UNIVID: Unified Vision-Language Model for Video Moderation

Kejuan Yang*, Yizhuo Zhang*, Mingyuan Du, Yue Zhang,
Dixin Zheng, Kaili Zhao, Yang Xiao, Hanzhong Liang, Kenan Xiao

Bytedance

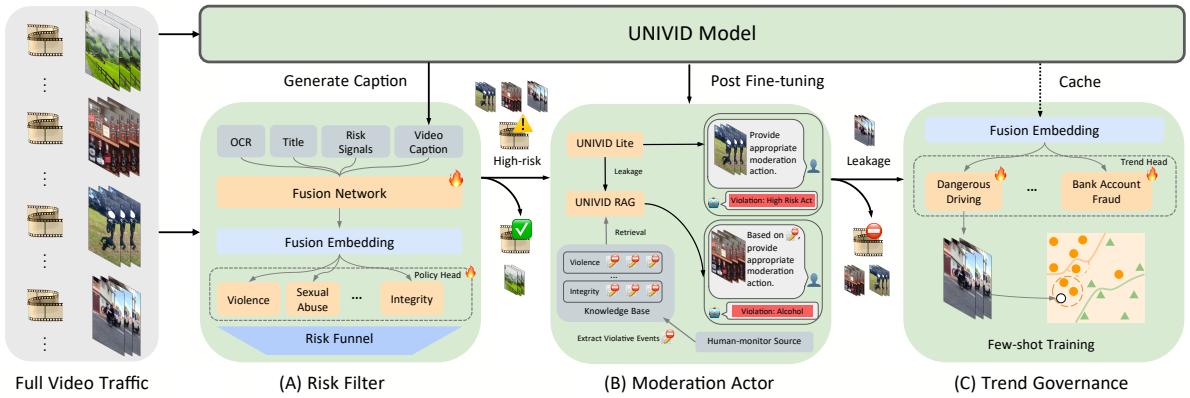

Figure 1: Our UNIVID-centric video moderation pipeline includes three cascaded stages: (A) Risk Filter acts as a multi-modal risk funnel that fuses UNIVID caption to filter potential high-risk videos; (B) Moderation Actor employs two finetuned downstream models, UNIVID-Lite and UNIVID-RAG, to predict moderation decisions and recall leakage based on prior violative events; (C) Trend Governance module utilizes cached UNIVID embeddings to adaptively detect emerging risks via tuning specific trend head.

Abstract

Global-scale video moderation faces a dual challenge: the need for fine-grained multi-modal reasoning and the demand for interpretable outputs to support downstream enforcement. Traditional moderation systems often rely on fragmented black-box classifiers that are difficult to maintain and lack transparency. In this paper, we present UNIVID, a ***UNI**fied **VI**sion-language model for video mo**D**eration*. Unlike standard classification models, UNIVID generates policy-aware captions that serve as an interpretable intermediate representation, enabling human-verifiable decisions and multi-task reusability. While existing open-source and commercial VLMs often suffer from safety-guardrail refusals and lack fine-grained policy alignment, we develop a specialized training data recipe that combines expert human-refined labels with synthetic data to align the model with our safety guidelines. By integrating UNIVID as the core captioner, we design a novel end-to-end video moderation system that reduces violation leakage by 42.7% and overkill rate by 37.0% relatively. Meanwhile, by replacing over 1,000 policy-specific models with a single UNIVID backbone, we recycled extensive computation resources while reducing engineering maintenance overhead. To our knowledge, this is one of the first reports of a high-efficiency captioning VLM successfully supporting industrial-scale moderation and cross-functional business.

1 Introduction

The rapid growth of short-video platforms such as Reels and YouTube Shorts necessitates accurate and operationally efficient moderation systems (Wang et al., 2025b; Liang et al., 2025). Prior systems rely on thousands of specialized, end-to-end classification models (Shi et al., 2024), each targeting a specific policy (e.g., violence, regulated activities). However, this approach faces three bottlenecks: (1) Lack of Interpretability, as black-box scores offer little to no rationales for human auditors (Levi et al., 2025; Bao et al., 2025); (2) Tedious Maintenance, where policy updates require retraining thousands of individual models (Liang et al., 2025); and (3) Resource Inefficiency,

*Equal contribution.
Correspondence: kendra.kejuanyang@bytedance.com

as these models lack the semantic flexibility to support cross-functional business such as advertisement and recommendation safety.

Vision-Language Models (VLMs) offer a promising alternative by transforming video into natural language descriptions, i.e., video captions. **Why using a Captioning VLM for moderation?** First, captions act as a unified, human-readable bridge that provides explicit evidence for policy violations (Huang et al., 2025). Second, a single VLM can replace thousands of specific classification models (Wang et al., 2026), simplifying our system infrastructure. Finally, these policy-aware captions can be cached and reused by downstream tasks, creating a multi-purpose content understanding feature (Wang et al., 2025a).

Despite their potential, existing open-source (Liu et al., 2023, 2024b; Li et al., 2024a) and commercial VLMs (Comanici et al., 2025; OpenAI, 2023) often fall short of industrial moderation requirements. They frequently refuse to describe sensitive or violative content (Bao et al., 2025; Lee et al., 2025) due to internal safety triggers. Furthermore, since these models are not grounded using our internal platform policies, they often fail to capture the precise enforcement boundaries, leading to inaccurate descriptions. Finally, the scale of commercial VLMs makes real-time inference for full-traffic moderation economically impractical.

To address these limitations, we develop UNIVID, which adopts the same architectural design as LLaVA-OV (Li et al., 2024a) while employing a task-specific training recipe for moderation. We use a hybrid training strategy that combines expert annotations with high-quality synthetic data, aligning the model with our detailed moderation policies. On top of UNIVID, we construct a multi-stage moderation system: a high-throughput Risk Filter leveraging UNIVID embeddings for early screening; a Moderation Actor deploying two finetuned variants of UNIVID to support moderation decisions; and a Trend Governance module that reuses cached UNIVID captions to detect emerging risks.

Our UNIVID-centric moderation system has been fully integrated into our platform, yielding the following contributions:

- **Unified System**: We propose the industry deployment of a unified video moderation infrastructure built on UNIVID, significantly reducing engineering overhead and simplifying end-to-end troubleshooting.

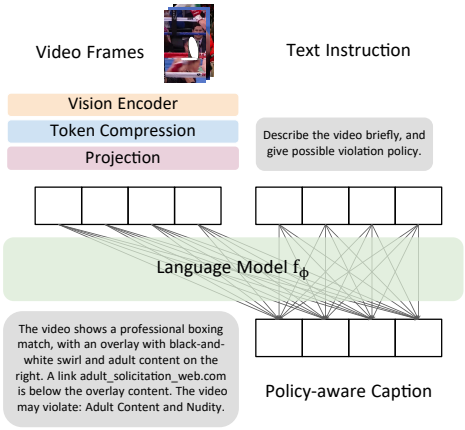

Figure 2: Model architecture of UNIVID following LLaVA-OneVision (Li et al., 2024a). We construct in-house data recipe focusing on safety violation content.

- **Data & Evaluation Pipeline**: We design a human-in-the-loop training recipe to ensure factual and policy alignment. Furthermore, we introduce *CapBench* for evaluation, which decomposes captions into atomic events to evaluate violation recall across key safety domains.
- **Platform Safety Governance**: By integrating UNIVID as the core captioner, our new moderation system reduces violation leakage by 42.7% and overkill rate by 37.0% relatively. Beyond that, UNIVID also achieves 81% matching accuracy on the beta simulation of Brand & Ads applications.

2 Method

2.1 Model Architecture

Our UNIVID model is built upon the LLaVA family of multimodal large language models (Liu et al., 2023, 2024a,b; Li et al., 2024a), with a particular focus on the LLaVA-OneVision (Li et al., 2024a) architecture, which has demonstrated strong scalability and flexibility for video content understanding. We choose Mistral-v0.3-7B (Jiang et al., 2023) as the LLM f_ϕ, as its permissive license ensures production compliance and its parameter scale is optimized for full-traffic deployment. We follow LLaVA-OneVision (Li et al., 2024a) for design choices on components shown in Figure 2.

To be more specific, for a video input V, and the associated caption C, we first extract frames uniformly from the video, providing a series of video frames $X_{1...N}$, where N is the maximum supported frames for our model. After appropriate pre-processing, the image is fed into vision encoder g, token compression module TC, and projection

Table 1: Overview of UNIVID training stages.

Stage	Data Source	Type	# Samples (M)
PT	LLaVA pretrain, our internal data	Single sentence caption	1.6
FT	Synthetic and human-refined caption	Caption	3.2
	Synthetic VQA	General VQA (e.g., summary, topic, keywords, etc.)	2.0
CFT	Hybrid caption	Caption	0.1

W. This process eventually convert visual information into tokens H_v, which is then concatenated with encoded text instructions H_t. For a caption C with length L we aim to maximize the following:

$$P(C \mid T_{\text{in}}, V_{\text{in}}) = \prod_{i=1}^{L} P(C_i \mid H_{\text{t}}, H_{\text{v}}, C_{<i})$$

2.2 Data Recipe

Constructing a large-scale training set for video moderation poses several challenges. Firstly, collecting real-world violative video is difficult and legally sensitive. Therefore, no open-source dataset is available. Moreover, we cannot completely rely on proprietary or open-source VLMs, which are safety-aligned and frequently refuse to generate sensitive labels, e.g., CSAM[1].

To address these limitations, we develop an in-house dataset tailored to the specific content styles and moderation requirements of our platform (see Table 1). Following the standard LLaVA instruction-tuning recipe (Liu et al., 2024b), we leverage GPT-4o (OpenAI, 2023) to generate one-paragraph captions and VQA pairs[2]. The VQA pairs encompass both general visual understanding and moderation-specific queries. Examples of training data are shown in Appendix A.1.

To optimize the trade-off between annotation costs and data quality, we internally train annotators to perform human-in-the-loop refinement of the raw GPT-4o outputs. The refinement includes two dimensions: (1) **Factual Correction**: Rectifying hallucinations and adding missing details regarding subjects, objects, actions, backgrounds, and OCR. (2) **Policy Grounding**: Ensuring that violation rationales are grounded in our internal safety policy playbook. For violative content, annotators are required to map the video to the corresponding pre-defined policy title.

Furthermore, to better capture edge cases in harmful trends, we use proprietary VLMs to normalize and enrich the violation rationale description. This step specifically targets multilingual OCR and adversarial visual hacking techniques (e.g., split-screen effect).

2.3 Training details

Following the LLaVA setup, our model training paradigm includes three stages: (1) Pretraining for modality alignment, (2) Instruction-tuning on caption and VQA data, (3) Continue finetuning on high-quality caption data to further enhance the model's generation ability (see Table 1).

In the pretraining stage, we freeze the whole model except the MLP projector to align the vision and text modality. In the next two finetuning stages, we train both the projector and LLM decoder. The model training procedure takes 120 hours on 32 H100 GPUs. We also develop UNIVID-1B as a lightweight variant by replacing the decoder with our proprietary LLM.

2.4 Video Moderation System

UNIVID functions as the core of our moderation system across following three stages.

Risk Filter Our risk filter module performs early-stage screening over full video traffic to identify potential policy violations while maintaining low latency and high throughput. We first generate video captions using UNIVID, augmented by auxiliary signals including OCR, titles, and risk embeddings predicted by lightweight policy-specific modules. Our fusion network integrates these multimodal inputs into a shared embedding, which is connected to multiple MLP policy heads. Each policy head applies a calibrated decision threshold to identify high-risk videos and route them to downstream stages.

Moderation Actor Our earlier moderation system suffered from precision dilution and high maintenance overhead. This is due to the reliance on multiple independent components (e.g., neural scoring models, vector-based retrievers, and heuristic rules) operating in parallel. Moreover, such a fragmented system complicates the threshold calibration and makes end-to-end troubleshooting tedious.

[1] CSAM: Child Sexual Abuse Material

[2] GPT-4o was the state-of-the-art VLM in multimodal understanding when we designed the system.

Table 2: Comparison of violative video captioning benchmarks.

Benchmark	# Total (Vio.)	Human-verified	Global Source	Uni-scoring	Credibility
Dream-1k (Wang et al., 2024)	1000 (0)	✓	✗	✗	✓
KuaiMod (Lu et al., 2025)	1000 (422)	✓	✗	✗	✗
CapBench (Ours)	17210 (11476)	✓	✓	✓	✓

To address these limitations, we transition to a recall-and-rank architecture composed of two models: UNIVID-Lite for primary enforcement and UNIVID-RAG for leakage mitigation.

UNIVID-Lite UNIVID-Lite serves as a unified actor model, providing a single decision layer for the moderation pipeline. It is finetuned on 1 million in-house videos based on the UNIVID-1B backbone. The training data consists of human-annotated moderation examples balanced at a positive-to-negative ratio of 1:5 to reflect the natural distribution of production traffic. Each sample includes video frames together with auxiliary textual signals which are concatenated directly into the instruction prompt (see Appendix D.2). UNIVID-Lite is trained with an autoregressive generation objective: rather than using a separate classification head, the model is prompted to produce a structured natural language output indicating the moderation decision (Approve or Violation) along with the corresponding violated policy. This formulation allows the model to inherit the general reasoning capabilities of the pretrained UNIVID backbone while specializing for binary enforcement.

UNIVID-RAG To mitigate leakage — violations missed by the primary actor — we introduce UNIVID-RAG, which augments the moderation decision with retrieval from a Violation Knowledge Base (VKB). The VKB contains approximately 100,000 structured violative events derived from past cases labeled by human moderators. When new violations are identified, Gemini-2.5-Pro converts the raw annotations into structured violative events capturing the policy title and violation rationale, which are then added to the VKB. The knowledge base is updated continuously to reflect the latest labeling.

At inference time, UNIVID-RAG retrieves the top-3 most semantically similar violative events from the VKB using cosine similarity over UNIVID caption embeddings. The retrieved events are inserted directly into the prompt as in-context examples, providing the model with concrete prior cases to reason against (see Appendix D.3). This retrieval-augmented approach is specifically designed to improve coverage of hard or low-frequency violations that the primary model may miss, at the cost of a moderately higher violation rate due to increased sensitivity.

Trend Governance Short-form video platforms are increasingly targeted by emerging trends, such as dangerous "hot water pouring challenges". To identify these real-time issues, we implement a lightweight adaptation strategy. By reusing cached fusion embeddings from the backbone, we train an MLP trend head to recognize these shared semantic videos via few-shot samples (< 50). This approach allows our moderation system to remain agile, capturing short-lived or rapidly evolving threats.

2.5 Evaluation

Existing video safety benchmarks are often limited to surveillance contexts (Sultani et al., 2018; Hassner et al., 2012) or synthetic video sources (Liu et al., 2025), failing to capture the broad spectrum of user-generated content. While the recent KuaiMod (Lu et al., 2025) benchmark addresses some of these gaps, it is localized to Chinese content and employs compliance-driven masking of key visual components, such as human faces. Such anonymization brings a distribution shift that deviates from the high-fidelity and unmasked real-world video posts.

To evaluate multimodal understanding across both general and harmful contexts, we design our in-house benchmark **CapBench**, which assesses VLMs on two dimensions: descriptive accuracy and violation recall. Inspired by the prior Dream-1k (Wang et al., 2024) benchmark, we employ a fine-grained metric that decomposes video captions into atomic "evidence." CapBench includes two subsets: **violative** and **healthy** sets. The violative subset is further divided into five domains: *Violence, Sexual Abuse, Mental Health, Regulated Activity,* and *Integrity*. Details are shown in Appendix A.3.

For each video, we establish a human-refined ground-truth caption, which is then segmented by Gemini-2.5-Pro into verifiable events. We then

Table 3: Evaluation results on CapBench.

	Violative Set							Healthy Set		
	Violence	Sex Abuse	Mental Health	Regulated Act	Integrity	Vio Rec.	Non-vio Rec.	Rec.	Prec.	F1
Proprietary Models										
GPT-4.1	45.9	17.4	32.3	42.5	57.6	36.1	32.8	31.0	65.5	37.4
Gemini-2.5-Pro	63.8	44.3	55.6	57.6	67.5	55.1	42.5	44.9	95.2	57.9
LLaVA-OV 8B	17.8	6.9	12.9	15.7	14.2	13.0	12.0	12.7	**86.3**	19.3
UNIVID-7B	**56.3**	**51.3**	**50.2**	**57.7**	**50.1**	**54.3**	**32.4**	**28.9**	82.3	**39.1**
UNIVID-1B	53.6	49.1	49.9	55.3	47.7	52.1	30.6	27.4	82.8	37.5
Ablation Study										
w/o Hybrid Data	37.9	35.5	33.5	41.1	30.4	37.5	18.4	15.8	81.8	23.1
w/o Human Data	29.1	22.9	18.9	29.4	23.9	26.1	15.8	15.8	82.2	23.0

prompt Gemini as judge to compute two symmetric metrics based on these segments: (1) Recall: Measures the coverage of ground-truth events within the predicted caption. (2) Precision: Measures the rate of hallucination or the proportion of predicted events supported by the ground truth. For violative samples, events serving as violation rationales are annotated with violation labels (e.g., "*The three men ride together on a single red motorcycle.*"), enabling separate evaluation of violative and non-violative recall. Compared to previous work in Table 2, our benchmark offers clearer insights into caption credibility and provides a unified scoring framework applicable to both violative and healthy video captioning.

3 Results

This section introduces our offline evaluation results and the simulated online impact of integrating UNIVID into our video moderation system.

3.1 Offline Experiments

We validate the captioning and violation recognition ability of UNIVID on our internal CapBench.

Experiment Settings We evaluate the leading commercial VLMs, specifically GPT-4o (OpenAI, 2023) and Gemini-2.5-Pro (Comanici et al., 2025) to explore the upper bound performace. Additionally, we benchmark LLaVA-OneVision-1.5 8B (An et al., 2025), which shares a similar architecture with UNIVID but has not been finetuned on our in-house moderation dataset. The model generation outputs are constrained to 150 words to align with our online deployment, as shown in Appendix D.

Analysis As demonstrated in Table 3, UNIVID-7B substantially outperforms LLaVA-OV 8B, despite their similar architectures. Notably, UNIVID-7B also achieves higher violation recall across all domains compared to GPT-4.1 (OpenAI, 2023). This gap is partially due to the safety guardrails of proprietary models, which refuse generation in certain cases (12.9%), yielding empty captions.

Table 4: Comparison of VLM deployment cost (per 1M videos). Note that we are unable to deploy LLaVA-OV due to compliance constraints.

Metric	GPT-4.1	Gemini-2.5-Pro	LLaVA-OV	UNIVID
Vio Rec.	36.1	55.1	13.0	54.3
Cost ($)	4830	3444	N/A	180

We further conduct ablation studies upon training data recipe. We find that removing hybrid training data leads to consistent violation recall degradation across all domains (-16.8% average). We observe a more pronounced decline when training with human-annotated data only, where violation recall further decreases to 26.1%, indicating limited generalization under narrowly distributed supervision using GPT-4o. Note that other ablation results on model architecture are presented in Appendix B.1.

3.2 Sandbox Moderation Experiments

UNIVID has delivered substantial improvements in moderation accuracy on production-scale traffic. Based on our simulated evaluation results, our new system achieved a 42.7% relative reduction in view-weighted violation leakage, falling from 0.255% to 0.146%. Simultaneously, the overkill rate decreased from 35.4% to 22.3%. Meanwhile, UNIVID replaces over 1000 policy-specific models with a single multimodal backbone, with 1900 A30 GPUs recycled. Consequently, our system reduces engineering overhead and simplifies long-term maintenance.

Deployment To support full video traffic, we deploy UNIVID model with FP8 quantization on H100 GPUs, reaching 5.7 QPS per device. As shown in

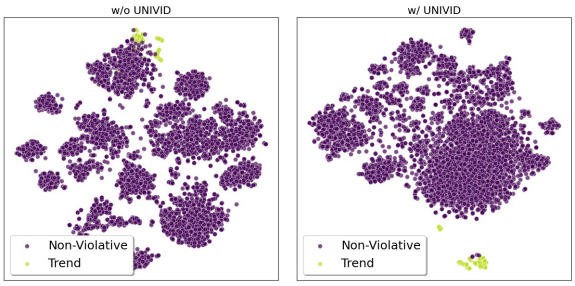

Figure 3: Visualization of trend detector embeddings.

Table 4, our production inference cost is approximately $180 per 1M videos, achieving a 15× reduction compared to commercial VLMs (> $3,000).

3.3 Downstream Modules

This section presents our simulated evaluation results for downstream systems that integrate UNIVID, including our video moderation main system and cross-functional applications.

Risk Filter As shown in Table 5, integrating UNIVID embeddings consistently improves violation recall (especially +23.9% on high-view leakage), highlighting enhanced sensitivity to rare but high-risk content.

After launching the Risk Filter with UNIVID embedding, our main moderation system observes an 11.5% increase in violation hit volume. Meanwhile, the moderation precision improves substantially, with violation precision from 35.9% to 75.6%, demonstrating better coverage and decision quality in live traffic.

Moderation Actor Our simulated moderation ablation in Table 6 demonstrates the effectiveness of UNIVID-Lite as a unified action layer in live moderation traffic. Compared to the recall-only pipeline, UNIVID-Lite reduces the violation rate (1.38% → 1.34%) while improving precision (+9.4%) and leakage recall (+11.8%), indicating more accurate enforcement with fewer unnecessary actions.

Introducing retrieval augmentation further increases leakage recall to 53.6%, improving coverage of hard or low-frequency violations. This gain comes at the cost of slightly reduced precision and a slightly higher violation rate, reflecting the trade-off between maximizing recall and enforcing conservatively. We further present our case study in Appendix B.1. Overall, UNIVID-Lite offers strong precision-oriented actions, with RAG serving as a complement when higher recall is prioritized.

Table 5: Simulated online ablation of UNIVID embedding on the risk filter and few-shot trend detector. Scores are reported at 65% precision threshold.

Embeddings	Vio Rec.	Leak. Rec.	Trend Rec.
w/o UNIVID	72.3	33.3	56.7
w/ UNIVID	**78.2**	**59.8**	**86.7**

Table 6: Simulated online ablation results of UNIVID-Lite and UNIVID-RAG on the moderation actor module.

Actor	Vio Rate	Vio Prec.	Leak. Rec.
Recall	1.38	76.0	39.3
UNIVID-Lite	1.34	**85.4**	51.1
UNIVID-Lite + RAG	1.48	78.3	**53.6**

Trend Governance We report the simulated moderation results of few-shot learning in Table 5. On the school dangerous behavior trend, incorporating UNIVID embeddings substantially improves moderation performance, boosting violation recall from 56.7% to 86.7%. The embedding visualization in Figure 3 further shows that UNIVID embeddings produce clearer separation between trend and non-violative cases. This suggests that UNIVID provides semantically rich representations that generalize in low-data regimes, enabling rapid adaptation to emerging trends without degrading precision.

Cross-functional Business Beyond moderation, UNIVID also generates fine-grained video keywords that are consumed by the advertising ranking system as semantic targeting signals, enabling brand-safe matching and custom-lineup construction. This integration achieved 81% accuracy in the Brand & Ads applications in our simulation tests.

4 Related Work

4.1 Video Moderation Systems

Video moderation is crucial for protecting users from detrimental content and maintaining a healthy platform, especially for minors (Gorwa et al., 2020; Gongane et al., 2022; Udupa et al., 2023; Lai et al., 2022). Plenty of efforts have been made to construct a video moderation ecosystem. Early work uses traditional ML methods on handcrafted features for hate speech, toxicity, or fake news detection (Gongane et al., 2022; Naseeb et al., 2025), typically in binary or multi-class setups. Similar patterns appear in visual moderation, where porn or anomaly detection in images/videos uses engineered visual features and classical classifiers (Wang et al., 2023; Yousaf and Nawaz, 2022).

4.2 MLMs for Content Moderation

Multimodal language models (MLMs) have demonstrated impressive capabilities in understanding images (Hudson and Manning, 2019; Goyal et al., 2017; Marino et al., 2019) and short videos (Caba Heilbron et al., 2015; Patraucean et al., 2023; Li et al., 2024b; Maaz et al., 2024). These models generally comprise three modules: a vision encoder, a large language model (LLM) decoder, and an adapter that aligns image-text modalities (Li et al., 2023; Zhu et al., 2023; Maaz et al., 2024; Lin et al., 2023; Liu et al., 2024b). Prior work explores MLM-based moderation models (Lu et al., 2025) and data generation frameworks validated on open-source VLMs (Wang et al., 2025c). These approaches focus on offline evaluation, without considering real-world deployment at production scale, while UNIVID is trained under our compliance regulations and operates in full-traffic production.

5 Conclusion

In this paper, we present UNIVID, a unified vision-language model that reframes video moderation from fragmented black-box classifiers to an interpretable, caption-driven framework. Built upon UNIVID, we construct a unified moderation pipeline comprising a high-throughput Risk Filter, a Moderation Actor, and a Trend Governance module. This UNIVID-based system has been deployed on our global-scale short-video platform, operating over full video traffic under strict compliance constraints. In our simulated production experiments, it achieves a 42.7% relative reduction in violation leakage and a 37.0% reduction in overkill rate, while attributing 81% accuracy on the Advertisement downstream business.

Ethics Considerations

The deployment of UNIVID for industrial-scale video moderation involves handling highly sensitive content, necessitating rigorous ethical and operational safeguards. First, we prioritize annotator welfare by enforcing strict daily exposure limits to violative content. Our Standard Operating Procedure (SOP) includes regular rotation shifts to prevent prolonged psychological strain and ensure mental well-being. For particularly sensitive categories like Child Sexual Abuse Material (CSAM), we implement dedicated protocols where access is restricted to a minimal, specially trained group of annotators operating under enhanced security and psychological support frameworks.

Regarding data privacy and retention, our system adheres to strict internal compliance standards. To minimize data footprint, deleted and private videos are permanently purged from our systems within a strictly defined retention window, while non-violative video data is retained only for a limited period. All data used for training and evaluation is stored in siloed environments with internal access controls. Any identified illegal content is handled through our internal reporting channels without permanent retention in any environment.

To prevent dual-use and adversarial risks, we have established a strict non-disclosure policy for our technical assets. We will not release the source code, model weights, or the specific training datasets used in this paper. Instead, we provide an overview of the system architecture and training objectives without offering a direct template for adversarial attacks. We are committed to ensuring that the specific enforcement logic remains protected against exploitation by bad actors.

Finally, we address bias and algorithmic fairness through our data recipe and model design. During the annotation process, our guidelines explicitly instruct annotators to avoid using or emphasizing terms related to specific ethnic groups to prevent the encoding of societal prejudices. Furthermore, we are committed to the continuous improvement of the model's multilingual capabilities, ensuring that safety enforcement is equitable and consistent across different linguistic and cultural contexts.

Limitations

Currently, UNIVID does not incorporate reinforcement learning methods such as Group Relative Policy Optimization (GRPO). The policy-aware captions can serve as a form of reasoning trace indicating how a video may violate specific policies; however, platform policy guidelines are not directly encoded as reward signals and therefore are not explicitly bound to this reasoning process.

Our system processes videos via frame sampling rather than full temporal modeling. As a result, videos that embed policy-violating content in only a single frame may evade detection if those frames are not selected. Therefore, the effectiveness of our system depends on keyframe selection quality, which remains a practical limitation under latency and computational constraints.

References

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, and 3 others. 2025. Llava-onevision-1.5: Fully open framework for democratized multimodal training. In *arXiv*.

Han Bao, Qinying Wang, Zhi Chen, Qingming Li, Xuhong Zhang, Changjiang Li, Zonghui Wang, Shouling Ji, and Wenzhi Chen. 2025. Vmoda: An effective framework for adaptive nsfw image moderation. *arXiv preprint arXiv:2505.23386*.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Vaishali U. Gongane, M. Munot, and A. Anuse. 2022. Detection and moderation of detrimental content on social media platforms: current status and future directions. *Social Network Analysis and Mining*, 12.

Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Tal Hassner, Yossi Itcher, and Orit Kliper-Gross. 2012. Violent flows: Real-time detection of violent crowd behavior. In *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 1–6. IEEE.

Guolei Huang, Qinzhi Peng, Gan Xu, Yuxuan Lu, and Yongjun Shen. 2025. Llavashield: Safeguarding multimodal multi-turn dialogues in vision-language models. *Preprint*, arXiv:2509.25896.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

Vivian Lai, Samuel Carton, Rajat Bhatnagar, Vera Liao, Yunfeng Zhang, Chenhao Tan, and Q. Liao. 2022. Human-ai collaboration via conditional delegation: A case study of content moderation. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*.

DongGeon Lee, Joonwon Jang, Jihae Jeong, and Hwanjo Yu. 2025. Are vision-language models safe in the wild? a meme-based benchmark study. *arXiv preprint arXiv:2505.15389*.

Adi Levi, Or Levi, Sardhendu Mishra, and Jonathan Morra. 2025. Ai vs. human moderators: A comparative evaluation of multimodal llms in content moderation for brand safety. *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 6024–6032.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Hanzhong Liang, Jinghao Shi, Xiang Shen, Zixuan Wang, Vera Wen, Ardalan Mehrani, Zhiqian Chen, Yifan Wu, and Zhixin Zhang. 2025. Embedding-based retrieval in multi-modal content moderation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4264–4268.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Xuannan Liu, Zekun Li, Zheqi He, Peipei Li, Shuhan Xia, Xing Cui, Huaibo Huang, Xi Yang, and Ran He. 2025. Video-safetybench: A benchmark for safety evaluation of video lvlms. *arXiv preprint arXiv:2505.11842*.

Xingyu Lu, Tianke Zhang, Chang Meng, Xiaobei Wang, Jinpeng Wang, Yi-Fan Zhang, Shisong Tang, Changyi Liu, Haojie Ding, Kaiyu Jiang, and 1 others. 2025. Vlm as policy: Common-law content moderation framework for short video platform. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 4682–4693.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Amna Naseeb, Muhammad Zain, Nisar Hussain, Amna Qasim, Fiaz Ahmad, Grigori Sidorov, and A. Gelbukh. 2025. Machine learning- and deep learning-based multi-model system for hate speech detection on facebook. *Algorithms*.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, and 1 others. 2023. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36:42748–42761.

Jinghao Shi, Xiang Shen, Kaili Zhao, Xuedong Wang, Vera Wen, Zixuan Wang, Yifan Wu, and Zhixin Zhang. 2024. Cpfd: Confidence-aware privileged feature distillation for short video classification. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4866–4873.

Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *Preprint*, arXiv:2303.15389.

Sahana Udupa, Antonis Maronikolakis, and Axel Wisiorek. 2023. Ethical scaling for content moderation: Extreme speech and the (in)significance of artificial intelligence. *Big Data & Society*, 10.

Hexu Wang, Wenlong Luo, Wei Wu, Fei Xie, Jindong Liu, Jing Li, and Shizhou Zhang. 2025a. Vips: Learning-view-invariant feature for person search. *Sensors*, 25(17):5362.

Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. 2024. Tarsier: Recipes for training and evaluating large video description models. *arXiv preprint arXiv:2407.00634*.

Mo Wang, Kaixuan Ren, Pratik Jalan, Ahmed Ashraf, Tuong Vy Vu, Rahul Seetharaman, Shah Nawaz, and Usman Naseem. 2026. From native memes to global moderation: Cros-cultural evaluation of vision-language models for hateful meme detection. *arXiv preprint arXiv:2602.07497*.

Wenxuan Wang, Jingyuan Huang, Chang Chen, Jiazhen Gu, Jianping Zhang, Weibin Wu, Pinjia He, and Michael R. Lyu. 2023. Validating multimedia content moderation software via semantic fusion. *Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis*.

Zixuan Wang, Jinghao Shi, Hanzhong Liang, Xiang Shen, Vera Wen, Zhiqian Chen, Yifan Wu, Zhixin Zhang, and Hongyu Xiong. 2025b. Filter-and-refine: A mllm based cascade system for industrial-scale video content moderation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 873–880.

Zixuan Wang, Yu Sun, Hongwei Wang, Baoyu Jing, Xiang Shen, Xin Luna Dong, Zhuolin Hao, Hongyu Xiong, and Yang Song. 2025c. Reasoning-enhanced domain-adaptive pretraining of multimodal large language models for short video content governance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1104–1112.

Kanwal Yousaf and Tabassam Nawaz. 2022. A deep learning-based approach for inappropriate content detection and classification of youtube videos. *IEEE Access*, 10:16283–16298.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A UNIVID

A.1 Training Data

A series of training data examples are provided in Figure 4.

A.2 Training Details

We report our training hyperparameters in Table 10.

A.3 CapBench

We show the detailed statistics of CapBench in Table 7. Note that since a video can violate multiple policies simultaneously, the sum of policy-specific counts exceeds the total number of unique samples.

The evaluation prompts for both stages are presented below, including event extraction and entailment judgment.

Event Extraction Prompt

```
You are given a description of a video clip: [caption]
Your tasks:
1. Extract at most 10 key events from the description.
2. From these events, identify those that may violate the video platform content policy labels: [policy_list]. Only include events that can serve as sufficient evidence of the listed policy violations. You will be given a document of policy instructions to help you better understand the policy definition (if any).

Requirements for Events:
- An event must include an action or motion.
- Merge semantically similar and disjunctive actions into a single event. DO NOT repeat same events.
- Every event is represented by a brief sentence within 20 words, with a subject with key attributes, a predicate and optionally an object. Include the action or attributes (such as minor's age, revealing outfit) that may violate the content policy!
- Every event must be atomic, meaning that it cannot be further split into multiple events.
- Scene cuts and camera movements are NOT events! - Substitute pronouns by the nouns they refer to.

Output Format:
Return a Python dictionary string with the following keys:
- "events": List[str]
the list of all extracted events.
- "violative_events": Dict[str, List[str]]
mapping from each policy name in policy_list to the list of events that violate it. If no events violate a policy, use an empty list.

Materials:
Policy Instructions: [policy_doc]
```

Entailemnet Judgement Prompt

```
You are given a video description and a list of events.
Your task is to classify the relationship between the video description and each event into one of three classes: - entailment: The video description entails the event.
- contradiction: Some detail in the video description contradicts the event.
- neutral: The relationship is neither entailment nor contradiction.

Video Description: [caption]
Events:[events]

Output Format:
Return a JSON list in the following format:
{"event": "copy an event here", "relationship": "put class name here"}

Requirements: - Classify every event independently.
- Use only one of the three predefined class names.
- Do NOT provide any additional text,
```

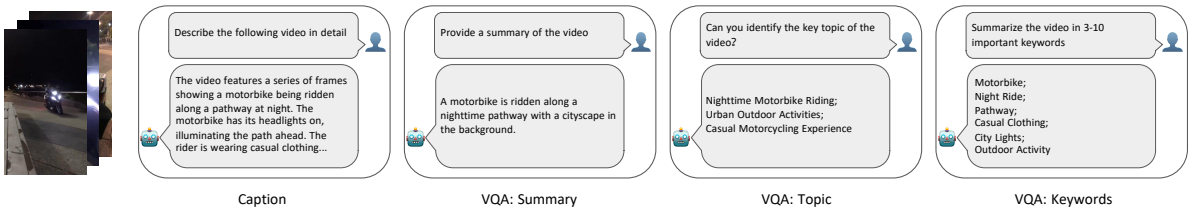

Figure 4: Examples of UNIVID training data. We introduce four different tasks: Caption, Summary, Topic, and Keywords.

Table 7: Distribution of safety domains and policies in CapBench.

Domain	Policy Category	# Samples
Violence	Violent Behaviors; Shocking & Graphic Content	2,700
Sexual Abuse	Exploitation & Abuse; Nudity & Sexual Activity	4,339
Mental Health	Mental Health; Harassment & Hateful Behavior	1,248
Regulated Activity	High-Risk & Regulated Activities	4,656
Integrity	Harmful Misinformation; Deceptive Behaviors	624

Figure 5: CapBench example.

```
explanation, or formatting.
- Output only the JSON.
```

A.4 Ablation Study

As shown in Table 9, we also examine the impact of visual and resolution components. Intuitively, replacing the vision encoder with less performant EvaCLIP (Sun et al., 2023) results in a decrease in violation recall (-9.5%). For resolution setting, given the on-par performance but substantially higher token cost, we choose not to use AnyRes in our model design.

B UNIVID-Lite

As shown in Appendix D.2, we fuse video frames with auxiliary textual features as model input, including titles, user profiles, OCR, and ASR, to provide enriched semantic context.

B.1 Case Study

We find that UNIVID-Lite exhibits a robust capacity for contextual inference, enabling the detection of latent risk signals that extend beyond explicit keywords. In High-Risk Driving scenarios, the model identifies danger by synthesizing environmental cues (such as dashboard speedometers) even in the absence of overt reckless actions. Similarly, in Severe Bullying cases, UNIVID-Lite effectively perceives hostile intent conveyed through suggestive or indirect language rather than explicit insults. Furthermore, in the domain of Serious Harm, the model recognizes subtle precursors to danger, successfully identifying implicit threats that often precede manifest harmful behavior.

C UNIVID-RAG

We show the example violative events in Table 8.

D Prompts

D.1 Evaluation

The captioner prompt for Capbench is as follows:

User Prompt

```
Given image frames uniformly sampled from a video clip, describe the video (not the individual images) in detail, focusing on the main subjects, their actions, and the background scene. You should also pay
```

Table 8: Examples of violative events from the knowledge base for UNIVID-RAG.

Policy	**Violative Events**
Nudity & Sexual Activity	The video shows two youths performing a dance that includes sexually suggestive movements, such as hip thrusting, squatting, and sticking their tongues out.
High-Risk & Regulated Activities	The video shows multiple people riding on the exterior of moving vehicles, including on the sides and tops of trucks within a convoy.
Mental Health	The video contains a real-life clip of a person standing on a train platform as a train passes, followed by animated scenes depicting bloodstains on the platform.
Harassment & Hateful Behavior	The video displays a person's full email address (xxx@gmail.com) and a verification code on a smartphone screen.

Table 9: Ablation results of UNIVID-7B on CapBench.

	Violative Set					Healthy Set			
Model	Violence	Sex Abuse	Mental Health	Regulated Act	Integrity	Non-vio Rec.	Rec.	Prec.	F1
UNIVID-7B	56.3	51.3	50.2	57.7	50.1	32.4	28.9	82.3	39.1
w/ EvaCLIP	46.6	44.2	39.4	46.9	35.8	24.3	21.6	77.1	30.0
w/ AnyRes	56.7	51.4	48.9	58.4	50.6	32.5	28.7	82.4	38.7

Table 10: UNIVID training hyper-parameters

Hyperparameter	**Value**
# Epochs	2
Per-device Batch Size	8
Gradient Accumulation	2
Learning Rate	1e-5
LR Scheduler	Cosine
Warmup Ratio	0.03

```
attention to details that may violate
the trust-and-safety platform content
policy. Don't describe feelings or
atmosphere.
Your output should be a single
coherent paragraph. Maximum length:
150 words.
```

D.2 UNIVID-Lite Inference

The online inference prompt of UNIVID-Lite is as follows:

User Prompt

```
Region: [region]
Title: [title]
User nickname: [nickname]
ASR Text: [asr]
OCR Text: [ocr]
Bio Text: [profie]

Based on the video frames and related
content above, please indicate
the severity of any inappropriate,
disruptive, or harmful content.
```

D.3 UNIVID-RAG Inference

The online inference prompt of UNIVID-RAG is as follows:

User Prompt

```
Region: [region]
Title: [title]
User nickname: [nickname]
ASR Text: [asr]
OCR Text: [ocr]
Bio Text: [profie]
```

Violative Events for Reference:
1. [policy title], violation reason: [reason]
2. [policy title], violation reason: [reason]
3. [policy title], violation reason: [reason]

Based on the video frames, the above content, and the judgment logic of the similar cases, please indicate the severity of any inappropriate, disruptive, or harmful content.