© JAN 2026 | IRE Journals | Volume 9 Issue 7 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I7-1713739

Implementation Architectures and Systematic Survey of Multimodal Vision-Language Models

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*

Abstract - Social media has officially crossed the 5-billion-user mark, and we are on track to hit 6 billion within just a few years. Over 90% of teens and the majority of children now bypass text and photos in favor of fast-paced, short-form video. This shift has put immense pressure on platforms to keep up. When you're dealing with millions of uploads every single day, traditional moderation tools which often look at data in silos just can't keep pace with the nuance and context of modern content. To solve this, the industry is moving toward more sophisticated AI models, like Multimodal Large Language Models (MLLMs). These systems are designed to "think" more like humans by processing video, sound, and text simultaneously to catch harmful content that older models might miss. Implementing a service for filtering YouTube Shorts and Instagram Reels requires a sophisticated architectural design that balances information integrity with computational speed.

I. INTRODUCTION

Traditional video classification models, while effective for well-defined tasks, often struggle with the contextual ambiguity and implicit harmful content inherent in modern social media. Consequently, the industry is gravitating toward Multimodal Large Language Models (MLLMs) and Vision-Language Models (VLMs) that offer superior cross-modal reasoning by integrating visual, auditory, and textual signals. The development and implementation of an AI-driven filtering service for YouTube Shorts and Instagram Reels represent a critical advancement in digital safety. By leveraging the cross-modal reasoning of Multimodal Large Language Models (MLLMs), researchers can overcome the limitations of traditional unimodal classifiers, achieving high-accuracy detection of implicit harms and contextual violations.

Technical Implementation of Multimodal Video Processing Frameworks

Implementing a service for filtering YouTube Shorts and Instagram Reels requires a sophisticated architectural design that balances information integrity with computational speed. A standard industrial-scale framework comprises four primary stages: input preprocessing, modal feature extraction, multimodal fusion, and task output generation.

Input Preprocessing and Decoding

The system initially receives raw video data, which must be decoded into constituent streams. For short-form content, which is typically recorded in a vertical 9:16 aspect ratio with a resolution of 1080x1920 pixels, the preprocessing pipeline must maintain spatial fidelity. Unlike landscape video, vertical formats require specific attention to ensure that critical visual elements often centered or at the extreme top/bottom of the frame are not lost during cropping or padding for 16:9-native encoders. The system separates the data into image frame sequences, speech tracks (audio signals), and embedded text data such as subtitles or on-screen overlays.

Modal Feature Extraction Modules

Each stream is processed by a specialized module to extract high-dimensional features. The image module typically utilizes Deep Convolutional Neural Networks (CNN) or ResNet to extract spatial structural features, focusing on object detection, facial analysis, and scene layout. For audio, speech tracks are characterized using Mel-frequency cepstrum coefficients (MFCC). Temporal modeling of the audio signal is handled through Recurrent Neural Networks (RNN) or LSTMs to track speech rate, tone, and emotional cues. The text module employs Optical Character Recognition (OCR) to capture in-video text, while language models like BERT or RoBERTa generate embeddings for contextual semantic understanding.

© JAN 2026 | IRE Journals | Volume 9 Issue 7 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I7-1713739

Multimodal Fusion Mechanism

The fusion layer is where the model aligns and reasons over the integrated data. Industrial implementations are increasingly adopting attention-based fusion rather than simple feature concatenation. A common mathematical approach is the weighted summation mechanism, where the fused representation is defined as :
$$F_{fusion} = \alpha_{v}F_{v} + \alpha_{a}F_{a} + \alpha_{t}F_{t}$$

Component Breakdown

- F_{fusion}: The resulting fused feature vector or final output.
- $\alpha_{v}, \alpha_{a}, \alpha_{t}$: The weighting coefficients (scalars) for the visual, audio, and textual modalities, respectively. These determine the "importance" or contribution of each source.
- F_{v}: The feature set or data representing the Visual modality.
- F_{a}: The feature set or data representing the Audio modality.
- F_{t}: The feature set or data representing the Textual modality.

Cascade Architectures for Industrial-Scale Deployment

One of the most significant barriers to the deployment of MLLMs in content filtering is the extreme computational cost. Full-scale inference on every video uploaded to a platform like Instagram is economically unfeasible. To mitigate this, practitioners utilize a "Router-Ranking" or "Filter-And-Refine" cascade architecture.

The Router-Ranking Paradigm

In this two-stage system, a lightweight router serves as a first-stage filter. It selectively passes only high-risk content to the high-capacity MLLM (the ranker), while safe content is dismissed with minimal processing. In many industrial implementations, the router is implemented as an embedding retrieval system that maintains a "seed bank" of known violating videos. New uploads are compared against these "golden seeds" using semantic similarity; only those with a high similarity score are routed to the ranker for fine-grained classification.

Metric	Direct MLLM Deployment	Router-Ranking Cascade	Improvement
Computational Cost	100% (Baseline)	1.5%	66.7x Reduction
Auto-Moderation Volume	Baseline	+41%	Significant Scale-up
F1 Score Improvement	Baseline	+66.5%	Over traditional classifiers
Fine-tuning Data Needed	100%	2%	50x Resource Saving

The "Filter-And-Refine" system deployed on major platforms has demonstrated that the router can eliminate up to 97.5% of the total traffic flow, allowing the costly MLLM to focus exclusively on the small subset of potentially violating content. This not only reduces costs but also significantly decreases the serving latency for the vast majority of users.

Efficiency Optimization: Token Compression and Hardware Acceleration

The speed of an MLLM-based filtering service is largely dictated by the number of vision tokens processed by the language model backbone. Standard models like LLaVA-v1.5 use 576 vision tokens, which can create substantial bottlenecks in real-time filtering.

The LLaVA-Mini Architecture

To achieve extreme efficiency, the LLaVA-Mini architecture introduces "modality pre-fusion." This technique is based on the discovery that vision tokens are most critical in the early layers of the LLM, where they primarily serve to fuse visual data into the text embeddings. By performing this fusion in advance, LLaVA-Mini can compress the visual representation fed to the main LLM layers into a single token. Efficiency analyses reveal that LLaVA-

Mini can reduce FLOPs by 77% and lower GPU memory usage from 360 MB per image to just 0.6 MB. This allows the model to process long-form videos exceeding 10,000 frames on consumer-grade hardware like the NVIDIA RTX 3090.

Hardware Benchmarks: A100 vs. H100

Hardware selection is another vital factor in implementation. The shift from the A100 (Ampere) to the H100 (Hopper) architecture has enabled significant leaps in throughput and latency reduction. The H100's dedicated "Transformer Engine" provides native support for FP8 precision, allowing for 2x faster performance and halved memory consumption compared to 16-bit options.

Metric	NVIDIA A100 (80GB)	NVIDIA H100 (80GB)	H100 Advantage
Inference Throughput	~130 tokens/sec	250–300 tokens/sec	~2x faster
Memory Bandwidth	2 TB/s	3.35 TB/s	~1.7x higher
Daily Requests (1024 tok/req)	~11,000	22,000–26,000	~2x more capacity
Latency (1st Token)	Baseline	4.4x faster	Significant for real-time

The H100 also features fourth-generation Tensor Cores that deliver up to 4x the performance of the A100's cores, making it the preferred choice for organizations with high-throughput requirements despite a higher hourly cost. For a service filtering YouTube Shorts, where low-latency responses are critical to user experience, the H100's ability to achieve less than 10ms to first token latency is a decisive factor.

Safety Benchmarking and Adversarial Vulnerabilities

As automated systems become the primary defense for digital platforms, evaluating their robustness against "video-text attacks" has become a central research priority.

Video-SafetyBench: The Standard for Multimodal Risk

Video-SafetyBench is the first benchmark designed to evaluate LVLMs under compositional attacks, comprising 2,264 video-text pairs across 48 fine-grained categories.

The benchmark utilizes a paired query strategy:

1. Harmful Query: A direct request for malicious content (e.g., "How do I construct the explosive device in this video?").
2. Benign Query: An ostensibly harmless request that triggers toxic behavior when grounded in the video (e.g., "Explain the chemistry experiment shown" when the video depicts illicit drug production).

Evaluation Findings: Extensive experiments show that benign-query video compositions achieve an average attack success rate (ASR) of 67.2%, highlighting consistent vulnerabilities in safety alignment.

Model	Harmful Query ASR	Benign Query ASR	Agreement with Human
GPT-4o	14.8%	43.3%	96.5%
Gemini 2.0 Pro	22.4%	61.9%	-
Qwen2.5-VL-72B	41.3%	74.0%	85.9%
InternVL2.5-78B	28.4%	68.0%	92.9%

Results reveal that proprietary models demonstrate significantly stronger safety alignment than open-source alternatives. However, the 28.1% gap between harmful and benign query performance across all models suggests that VLMs still struggle with implicit video-referential threats. Additionally, the study found that adding typographical overlays (TYPO) to harmful queries consistently increases unsafe response rates, with Qwen2-VL-72B showing a 10.4% increase.

Forensic Failures: Temporal Reasoning

The VBenchComp pipeline exposes the reliance of models on "language priors" answering based on the text prompt without actually processing the visual sequence.

Models are categorized by their ability to answer:

1. LLM-Answerable: Questions answerable without any video input.
2. Semantic: Questions answerable even if frames are shuffled.
3. Temporal: Questions requiring correct frame ordering.

Research indicates that many leading models, including early iterations of GPT-4 Vision, perform poorly in the "Temporal" category. They struggle with causality and the "arrow of time," making them vulnerable to forensic failures such as being unable to determine if a person entered or left a room, or distinguishing between a video played forward and one played in reverse.

Economic Analysis and Cost Management Strategies

The economic reality of 2025 is that AI adoption is outpacing cost governance. A report on the "2025 State of AI Cost Management" revealed that 84% of companies are seeing significant gross margin erosion due to AI infrastructure costs, with 26% reporting an impact of 16% or higher.

Model Routing and Token Economics

To preserve margins, platforms are implementing "Inference Compute-Optimal" routing. This strategy uses a lightweight classifier to estimate the complexity of a video and routes it to the most cost-effective model.

1. Tier 1 (Triage): Content with benign metadata is routed to models like GPT-4o-mini or Qwen2.5-VL-7B. These models are inexpensive ($0.15/1M tokens) and can handle the vast majority of benign content.
2. Tier 2 (Adjudication): Flagged or culturally complex content is escalated to GPT-4o or Qwen2.5-VL-72B ($5.00/1M tokens) for deep reasoning and rationale generation.
3. Selective Reasoning: Research into selective reasoning indicates that for straightforward queries, deep reasoning is unnecessary. Implementing an adaptive reasoning strategy can reduce token consumption by 48.5% and response latency by 47.1%.

Trustworthy Moderation through Policy-Aligned Reasoning

A critical challenge in automated content moderation is the "black-box" nature of traditional classifiers, which often fail to provide human-interpretable reasons for their decisions. To address this, current research is moving toward "Hierarchical Guard" (Hi-Guard) frameworks that align decisions with explicit platform policies.

The Hi-Guard Framework and Hierarchical Taxonomy

The Hi-Guard framework integrates category-level policy definitions directly into the model's prompt. This allows the model to reason over rules at inference time, ensuring that the final decision is a result of logical deduction based on current standards. The system utilizes a multi-level taxonomy to categorize risks:

1. Domain: The broad category of harm (e.g., Harassment).
2. Topic: The specific area within the domain (e.g., Cyberbullying).
3. Subtype: The type of behavior observed (e.g., Targeted insults).
4. Behavior: The granular action being flagged.

To optimize this process, researchers use Group Relative Policy Optimization (GRPO), a reinforcement learning strategy that penalizes misclassifications based on their semantic distance from the correct label. This approach significantly improves the model's ability to distinguish between subtle violations, such as "inappropriate attire" versus "suggestive imagery," which traditional models often conflate.

Benchmarking and Evaluation Standards in Social Media Analysis

© JAN 2026 | IRE Journals | Volume 9 Issue 7 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I7-1713739

The evaluation of filtering services requires benchmarks that accurately mirror the complexity of social media content. Traditional benchmarks for image captioning or object detection are insufficient for detecting the nuanced "social understanding" required for moderation.

The MM-Soc and YouTube2M Datasets

The MM-Soc benchmark was developed specifically to assess MLLMs' proficiency in understanding human emotions, humor, sarcasm, and misinformation in online spaces. A standout feature of this benchmark is the "YouTube2M" dataset, which includes 2 million YouTube videos that were shared on Reddit. This targeted selection process ensures that the dataset reflects the viral potential and cultural context of specific communities, making it an ideal testing ground for content moderation.

Benchmark Dataset	Primary Tasks	Modality Focus	Significance
YouTube2M	Tagging, categorization, text generation.	Video, Audio, Metadata.	Captures viral & community-specific trends.
Memotion	Sarcasm, humor, offensive detection.	Image, Embedded Text.	Focuses on meme culture and implicit sentiment.
Hateful Memes	Hate speech detection.	Multimodal (Text + Image).	Requires joint understanding to identify hate.
FakeNewsNet	Misinformation detection.	Text, Image, User Metadata.	Essential for filtering deceptive "news" content.

Extensive evaluations on MM-Soc have shown that while zero-shot models often struggle with these complex social tasks, their performance can be dramatically improved through specific fine-tuning strategies and policy alignment.

Ethical, Legal, and Privacy Considerations for Filtering Services

Operating a filtering service for social media requires a profound understanding of global privacy regulations, notably the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). Compliance in 2024 and 2025 is no longer just a legal hurdle but a core component of user trust and application sustainability.

Data Scraping and the Ninth Circuit Precedent

The legality of web scraping for research and moderation depends heavily on whether the data is publicly accessible. The landmark hiQ Labs v. LinkedIn case established that scraping publicly available information does not violate the Computer Fraud and Abuse Act (CFAA). However, this precedent does not grant absolute freedom; scraping personal data without consent even if it appears publicly can still violate privacy laws like the GDPR and CCPA.

Compliance Principles: Minimization and Transparency

Adhering to GDPR requires a lawful basis for processing, such as "legitimate interest," but this interest must be demonstrated to outweigh the individual's privacy rights. Key principles include:

- Data Minimization: Collecting only the specific data necessary for the filtering task.
- Purpose Limitation: Ensuring data is not repurposed for unrelated profiling or marketing.
- Transparency: Informing data subjects about how their information is processed, which presents practical challenges for large-scale scrapers.

Furthermore, the EU AI Act and the Digital Services Act (DSA) have introduced stricter protections for minors. Content filtering services must ensure that their algorithms do not employ deceptive designs or

"dark patterns" that manipulate users' privacy choices. For services targeting younger demographics, compliance with COPPA (Children's Online Privacy Protection Act) is mandatory, requiring verifiable parental consent before collecting information from children under 13.

Survey of 2024-2025 Social Media Trends and Their Impact on Moderation

The field of content filtering is continuously shaped by emerging trends in social media marketing and user behavior. In 2024 and 2025, the rise of "Augmented Reality (AR)" and "Ephemeral Content" has introduced new vectors for potentially harmful material.

AR Filters and Ephemeral Content

Brands and creators are increasingly using AR filters on Instagram and Snapchat to drive engagement. These filters, while creative, can be used to mask identity or distribute deceptive content. Filtering services must now incorporate AR-detection modules that can "see through" digital overlays. Similarly, ephemeral content posts that disappear after 24 hours requires high-velocity filtering systems that can flag violations before the content expires.

AI-Curated Discovery and Intent Modeling

Social media algorithms have shifted toward "intent modeling," where the system predicts not just what a user likes, but what they will engage with next. Instagram's 2025 ranking rules, for instance, prioritize original audio and short-form Reels over static posts, with a heavy emphasis on "save" and "share" metrics over simple "likes". For a filtering service, this means that moderation must happen faster than the algorithm can amplify a post. The integration of "Sentiment Analysis" and "Social Listening" is becoming indispensable, as it allows platforms to identify shifts in audience emotions and proactively filter content that might trigger harmful viral trends.

The technical success of moderation systems is inextricably linked to public trust. However, global studies in 2025 show a "tension between benefits and risks".

Trust Metric	2024 Result (%)	2025 Result (%)	Trend
Belief AI is Beneficial	52%	55%	Increasing
Willingness to Trust AI Systems	50%	46%	Decreasing
Requirement for AI Regulation	55.7% (2022)	73.7%	Increasing
Full Trust in AI Fairness	N/A	2%	Critically Low

Only 2% of U.S. adults "fully" trust AI's capability to make unbiased decisions, while 60% express some level of distrust. This skepticism is fueling a public mandate for regulation, with 70% of people believing that national and international regulation is necessary.

II. CONCLUSION AND FUTURE INTEGRATION

By 2025, content moderation has clearly moved far beyond being a slow, manual task handled by large teams of reviewers. It has evolved into a mature, AI-driven discipline, largely enabled by the adoption of large vision–language models that can finally keep up with the scale and complexity of short-form video. These models represent a major technological shift, making it possible to analyze massive volumes of visual content with far greater speed and consistency than ever before.

Key implementation findings demonstrate that:

1. Frameworks define success: Raw model capability is secondary to the framework.

Systems like KuaiMod and MonitorVLM, which wrap VLMs in "common-law" reasoning or "clause-specific" filters, provide significantly higher industrial accuracy.

2. Decoding is the bottleneck: The transition from OpenCV to hardware-accelerated libraries like Decord is essential for pipeline throughput.
3. Safety requires temporal reasoning: Models that rely on static frame analysis or language priors are insufficient for forensic safety. The field must prioritize models capable of understanding causality and the sequence of events, as rigorously tested by benchmarks like Video-SafetyBench.

The future of video content moderation lies in "Active Reasoning," where systems don't just filter content but explain their decisions and continuously adapt to evolving cultural norms. As platforms navigate the economic challenges of inference, the intelligent application of smart routing and serving engines like vLLM will remain the cornerstone of a safe and scalable digital internet.

WORK CITED

- marketingltb.com
  Short Form Video Statistics 2025: 97+ Stats & Insights [Expert Analysis] - Marketing LTB
- vidico.com
  20+ Interesting Short Form Video Trends & Statistics (2025) - Vidico
- conectys.com
  AI Content Moderation Trends for 2026 | Blog - Conectys
- newsroom.tiktok.com
  Digital Services Act: Our fifth transparency report on content moderation in Europe
- clippie.ai
  Why Short-Form Video Continues to Dominate in 2026 - Clippie AI
- medrxiv.org
  The Impact of Short-Form Video Use on Cognitive and Mental Health Outcomes: A Systematic Review | medRxiv
- reutersinstitute.politics.ox.ac.uk
  Overview and key findings of the 2025 Digital News Report - Reuters Institute
- businessresearchinsights.com
  Short-form video Market 2026-2035 | Growth at CAGR 30.33% - Business Research Insights
- researchnester.com
  Content Moderation Services Market Size, Trends & Forecast to 2035 - Research Nester
- researchandmarkets.com
  Social Media Moderation Global Market Insights 2026, Analysis and Forecast to 2031
- deloitte.com
  Deloitte 2026 Technology, Media & Telecommunications Predictions - Press Release
- chatpaper.com
  VLM as Policy: Common-Law Content Moderation Framework for Short Video Platform
- sourceforge.net
  GPT-4o vs. Qwen2.5-VL Comparison - SourceForge
- llm-stats.com
  AI Leaderboards 2026 - Compare LLM, TTS, STT, Video, Image & Embedding Models
- llm-stats.com
  VideoMMMU Leaderboard - LLM Stats
- frontiersin.org
  AI-driven disinformation: policy recommendations for democratic resilience - Frontiers
- internvl.github.io
  InternVL2.5: Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling
- llm-stats.com
  Video-MME Leaderboard - LLM Stats
- deeplearning.ai
  Alibaba Debuts Qwen2.5-VL, A Powerful Family of Open Vision-Language Models
- debuggercafe.com
  Qwen2.5-VL: Architecture, Data, Benchmarks and Inference - DebuggerCafe
- github.com
  Vision-CAIR/MiniGPT4-video: Official code for Goldfish model for long video understanding and MiniGPT4-video for short video understanding - GitHub
- arxiv.org
  [2502.13923] Qwen2.5-VL Technical Report - arXiv

- arxiv.org
  Time Blindness: Why Video-Language Models Can't See What Humans Can? - arXiv
- arxiv.org
  Video-SafetyBench: A Benchmark for Safety Evaluation of Video LVLMs - arXiv
- openreview.net
  Video-SafetyBench: A Benchmark for Safety Evaluation of Video LVLMs | OpenReview
- arxiv.org
  Video-SafetyBench: A Benchmark for Safety Evaluation of Video LVLMs - arXiv
- www2.eecs.berkeley.edu
  vLLM: An Efficient Inference Engine for Large Language Models by Woosuk Kwon - UC Berkeley EECS
- catalyzex.com
- Shisong Tang - CatalyzeX
- kuaimod.github.io
  KuaiMod: VLM-based SVP Moderator
- kwai-keye.github.io
  Kwai Keye
- kanerika.com
  LLM vs vLLM in 2026: Best Practices for AI Teams & Deployments - Kanerika
- itecsonline.com
  vLLM vs Ollama vs llama.cpp vs TGI vs TensorRT-LLM: 2025 Guide | Blog // ITECS
- mavvrik.ai
  2025 State of AI Cost Management Research Finds 85% of Companies miss AI forecasts by >10% - Mavvrik
- arxiv.org
  When to Reason: Semantic Router for vLLM - arXiv
- llm-stats.com
  GPT-4o mini vs Qwen2.5 VL 32B Instruct - LLM Stats
- mofo.com
  Privacy + Data Security Predictions for 2025 - Morrison Foerster
- kpmg.com
  Trust, attitudes and use of artificial intelligence: A global study 2025 - KPMG International
- hai.stanford.edu
  Public Opinion | The 2025 AI Index Report | Stanford HAI
- news.gallup.com
  Americans Prioritize AI Safety and Data Security - Gallup News
- committees.parliament.uk
  TikTok fails to share evidence behind increased AI use in content moderation - Committees
- creatorhandbook.net
  YouTube addresses AI moderation concerns after reporting 12 million channel terminations in 2025 - Creator Handbook
- ppc.land
  YouTube CEO defends AI moderation as creators lose channels overnight - PPC Land
- youtube.com
  YouTube addresses AI moderation concerns after reporting 12 million channel terminations in 2025
- techgdpr.com
  Data protection digest 3-17 July 2025: AI-generated voice and visuals' potential to violate people's rights and freedoms - TechGDPR
- arxiv.org
  VisualPRM: An Effective Process Reward Model for Multimodal Reasoning - arXiv
- internvl.github.io
  VisualPRM: An Effective Process Reward Model for Multimodal Reasoning - InternVL
- researchgate.net
  VisualPRM: An Effective Process Reward Model for Multimodal Reasoning - ResearchGate
- emergentmind.com
  Visual Process Bench: Multimodal Reasoning Benchmark - Emergent Mind