Content Moderation System Write-Up

Approach

This content moderation system leverages advanced AI models to analyze video content for inappropriate material. It uses **CLIP (ViT-B/32)** for visual analysis, extracting frames at one-second intervals and evaluating them against predefined prompts like "nudity," "graphic violence," or "a safe scene." CLIP's zero-shot classification assigns confidence scores to each prompt, flagging frames with unsafe content above a 0.5 threshold. For audio, **Whisper (base model)** transcribes the video's audio, and a predefined list of vulgar words is checked against the transcript. The system runs on Google Colab with GPU support for efficiency, using libraries like OpenCV, MoviePy, and Pillow for frame extraction and processing. The final decision flags a video as "unsafe" if either visual or audio issues are detected, displaying up to five flagged frames or a random safe frame for verification.

Challenges

1. **Threshold Sensitivity**: Setting a 0.5 threshold for CLIP's confidence scores required balancing false positives and negatives. Too high, and subtle issues might be missed; too low, and safe content could be flagged.
2. **Contextual Nuance**: CLIP struggles with contextual interpretation (e.g., distinguishing artistic nudity from explicit content), and Whisper's transcription may miss slang or misinterpret accents.
3. **Resource Constraints**: Frame extraction and CLIP inference are computationally intensive, especially for long videos, even with GPU acceleration.
4. **Vulgar Word List**: The hardcoded list of vulgar words is limited and may miss synonyms or context-specific phrases, requiring manual updates.
5. **Frame Sampling**: Extracting one frame per second might skip brief but problematic content, necessitating finer granularity for short, high-risk videos.

Why It's Awesome

This system is a robust, scalable solution for automated content moderation, combining **multimodal AI** (CLIP for visuals, Whisper for audio) to tackle both visual and auditory risks comprehensively. Its **zero-shot capability** via CLIP allows it to generalize to new content types without retraining, making it adaptable to evolving moderation needs. The system is **user-friendly**, running in Colab with clear outputs like flagged frames and transcripts, enabling quick human review. It’s also **efficient**, processing only one frame per second to balance accuracy and speed. By catching explicit visuals and vulgar language, it addresses key platform safety concerns, reducing manual moderation workloads. Future enhancements could include dynamic thresholds, contextual analysis, or real-time processing, but the current setup already offers a powerful, accessible tool for safer online spaces.