ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25dec1148

Advance Framework and Methodologies for Automated Video Content Moderation

A Comprehensive Analysis of Vision-Language Models, Safety Benchmarking, and Pipeline Optimization

Niharika Patidar[1]; Dr. Sachin Patel[2]

[1]Student, Institute of Sciences, Department of Computer Science, SAGE University, Indore
[2]Head of Department (CSIT), SAGE University, Indore

Publication Date: 2025/12/27

Abstract: Today video content now defines over 82% of global internet traffic, the explosion of user-generated content (UGC) has fundamentally overwhelmed our ability to keep platforms safe. Traditional moderation systems were simply not designed for this unprecedented volume of content. They are too slow, too rigid, and completely blind to the cultural context that defines modern toxicity. This is where the true innovation lies: CCHE (Content Classification and Harm Evaluation) is an open-source pipeline for downloading short-form video from public sources, sampling representative frames, and inferring content typology and age suitability using a large vision-language model. The shift towards Large Vision-Language Models (LVLMs) is not an option; it is an urgent necessity. This paper furnishes a comprehensive technical examination of this transition, contrasting proprietary titans like GPT-4o with the thrilling advancements in open-source alternatives, and critically dissecting the engineering frameworks from ingestion efficiency to industrial deployment required for robust, real-world content safety.

How to Cite: Niharika Patidar; Dr. Sachin Patel (2025) Advance Framework and Methodologies for Automated Video Content Moderation. *International Journal of Innovative Science and Research Technology*,10(12), 1804-1810. https://doi.org/10.38124/ijisrt/25dec1148

I. INTRODUCTION

User-generated video—especially shorts and reels—poses scaling challenges for platforms attempting to enforce nuanced content policies. We are looking closely at the best modern AI systems that can handle text, pictures, and videos. We're comparing the systems owned by big companies, like OpenAI's GPT-4o, with the free and open-to-all options, such as Qwen2.5-VL and InternVL2.5. Our analysis goes beyond just the AI's internal structure; we also examine the practical tools needed to actually use these AIs in the real world.

Finally, we are breaking down the biggest technical problems when dealing with video, specifically the time it takes to get a video, unpack it into individual pictures (frames), and then intelligently choose which pictures the AI should look at. We compare how efficient different software libraries, like Decord versus OpenCV, are at handling this task.

- *In essence:*

We're reviewing the top AIs that see and talk, how to safely deploy them in businesses, and how to fix the main speed bumps involved in processing video data.

II. THE STATE OF DIGITAL VIDEO IN 2025: TRAFFIC, TRENDS, AND THE SAFETY CRISIS

- *The Hegemony of Short-Form Video*

By 2025, the transformation of the internet into a video-first medium is complete. Data indicates that video content now constitutes approximately 82% of all global internet traffic, a statistic that underscores the massive bandwidth and storage challenges facing infrastructure providers.

A notable trend in 2025 is the surge of "faceless content"—videos that focus on products, processes, or automated visualizations without featuring a human presenter. This genre, often utilized for rapid monetization, presents unique moderation challenges as it lacks the facial cues often used by legacy safety systems to detect age or identity.

- *The Moderation Gap*

We lost the war on volume. Millions of videos are uploaded daily, creating a "moderation gap"—a logistical sinkhole where harmful content, ranging from hate speech and violence to non-consensual sexual imagery, can proliferate virally before human teams can intervene.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25dec1148

➢ *Legacy Automated Systems Suffer from Critical Deficiencies:*

- *The Tragedy of a Static Frame (Contextual Blindness):*

Traditional computer vision models operate on isolated frames, failing to grasp temporal context. They look at one frame and miss the whole movie. Classifiers can flag a staged theatrical performance as a genuine assault, leading to high rates of False Positives (FP).

- *Scalability Limits:*

The computational cost of processing every frame of a high-resolution video is prohibitive. Aggressive down-sampling often misses brief but critical violations (e.g., a single frame of subliminal hate imagery).

- *Adaptability Latency:*

The "industrial moderation regulation" cycle is painfully slow. Updating a standard classifier to detect a new visual symbol of hate requires weeks of retraining the entire backbone, whereas viral trends evolve in hours.

The integration of LVLMs offers a transformative solution. By fusing visual encoders with the reasoning capabilities of Large Language Models (LLMs), these systems can interpret complex, multimodal narratives.

III. THE VISION-LANGUAGE MODEL LANDSCAPE: A TECHNICAL COMPARATIVE ANALYSIS

The year 2025 marks the maturity of "native" multimodal models. The current generation features deeply integrated architectures capable of genuine cross-modal reasoning.

➢ *The Reigning Champion: GPT-4o ("Omni")*

OpenAI's GPT-4o represents the current apex of proprietary multimodal capability.

- *Native Multimodality:*

The "omni" architecture is a game-changer. It processes text, audio, image, and video inputs in a single neural network, preserving the rich signal often lost in translation between separate modules.

- *Latency:*

This unified architecture drastically reduces latency. GPT-4o can respond to audio inputs in as little as 232 milliseconds (average 320ms), approximating human conversational response times, which is critical for real-time live stream moderation.

- *The Price Tag:*

The brilliance comes with a brutal cost. Standard pricing is approximately $5.00 per 1 million input image tokens, making it an economic non-starter for high-volume applications.

- *Performance:*

In safety benchmarks, GPT-4o demonstrates superior "agreement" with human annotators (over 88%) compared to specialized fine-tuned models. It excels in complex reasoning tasks where the "safety" of a video depends on subtle interplay between audio and visual cues.

➢ *The Open-Source Revolution*

For platforms struggling with cost and data privacy, the ability to self-host open-source VLMs offers a compelling alternative.

- *Qwen2.5-VL (Alibaba)*

Qwen2.5-VL serves as a flagship open-source model, designed to rival GPT-4o in visual understanding capabilities.

✓ *Adaptive Resolution:*

Unlike models that resize all inputs to a fixed square (e.g., 224x224 or 336x336), Qwen2.5-VL handles variable resolutions and aspect ratios dynamically. This is crucial for moderating mobile-first vertical video (9:16 aspect ratio) without introducing distortion artifacts.[10]

✓ *Long-Context Extrapolation:*

Utilizing the YaRN (Yet another RoPE extension) technique, Gwen 2.5-VL supports context windows up to 32,768 tokens. This allows it to ingest and reason over videos exceeding 20 minutes in length, a capability often lacking in standard VLMs.

✓ *Agentic Capabilities:*

Beyond passive analysis, the model demonstrates "agentic" abilities, such as interpreting GUI environments and executing actions. In a moderation context, this could theoretically allow the model to navigate internal moderation tools directly.

- *InternVL 2.5 (OpenGVLab)*

InternVL 2.5 adopts a "Progressive Scaling Strategy" to optimize the alignment between its vision encoder (InternViT) and the language model.

✓ *Architecture:*

It pairs a massive vision encoder (up to 6B parameters) with LLMs of varying sizes (from 1B to 78B). This decoupling allows for flexible deployment: a smaller LLM can be used for rapid triage, while the 78B variant is reserved for complex policy adjudication.

✓ *Performance:*

On the MMMU benchmark, InternVL 2.5 achieves over 70% accuracy, matching the performance of proprietary leaders like GPT-4o and Claude 3.5 Sonnet. It is explicitly optimized to reduce "hallucination," a critical safety feature to prevent the wrongful flagging of benign content.

✓ *Visual PRM:*

The release includes VisualPRM, a Process Reward Model that enhances reasoning by evaluating the intermediate steps of the model's thought process. This improves reliability in complex "Chain-of-Thought" moderation tasks.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25dec1148

- *MiniGPT4-Video*

For scenarios where latency and throughput are prioritized over deep reasoning (e.g., initial pre-filtering), MiniGPT4-Video offers a specialized architecture.

✓ *Joint Space-Time Attention:*

It utilizes a custom vision transformer that attends to both spatial and temporal dimensions simultaneously.

✓ *Benchmark Efficiency:*

This specialized focus allows it to achieve 41.78% accuracy on the TVQA-long benchmark, outperforming previous generalist methods by nearly 15%. It proves that smaller, domain-specific architectures can outperform larger generalist models in targeted video understanding tasks.

➢ *comparative Feature Matrix*

Table 1 Comparative Feature Matrix

Feature	GPT-4o	Qwen2.5 -VL	InternVL 2.5	MiniGPT4-Video
Type	Proprietary API	Open Source (Apache 2.0)	Open Source	Open Source
Max Context	128k Tokens	32k Tokens (YaRN)	Variable (Long Context)	Optimized for VideoQA
Input Modalities	Text, Audio, Image, Video	Text, Image, Video	Text, Image, Video	Text, Video
Key Strength	Unified multimodal reasoning, low latency	Adaptive resolutio n, GUI agent capabiliti es	Strong vision encoder, hallucination reduction	High efficiency on long video QA
Pricing/Cost	High ($5/1M img tokens)	Self-host ed (GPU depende nt)	Self-hosted (Scalable)	Low (Efficient inference)
Deployment	Cloud API	Cloud/O n-Prem/ Edge	Cloud/On-Prem	Cloud/On-Prem

IV. SPECIALIZED FRAMEWORKS FOR AUTOMATED MODERATION

General-purpose VLMs provide the "brain", but effective moderation requires a "body"—a framework that integrates the model into an industrial workflow.Two prominent frameworks emerging in 2024-2025 demonstrate distinct approaches to this challenge.

➢ *KuaiMod: The "Common-Law" Framework for Social Media*

KuaiMod conceptualizes moderation not as binary classification, but as a dynamic legal system where "policies" (laws) evolve and models (judges) apply reasoning.

- *VLM as Policy Agent:*

It treats the VLM as an agent that interprets natural language policy documents. This means platform administrators can update moderation rules simply by updating the text prompt, without retraining the model.

- *Chain-of-Thought (CoT) Reasoning:*

KuaiMod mandates that the model generates a rationale for its decision before outputting a verdict. This "interpretable critique" forces the model to ground its decision in specific visual evidence.

- *Sparse Feedback Learning:*

Recognizing that high-quality labeled data is scarce, KuaiMod utilizes a reinforcement learning loop based on "sparse" signals—user reports, successful appeals, and moderator overturns. This allows the system to continuously adapt to the evolving "case law" of the platform.

✓ *Impact:*

Deployment on Kuaishou platforms demonstrated a 20% reduction in user reporting rates, suggesting that the system effectively removed toxic content before users encountered it. Furthermore, the improved environment quality led to increases in Daily Active Users (DAU) and Average Usage Time (AUT).

➢ MonitorVLM: Industrial Safety and Compliance

MonitorVLM targets the rigid, high-stakes environment of industrial safety (mining, construction).

- *Clause-Specific Filtering:*

It does not vaguely look for "unsafe" acts. It uses a Clause Filter (CF) to map specific regulatory clauses to visual triggers, constraining the VLM's search space and improving accuracy.

- *Temporal Order Verification:*

MonitorVLM is explicitly trained to recognize the sequence of events. For safety protocols, order is paramount—a nuance generic VLMs often fail to distinguish.

- *Lightweight Integration:*

The framework includes a web-based interface that converts VLM outputs into timestamped, legally compliant violation reports. This bridges the gap between AI inference and the bureaucratic requirements of safety audits.

V. SAFETY BENCHMARKING AND EVALUATION METHODOLOGIES

As reliance on automated systems grows, the methodology for evaluating their "safety" has become a critical field of research.

➢ *Video-SafetyBench: The Standard for Multimodal Risk*

Introduced to address the lack of video-specific safety evaluations, Video-SafetyBench is the first comprehensive benchmark designed to test LVLMs against "video-text attacks".

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25dec1148

- *Risk Taxonomy:*

The benchmark categorizes safety risks into 13 primary categories and 48 fine-grained subcategories 20.

✓ S1-Violent Crimes: Terrorism, assault, animal abuse.
✓ S2-Non-Violent Crimes: Fraud, theft, cybercrime.
✓ S3-Sex-Related Crimes: Harassment, non-consensual sexual imagery.
✓ S11-Suicide & Self-Harm: Promotion of self-injury or eating disorders.
✓ S10-Hate Speech: Visual or auditory slurs targeting protected groups.

- *The "Benign vs. Harmful" Protocol:*

A key innovation is the paired query strategy 19:

✓ *Harmful Query:*

A direct request for malicious content (e.g., "How do I construct the explosive device shown in this video?").

✓ *Benign Query:*

An ostensibly harmless request that becomes toxic in context (e.g., "Explain the chemistry experiment shown" when the video depicts meth production). This duality tests the model's ability to infer intent from the relationship between text and video, rather than just keywords.

- *Evaluation Findings:*

Evaluations of 24 models revealed that even top-tier models like GPT-4o are not immune. While GPT-4o generally achieves high safety scores and agreement with human judges, it exhibits specific vulnerabilities in categories like S2-Non-Violent Crimes (e.g., cybercrime instructions). Interestingly, Claude 3.7 Sonnet was identified as having the strongest robustness in violent crime categories.The benchmark utilizes RiskJudgeScore, an LLM-based metric that evaluates the probability of toxic output tokens, providing a more granular risk assessment than binary labels.

➢ *Deconstructing Temporal Reasoning: VBenchComp*

A major critique of current VLMs is their reliance on "language priors"—answering questions based on the text prompt without actually "watching" the video. VBenchComp is a pipeline designed to expose this flaw.

- *Classification of Vulnerabilities:*

✓ *LLM-Answerable:*

Questions the model can answer correctly without video input (e.g., "What color is a fire truck?"). High performance here indicates the model is ignoring visual data.

✓ *Semantic:*

Questions answerable even if video frames are shuffled (e.g., "Is there a dog in the video?").

✓ *Temporal:*

Questions that require correct frame ordering (e.g., "Did the person enter or leave the room?").

Research utilizing this pipeline indicates that many models, including early GPT-4 Vision iterations, perform poorly on the "Temporal" category. They struggle with causality, failing to distinguish between a video played forward and one played in reverse—a critical failure point for forensic moderation.

VI. ENGINEERING THE PIPELINE: INGESTION TO INFERENCE

The theoretical capabilities of VLMs are moot without a robust engineering pipeline capable of feeding them data at scale. Video ingestion is notoriously resource-intensive, and the choice of tools defines the system's throughput.

➢ *Video Decoding: The Bottleneck*

Decoding compressed video (e.g., H.264, HEVC) into raw pixel data is the first and often most expensive step.

➢ *Comparative Analysis of Libraries:*

Table 2 Comparative Analysis of Libraries

Library	Mechanism	Pros	Cons	Use Case
OpenCV (cv2)	Sequential Decoding (ffmpeg wrapper)	Ubiquitous, easy to use	High CPU overhead, slow seeking, Python GIL bottlenecks	Prototyping, simple linear playback
FFmpeg (CLI/Python)	Command Line / Pipe	Robust, supports all formats	Overhead of subprocess piping, complex arguments	Batch processing, transcoding
Decord	Hardware Wrapper (NVDEC/Intel)	Smart shuffling, extremely fast random access, GPU decoding	CPU decoding can be slower than optimized OpenCV on some hardware	Deep Learning Data Loaders, Training

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25dec1148

- *Performance Insight:*

Decord is superior for deep learning tasks requiring random access (e.g., sampling frames 1, 50, and 100). OpenCV forces the decoder to seek through the stream sequentially, which is inefficient. For maximum efficiency, pipelines should utilize Decord's GPU decoding to keep data on the VRAM.

➢ *Intelligent Frame Sampling Strategies*

Analyzing every frame of a 30fps video is redundant.

- *Entropy-Guided Motion Enhancement Sampling (EGMESampler):*

This method calculates the spatio-temporal information entropy of motion vectors. Frames with high entropy (significant movement or scene changes) are prioritized. This ensures the VLM focuses on "events" rather than static backgrounds.

- *Semi-Optimal Policy Approach:*

This strategy trains a lightweight "policy network" to estimate the value of each frame for the moderation task, reducing the search space and intelligently selecting the most discriminative frames.

➢ *Navigating Platform Throttling (The yt-dlp Case)*

For researchers and moderators scraping content from public platforms like YouTube or Instagram, 2025 has brought intensified technical countermeasures. The standard tool yt-dlp has faced significant hurdles.

- *Common Throttling Mechanisms:*

✓ *HTTP 403/400 Errors:*

YouTube now aggressively blocks requests that lack specific browser masquerading signatures.

✓ *Bandwidth Throttling:*

IP addresses identified as scrapers are throttled to~1MB/s, making video downloads impractically slow.

✓ *Signature Extraction Failures:*

Changes to the player JavaScript frequently break the "nsig" extraction logic used to decrypt video URLs.

- *Mitigation:*

Successful scraping now requires:

✓ *Variable Delays:*

Implementing random sleep intervals between requests to mimic human behavior.

✓ *Android Player API:*

Configuring yt-dlp to mimic the Android mobile client often bypasses web-client specific restrictions.

✓ *Verbose Debugging:*

Using the -vU flag is essential to identify whether a failure is a network error or a specific API change (e.g., "Precondition check failed").

VII. ECONOMIC AND LATENCY OPTIMIZATION STRATEGIES

The unit economics of moderating video at scale are daunting. A single viral video might be viewed millions of times, but it only needs to be moderated once. However, with millions of *uploads* per day, the inference costs can easily exceed revenue.

➢ *Token Economics and API Optimization*

When using providers like OpenAI, costs are driven by token count.

- *Resolution Management:*

A high-definition image consumes significantly more tokens. Downscaling frames to 512x512 pixels can reduce token usage by 4x while often retaining sufficient detail for safety analysis (e.g., detecting a gun or nudity does not usually require 4K resolution).

- *Batch API:*

For content that does not require real-time clearance (e.g., archival backlog), using the OpenAI Batch API provides a 50% discount. This allows platforms to process vast queues of data during off-peak hours.

➢ *Efficient Serving: The vLLM Revolution*

For self-hosted open-source models (Qwen, InternVL), the inference engine is the key variable. vLLM has emerged as the standard for high-performance serving.

- *Key Optimizations:*

✓ *Paged Attention:*

Inspired by OS virtual memory, this algorithm manages the Key-Value (KV) cache in non-contiguous memory blocks. This eliminates memory fragmentation, allowing for much larger batch sizes.

✓ *Continuous Batching:*

Traditional batching waits for all requests in a batch to finish. Continuous batching inserts new requests immediately as others complete. For video moderation, where request processing times vary wildly (short vs. long videos), this can improve GPU throughput by up to 24x.

✓ *Prefix Caching:*

If the system uses a shared system prompt (e.g., a long definition of "Hate Speech"), vLLM caches the KV states of this prompt, so it doesn't need to be recomputed for every video.

- *Smart Model Routing*

The "Inference Compute-Optimal Frontier" suggests that using a single model for all tasks is inefficient.

✓ *The Routing Architecture:*

A lightweight classifier (e.g., a text-only model analyzing the title/transcript) estimates the "complexity" of the video.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25dec1148

✓ *Tier 1 (Low Risk):*
Content with benign metadata is routed to MiniGPT4-Video or GPT-4o-mini. These models are cheap ($0.15-$0.60 per 1M tokens) and fast.

✓ *Tier 2 (High Risk/Ambiguous):*
Content flagged as potentially violative or culturally complex is routed to GPT-4o or Qwen2.5-VL-72B. While expensive, their deep reasoning capabilities reduce the risk of costly false positives (appeals) or false negatives (PR scandals).

VIII. ETHICAL, LEGAL, AND SOCIETAL FRONTIERS

The deployment of these powerful systems operates within a tightening mesh of ethical and legal constraints.

➢ *The Ethics of Scraping and Data Use*
Constructing datasets like Video-SafetyBench necessitates scraping user content.

- *Regulatory Friction:*
Regulations like GDPR (Europe) and CCPA (California) impose strict limits on processing "personal data." Even public social media posts are considered personal data if the individual is identifiable. Scrapers must now implement robust anonymization pipelines (blurring faces, scrubbing metadata) to remain compliant.

- *Contextual Integrity:*
Researchers face the ethical dilemma of "Contextual Integrity." A user uploading a video to YouTube consents to the platform's Terms of Service, but arguably does not consent to having their content used to train a surveillance system. Ethical guidelines now recommend "attribution with anonymization"—acknowledging the source while protecting identity—though this is technically difficult with video data.

➢ *Bias and Fairness in Automated Adjudication*
Automated moderation systems inevitably inherit the biases of their training data.

- *Demographic Disparities:*
There is a documented risk of VLMs flagging content from marginalized communities as "toxic" due to linguistic or cultural markers (e.g., AAVE, LGBTQ+ vocabulary) that are overrepresented in "hate speech" training sets.

- *Mitigation via CoT:*
Chain-of-Thought prompting has shown promise in mitigating this. By forcing the model to explain *why* it flagged a video, biases often become explicit in the reasoning trace, allowing for easier auditing and correction. Research indicates that CoT can significantly reduce false positives in sensitive categories.

IX. CONCLUSION

The transformation of content moderation from a manual, labor-intensive burden to a sophisticated, AI-driven discipline is well underway in 2025. The integration of Large Vision-Language Models has provided the necessary technological leap to address the scale and complexity of the short-form video era.

➢ *Key Findings:*

- *Model Parity:*
The gap between proprietary models (GPT-4o) and open-source challengers (Qwen2.5-VL, InternVL 2.5) has closed. Open-source models now offer viable, cost-effective alternatives for self-hosted moderation pipelines, particularly when combined with advanced serving engines like vLLM.

- *Frameworks over Models:*
The success of moderation depends less on the raw model and more on the framework—systems like KuaiMod and MonitorVLM that wrap the VLM in legal, temporal, and feedback-driven logic are the future of the field.

- *The Temporal Imperative:*
"Video understanding" systems that rely on static frames are obsolete. True safety requires models that understand causality, sequence, and the "arrow of time," a capability now being rigorously tested by benchmarks like Video-SafetyBench and VBenchComp.

The next major step is Active Reasoning. Future moderation AIs won't just label something as good or bad; they will explain *why* they made that decision and continuously improve themselves based on what they learn. They will become a dynamic, watchful protector of the internet, not just a simple filter. The crucial challenge will always be using this powerful technology responsibly, protecting people's privacy and their right to speak freely.

REFERENCES

[1]. Short Form Video Statistics 2025: 97+ Stats & Insights [Expert Analysis] - Marketing LTB, https://marketingltb.com/blog/statis tics/short-form-video-statistics/
[2]. The State of Short-Form Video in 2025: A Business Guide to Growth - Performance Digital, https://www.performancedigital.co m/the-state-of-short-form-video-in- 2025-a-business-guide-to-growth
[3]. VLM as Policy: Common-Law Content Moderation Framework for Short Video Platform, https://www.researchgate.net/publi cation/390991722_VLM_as_Policy_Common-Law_Content_Moderati on_Framework_for_Short_Video_ Platform
[4]. Temporal-Spatial Redundancy Reduction in Video Sequences: A Motion-Based Entropy-Driven Attention Approach - PubMed Central, https://pmc.ncbi.nlm.nih.gov/article s/PMC12025262/
[5]. GPT-4o Guide: How it Works, Use Cases, Pricing,

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25dec1148

Benchmarks | DataCamp, https://www.datacamp.com/blog/w hat-is-gpt-4o
[6]. GPT-4o vs. Qwen2.5-VL Comparison - SourceForge, https://sourceforge.net/software/co mpare/GPT-4o-vs-Qwen2.5-VL/
[7]. Video-SafetyBench: A Benchmark for Safety Evaluation of Video LVLMs - arXiv, https://arxiv.org/html/2505.11842v3
[8]. API Pricing - OpenAI, https://openai.com/api/pricing/
[9]. Rate limits - OpenAI API, https://platform.openai.com/docs/g uides/rate-limits
[10]. Video Understanding: Qwen2-VL, An Expert Vision-language Model, https://www.edge-ai-vision.com/20 25/03/video-understanding-qwen2- vl-an-expert-vision-language-mode l/
[11]. Best Open Source Multimodal Vision Models in 2025 - Koyeb, https://www.koyeb.com/blog/best- multimodal-vision-models-in-2025
[12]. Multimodal AI: A Guide to Open-Source Vision Language Models - BentoML, https://www.bentoml.com/blog/mult imodal-ai-a-guide-to-open-source- vision-language-models
[13]. InternVL2.5: Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling, https://internvl.github.io/blog/2024- 12-05-InternVL-2.5/
[14]. The Top Challenges of Using LLMs for Content Moderation (and How to Overcome Them), https://www.musubilabs.ai/post/the-top-challenges-of-using-llms-for-c ontent-moderation-and-how-to-ove rcome-them
[15]. Vision-CAIR/MiniGPT4-video: Official code for Goldfish model for long video understanding and MiniGPT4-video for short video understanding - GitHub, https://github.com/Vision-CAIR/Min iGPT4-video
[16]. VLM as Policy: Common-Law Content Moderation Framework for Short Video Platform, https://arxiv.org/html/2504.14904v 1
[17]. Kwai Keye, https://kwai-keye.github.io/
[18]. MonitorVLM: A Vision–Language Framework for Safety Violation Detection in Mining Operations - arXiv, https://arxiv.org/html/2510.03666v 1
[19]. Video-SafetyBench: A Benchmark for Safety Evaluation of Video ..., https://liuxuannan.github.io/Video-SafetyBench.github.io/
[20]. BAAI/Video-SafetyBench · Datasets at Hugging Face, https://huggingface.co/datasets/BA AI/Video-SafetyBench
[21]. Large Language Models for Crash Detection in Video: A Survey of Methods, Datasets, and Challenges - arXiv, https://arxiv.org/html/2507.02074v 1
[22]. Breaking Down Video LLM Benchmarks: Knowledge, Spatial Perception, or True Temporal Understanding? - Apple Machine Learning Research, https://machinelearning.apple.com/research/breaking-down
[23]. Video understanding limitations - Amazon Nova - AWS Documentation, https://docs.aws.amazon.com/nov a/latest/userguide/prompting-visio n-limitations.html
[24]. OpenCV vs FFMPEG Efficiency - Third party integrations - Home Assistant Community, https://community.home-assistant.i o/t/opencv-vs-ffmpeg-efficiency/21 4085