open AI, which enables zero-shot learning by aligning visual and textual representations in a shared embedding space (Radford et al., 2021). This capability allows systems to classify previously unseen content categories without the need for retraining on domain-specific datasets. Building on this innovation, the present study adopts the CLIP model to develop a real-time, in-browser content moderation system specifically designed to enhance child protection in online video environments.

3. Methodology

The proposed system leverages the CLIP (Contrastive Language–Image Pretraining) model developed by Open AI to enable real-time moderation of YouTube video content through

the detection of inappropriate visual elements. CLIP facilitates zero-shot classification by embedding both images and textual descriptions into a shared semantic space, allowing for direct comparison without requiring task-specific training data (Radford et al., 2021). This capability enables the system to identify harmful content dynamically while maintaining flexibility and scalability.

3.1 Video Frame Extraction and Pre-processing

The system employs a browser extension that continuously captures video frames from the YouTube player at one-second intervals. Each captured frame is resized to a resolution of 224 × 224 pixels and normalized according to the input specifications required by the CLIP model (Radford et al., 2021).

Unlike traditional machine learning approaches, the proposed system does not rely on a predefined or stored training dataset. Instead, it performs real-time inference directly on incoming frames, thereby preserving user privacy and minimizing storage and computational overhead.

3.2 Feature Extraction Using CLIP

The CLIP model consists of two jointly trained components: a visual encoder and a text encoder, both designed to project inputs into a shared latent embedding space (Radford et al., 2021). This architecture enables semantic alignment between images and textual descriptions.

3.2.1 Visual Encoder

Each sampled video frame is processed through the CLIP visual encoder to generate a high-dimensional feature vector that captures the semantic characteristics of the image.

3.2.2 Text Encoder

A predefined set of textual labels representing both harmful and neutral categories—such as *“porn,” “hentai,” “sexy,” “blood,”* and *“neutral”*—is encoded using the CLIP text encoder. These embeddings serve as reference vectors for classification.

3.2.3 Zero-Shot Similarity Matching

For each frame, cosine similarity is computed between the visual embedding and all corresponding text embeddings. The label with the highest similarity score is selected as the predicted class. If the similarity score exceeds a predefined threshold and corresponds to a harmful category, the system flags the frame as inappropriate.

3.3 System Architecture

The proposed system follows a modular architecture composed of several interconnected components that collectively enable real-time content analysis and response. The primary components include:

- **Browser Extension Frame Sampler:** Captures video frames periodically during playback.
- **CLIP Visual Encoder:** Converts frames into semantic feature representations.
- **CLIP Text Encoder:** Transforms predefined content labels into embedding vectors.
- **Similarity Engine:** Computes cosine similarity between image and text embeddings.
- **Classification Module:** Determines whether content is appropriate based on similarity scores and predefined thresholds.

This modular design enhances system scalability and allows for efficient integration of additional components or future improvements.

3.4 Real-Time Integration

The system is deployed as a backend inference service that communicates directly with the browser extension. Upon detecting inappropriate content, the system triggers predefined actions such as blurring the video, pausing playback, or sending alerts to a guardian interface.

All processing is performed in real time, ensuring immediate response to harmful content without requiring local storage or offline analysis. This design supports continuous monitoring while maintaining user privacy and system efficiency.

4. Algorithm

4.1 System Architecture Overview

The proposed Parental Monitor system is designed using a microservices-based architecture that ensures scalability, modularity, and efficient real-time processing. The system comprises three primary components: a **Content Capture Module** implemented as a browser extension, an **AI Classification Engine** deployed using a Flask-based backend, and a **Decision Management Dashboard** developed by Django for monitoring and control.

These components operate collaboratively to capture video frames, perform semantic analysis, and manage detection outcomes in real time. The distributed architecture enables efficient workload handling and facilitates future system expansion.

4.2 CLIP-Based Content Classification

4.2.1 Model Selection and Mathematical Foundation

The system uses OpenAI's **CLIP ViT-B/32** model for zero-shot classification of video frames. CLIP maps both images and text descriptions into a shared embedding space using contrastive learning:

$$\text{similarity}(I, T) = \cos(\phi(I), \psi(T))$$

Here, $\phi(I)\phi(I)$ represents the embedding of the input image, and $\psi(T)\psi(T)$ denotes the embedding of the corresponding text label. The cosine similarity function quantifies the semantic alignment between the two modalities.

The classification framework adopts a dual-pass strategy to improve decision robustness. In the first pass, a predefined set of harmful labels—such as *violence*, *gore*, and *explicit content*—is used to evaluate potential risk. In the second pass, safe labels—including *family-friendly* and *child-safe*—are applied to validate and refine the classification outcome.

The system computes the **maximum similarity score** among harmful labels and the **average similarity score** across safe labels. A threshold-based decision rule is then applied:

- If the harmful similarity score exceeds **0.30**, the frame is classified as inappropriate.
- If the harmful score is low and the average safe score exceeds **0.10**, the frame is classified as safe.

This dual-path decision mechanism enhances classification accuracy and reduces ambiguity in borderline cases.

4.2.2 Temporal Consistency

To improve reliability and reduce false positives, the system incorporates a temporal consistency mechanism that evaluates consecutive frame predictions within the same video stream.

- Safe content is recorded immediately upon detection (once per video).
- Inappropriate content is only confirmed if detected across **five consecutive frames**.

This approach minimizes the impact of transient visual noise and ensures that only persistent harmful content triggers system actions.

4.3 Image Processing Pipeline

Captured frames are transmitted in base64 format and undergo a structured pre-processing pipeline prior to classification. The frames are decoded into PIL image format, converted to RGB color space, resized to 224 × 224 pixels, and normalized according to the input requirements of the CLIP model (Radford et al., 2021).

These pre-processing steps ensure compatibility with the model architecture and maintain consistency in inference performance across varying input conditions.

4.4 Decision Fusion and Confidence Scoring

The final classification decision is derived using a fusion mechanism that combines both harmful and safe similarity scores. The system dynamically evaluates the most reliable prediction by comparing the maximum harmful score with the aggregated safe score.

Threshold values of **0.30** for harmful classification and **0.10** for safe classification were determined empirically to balance precision and recall. Additionally, the requirement of consecutive detections for harmful content further strengthens classification confidence.

This decision fusion strategy enables the system to achieve robust performance while minimizing both false positives and false negatives, thereby aligning with the primary objective of ensuring child safety in real-time video environments.

5. System Interact Flow

The Parental Monitor system adopts a distributed interaction model in which responsibilities are divided between a **Django-based web dashboard**, designed for parental use, and a **Flask-based backend**, responsible for device-side monitoring and communication. This separation ensures efficient user management, real-time monitoring, and secure system configuration.

The interaction process begins with **parent registration**, where users create an account through the Aman Django dashboard. This step establishes the primary user identity and provides access to all system functionalities, as illustrated in **Figure 1**.

The interaction process begins with **parent registration**, where users create an account through the Aman Django dashboard. This step establishes the primary user identity and provides access to all system functionalities, as illustrated in **Figure 1**.

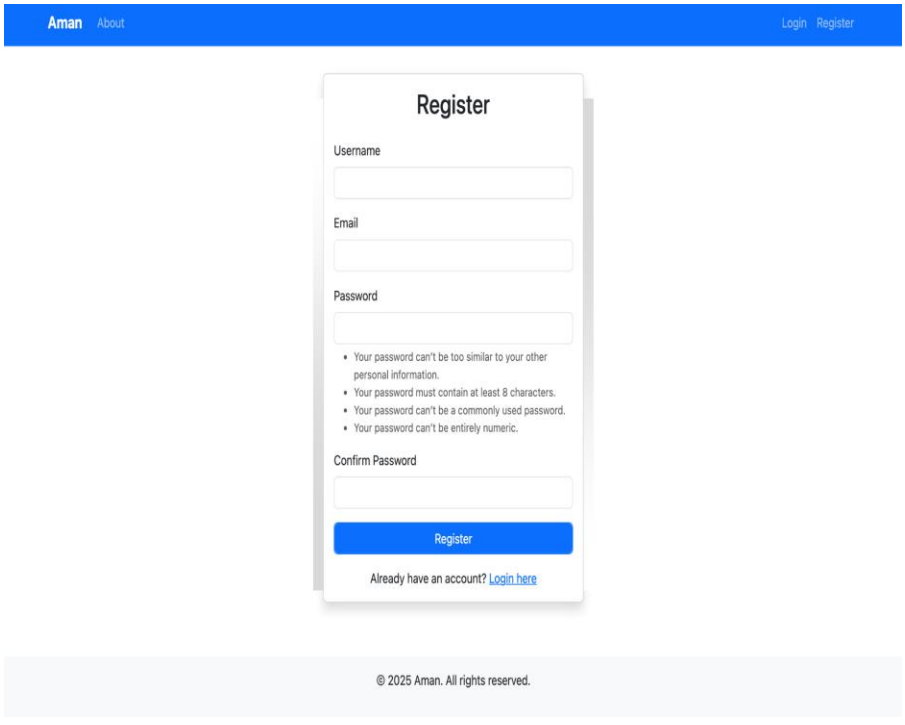

Figure 1: Parent Registration Interface on the Django-Based Dashboard

After registration, parents can create individual profiles for their children through the *Manage Children* interface. This functionality allows the system to associate monitoring activities with specific users by recording essential information such as name and age, as shown in **Figure 2**.

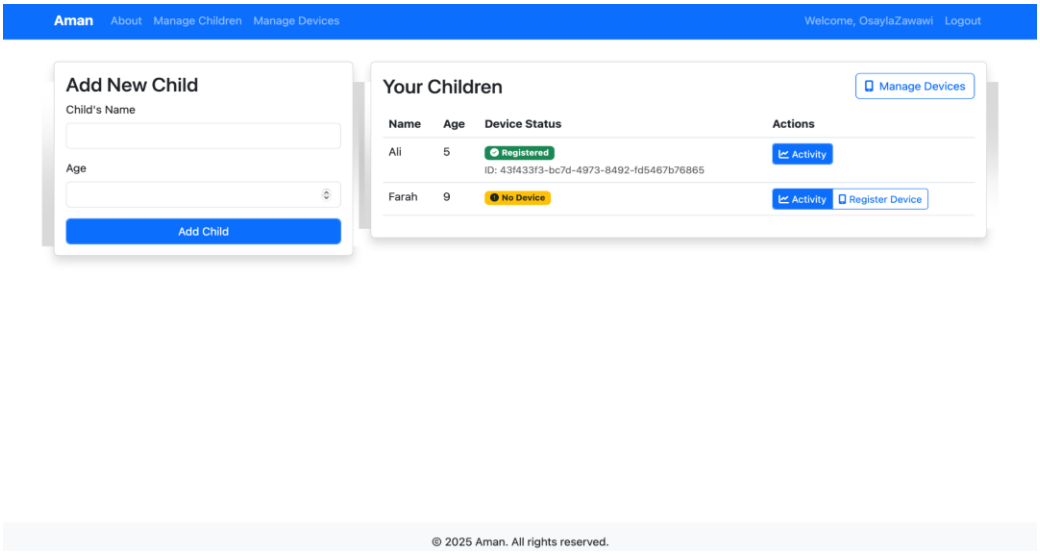

Figure 2: Child Profile Management Interface for Adding and Viewing Registered Children

The system then supports **device association**, where each child profile can be linked to a dedicated monitoring device via the *Manage Devices* interface. The device is registered through the Flask backend, enabling independent tracking of activity for each user, as illustrated in **Figure 3**.

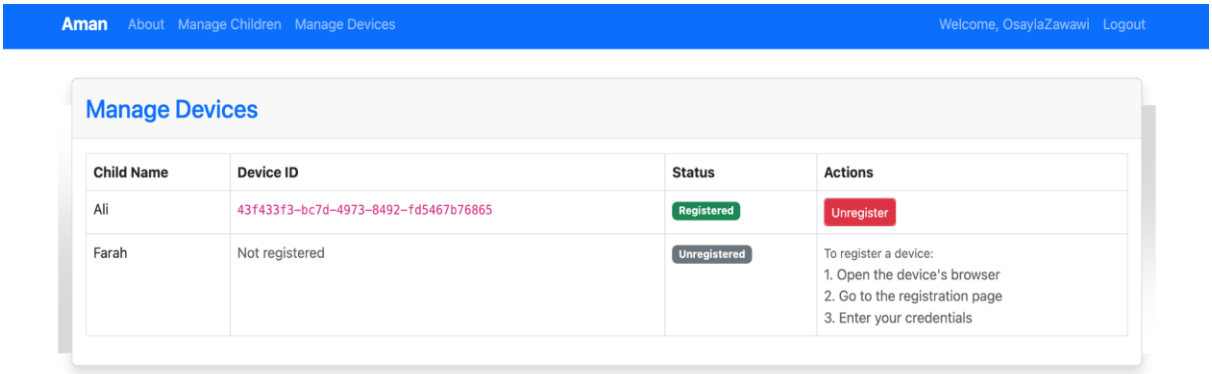

Figure 3: Device Management Interface Showing Device Registration and Status

Next, the system performs **device login and activation**, where the parent authenticates the monitoring device using the Flask backend interface. Once verified, the device is designated

as the active monitoring device for the selected child, allowing the system to initiate real-time monitoring, as shown in **Figures 4**.

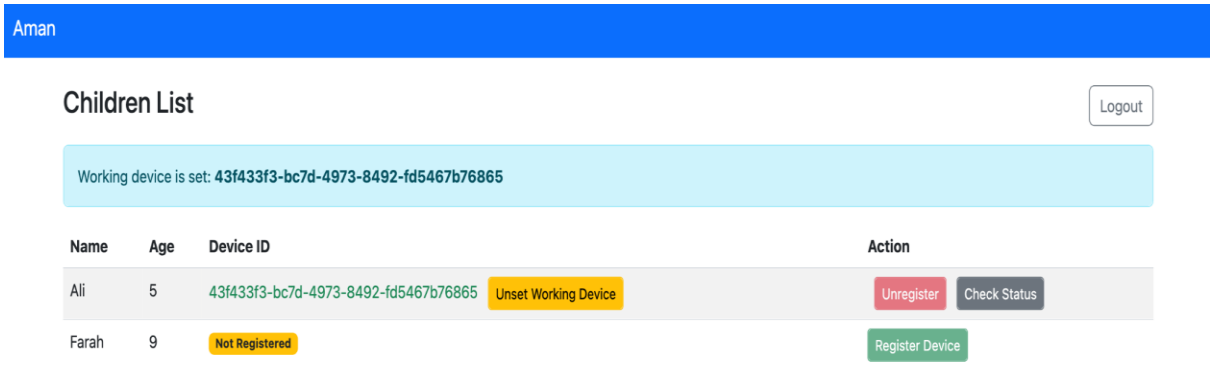

Figure 4: Successfully Registered Device Set as the Active Monitoring Device

Following activation, the system continuously tracks user activity and presents the results through a centralized **child activity dashboard**. This dashboard displays detailed logs of both normal and flagged content, including timestamps, video titles, predicted content categories, model confidence scores, and system actions such as blurring. This functionality is illustrated in **Figure 5**, enabling guardians to monitor user behavior effectively.

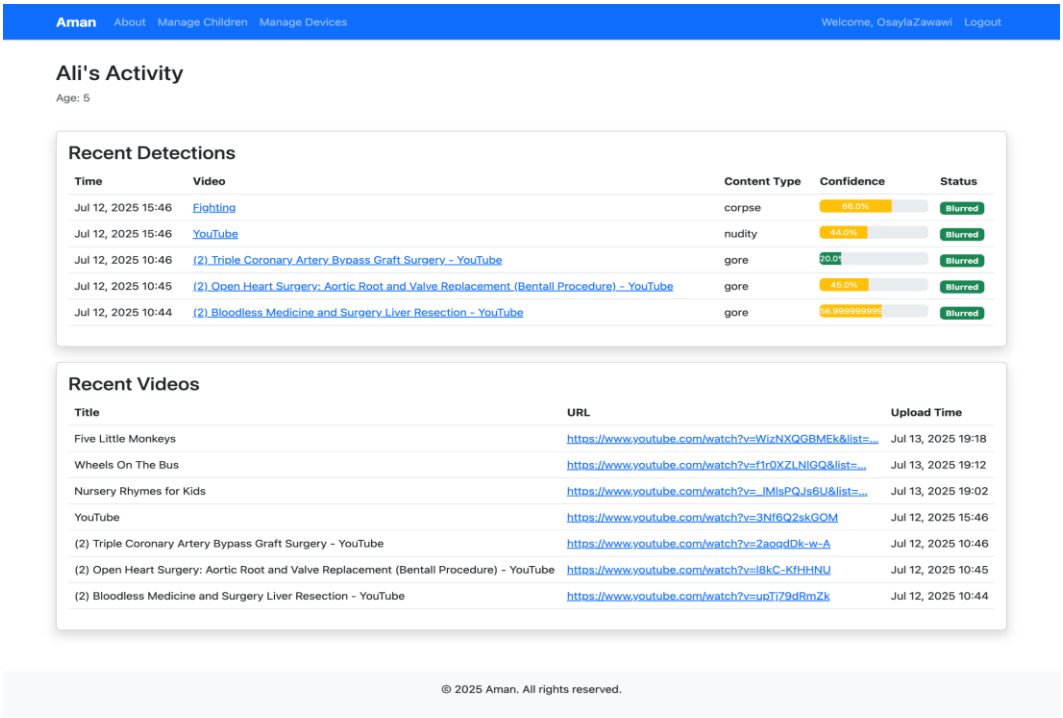

Figure 5: Child Activity Dashboard Displaying Detected Content and Viewing History

When inappropriate content is detected with high confidence, the system performs **real-time visual intervention** by applying a blur overlay to the video content while preserving contextual elements such as the video title and platform interface. This process is demonstrated

in **Figure 6**, ensuring immediate protection without completely interrupting the viewing experience.

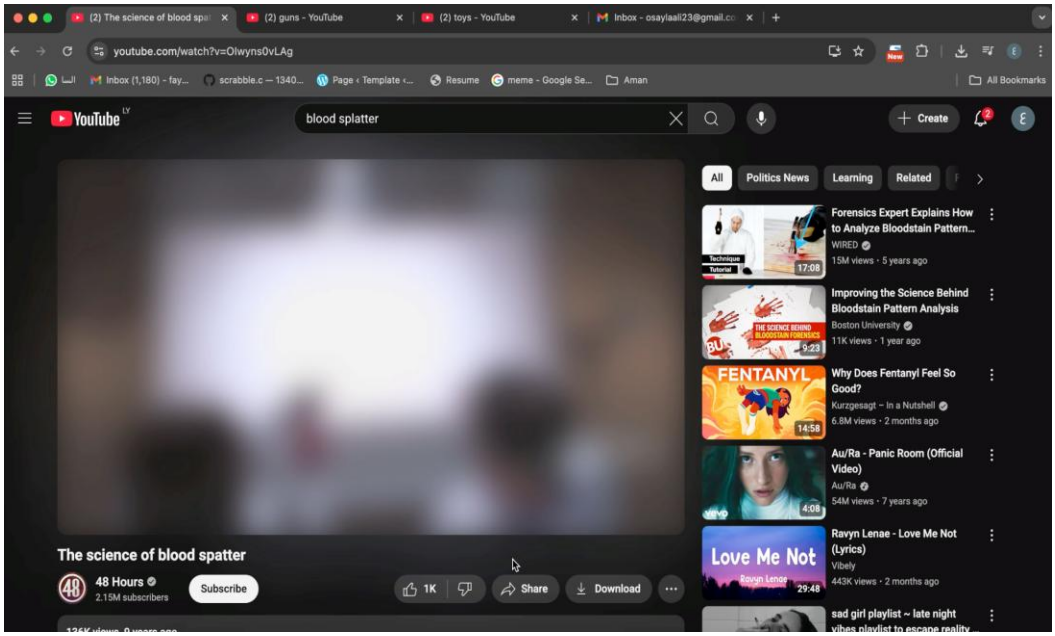

Figure 6: Real-Time Video Blurring Applied to Detected Inappropriate Content

In addition to visual intervention, the system generates **automated email notifications** sent to the registered parent. These alerts include detailed information such as the video URL, detection timestamp, identified content type, and model confidence score. This mechanism, illustrated in **Figure 7,** enables timely parental awareness and appropriate response.

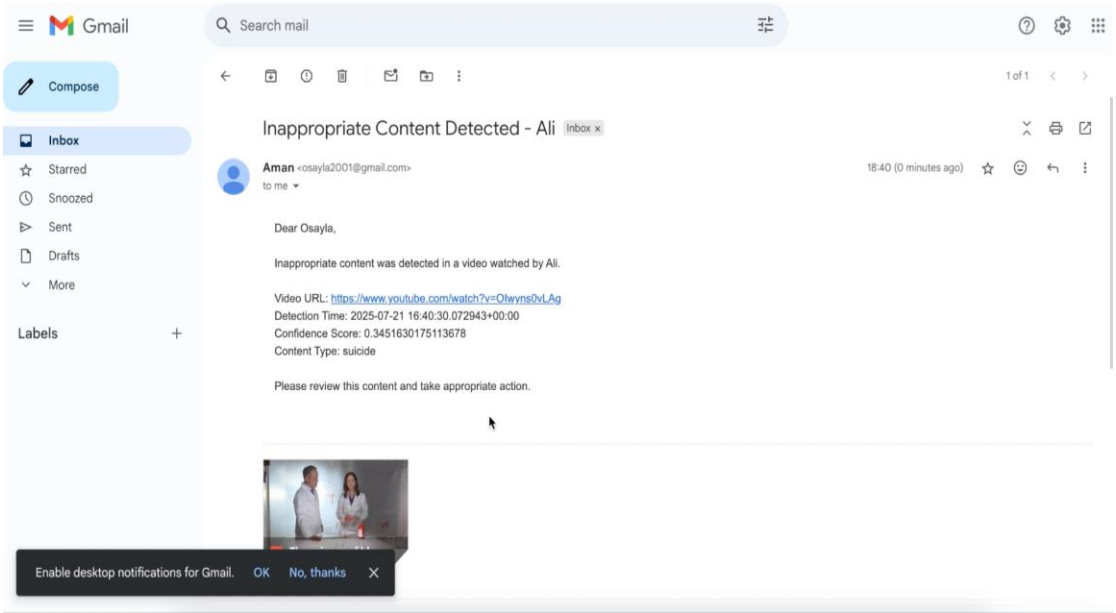

Figure 7: Automated Email Notification Sent Upon Detection of Harmful Content

Overall, the system interaction flow integrates user registration, profile management, device configuration, real-time monitoring, and alert mechanisms into a unified framework that enhances child safety in online video environments.

6. Model Performance

The proposed model was evaluated using a labelled test dataset consisting of 130 images, including both harmful and safe content categories. The evaluation results show that the system achieved 56 true positives (correctly detected harmful content), 9 false negatives (harmful content not detected), 20 false positives (safe content incorrectly classified as harmful), and 45 true negatives (correctly identified safe content). Based on these results, the model attained an overall accuracy of 78%.

To further assess performance, standard evaluation metrics were calculated, including precision and recall (Sokolova & Lapalme, 2009). The recall for harmful content was 0.86, indicating that 86% of actual harmful instances were successfully detected by the model. The precision for harmful predictions was 0.74, demonstrating that approximately three-quarters of the flagged content was indeed inappropriate.

In contrast, the model achieved a recall of 0.69 for safe content and a precision of 0.83, indicating that when the system classified content as safe, it was correct in 83% of cases. These results reflect a balanced but safety-oriented classification behaviour.

Overall, the evaluation metrics highlight that the system prioritizes minimizing false negatives, thereby reducing the likelihood of undetected harmful content, even at the expense of occasionally misclassifying safe material. This trade-off is consistent with the primary objective of the system, which is to enhance child protection in online video environments by favouring safety over strict precision.

6.1 Ablation Study and Baseline Comparison

To address the contribution of the proposed Dual-Pass and Temporal Consistency modules, we conducted a comparative analysis against a baseline. The results (summarized in Table 1) demonstrate that the proposed system significantly outperforms the baseline, particularly in reducing false positives and improving the stability of detections.

Table 1: Performance comparison between the baseline CLIP model and the proposed system

Configuration	Harmful Recall (%)	Harmful Precision (%)	Overall F1-Score
Baseline	72	61	66
Proposed (CLIP + Dual-Pass + Temporal)	86	74	79

7. Limitations

While the results of this study demonstrate the potential of utilizing CLIP for real-time detection of inappropriate content on YouTube, several limitations must be acknowledged. First and foremost, this work is intended as a Proof of Concept (PoC) rather than a production-ready system. The primary constraints are outlined below:

- **Sample Size and Diversity:** The evaluation was conducted on a relatively small dataset of 130 images. While these preliminary results are promising, they may not fully represent the vast and heterogeneous nature of YouTube's visual content, which includes diverse lighting conditions, cultural contexts, and varying video qualities.
- **Preliminary Nature of Results:** The findings presented herein are **preliminary**. Extensive benchmarking against larger, more diverse datasets is required to establish statistical significance and ensure the model's robustness across different categories of prohibited content.

8. Conclusion

The proposed Parental Monitor system represents a significant advancement in AI-driven parental control solutions by integrating zero-shot content classification with real-time deployment within a scalable microservices architecture. By leveraging the CLIP model (Radford et al., 2021), the system effectively detects inappropriate content with a high recall for harmful material, thereby minimizing the risk of exposure to children.

The incorporation of a dual-pass classification strategy, combined with temporal consistency and a structured decision fusion mechanism, enhances the robustness of the system against false positives while improving the overall reliability of detection outcomes. These design choices contribute to a balanced and safety-oriented classification approach.

Experimental evaluation demonstrates that the system achieves reliable accuracy and real-time responsiveness while maintaining efficient resource utilization and a user-friendly interface. Furthermore, the modular architecture enables seamless interaction between the content capture module, AI inference backend, and guardian dashboard, ensuring low-latency processing and effective monitoring.

This study highlights the practical applicability of zero-shot learning techniques in dynamic, real-world environments where adaptability and semantic understanding are essential. Although certain challenges remain, such as the misclassification of safe content and platform-related constraints, the system architecture supports continuous improvement and scalability.

Overall, the Parental Monitor provides a comprehensive and effective solution for enhancing child safety in online video environments. In addition to offering actionable insights and real-time alerts for parents, this work contributes to the broader field of ethical artificial intelligence and establishes a flexible foundation for future advancements in intelligent content moderation systems.

9. References

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y. (2014) 'Learning phrase representations using RNN encoder–decoder for statistical machine translation', *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734.

Graves, A. and Schmidhuber, J. (2005) 'Framewise phoneme classification with bidirectional LSTM', *Neural Networks*, 18(5–6), pp. 602–610.

Hochreiter, S. and Schmidhuber, J. (1997) 'Long short-term memory', *Neural Computation*, 9(8), pp. 1735–1780.

Papadamou, S., Zannettou, S., Blackburn, J., De Cristofaro, E., Stringhini, G. and Sirivianos, M. (2020) 'Disturbed YouTube for Kids: Characterizing and detecting inappropriate videos targeting young children', *Proceedings of the IEEE Symposium on Security and Privacy (SP)*, pp. 1321–1338.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I. (2021) 'Learning transferable visual models from natural language supervision', *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 8748–8763

Raymond, D. and Marchany, S. (2012) 'The evolution of cyber security in the home', *IEEE Security & Privacy*, 10(5), pp. 50–56.

Tan, M. and Le, Q.V. (2019) 'EfficientNet: Rethinking model scaling for convolutional neural networks', *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 6105–6114.

Tran, D., Bourdev, L., Fergus, R., Torresani, L. and Paluri, M. (2015) 'Learning spatiotemporal features with 3D convolutional networks', *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 4489–4497.

Yao, L., Mamitsuka, H. and Zhu, S. (2018) 'Deep learning for multi-label video classification using structured label representation', *Proceedings of the AAAI Conference on Artificial Intelligence*.

YouTube Help (n.d.) *YouTube Kids Parental Guide*. Available at:

https://support.google.com/youtubekids/answer/6172308

Wang, X., Shrivastava, A. and Gupta, A. (2017) 'A-Fast-RCNN: Hard positive generation via adversary for object detection', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2606–2615.