Filter-And-Refine: A MLLM Based Cascade System for Industrial-Scale Video Content Moderation

Zixuan Wang*, **Jinghao Shi***
Hanzhong Liang, **Xiang Shen**, **Vera Wen**, **Zhiqian Chen**, **Yifan Wu**
Zhixin Zhang, **Hongyu Xiong**
TikTok
{zixuan.wang1, jinghao.shi}@tiktok.com

Abstract

Effective content moderation is essential for video platforms to safeguard user experience and uphold community standards. While traditional video classification models effectively handle well-defined moderation tasks, they struggle with complicated scenarios such as implicit harmful content and contextual ambiguity. Multimodal large language models (MLLMs) offer a promising solution to these limitations with their superior cross-modal reasoning and contextual understanding. However, two key challenges hinder their industrial adoption. First, the high computational cost of MLLMs makes full-scale deployment impractical. Second, adapting generative models for discriminative classification remains an open research problem. In this paper, we first introduce an efficient method to transform a generative MLLM into a multimodal classifier using minimal discriminative training data. To enable industry-scale deployment, we then propose a router-ranking cascade system that integrates MLLMs with a lightweight router model. Offline experiments demonstrate that our MLLM-based approach improves F1 score by 66.50% over traditional classifiers while requiring only 2% of the fine-tuning data. Online evaluations show that our system increases automatic content moderation volume by 41%, while the cascading deployment reduces computational cost to only 1.5% of direct full-scale deployment.

1 Introduction

The rapid expansion of short video platforms such as YouTube Shorts and Instagram Reels has transformed online content consumption. As user engagement and content volume continue to grow massively, effective content moderation has become more and more important.

Content moderation generally falls into two categories: human moderation and machine-driven auto-moderation. While human moderation provides good judgment, it is inherently slow, expensive, and difficult to scale. As a result, machine learning (ML)-based auto-moderation has become crucial, offering scalable and efficient solutions for content moderation.

Currently, video content moderation is mostly handled by video classification models (Shi et al., 2024), which process video inputs and tag videos based on a predefined taxonomy. While traditional video classification models effectively handle well-defined moderation tasks, they struggle with more complicated and context-dependent moderation challenges. For instance, they can reliably flag explicit harmful content but often fail to recognize implicit violations, such as subtle forms of misinformation or suggestive imagery. Multimodal Large Language Models (MLLMs) can be a promising alternative due to their superior reasoning and contextual understanding capabilities.

Despite the potential of MLLMs in content moderation, two key challenges make their industrial deployment difficult. First, the high computational cost of large-scale MLLMs poses a big barrier for real industry deployment. To enable scalable deployment, we introduce a router-ranking cascade system. Inspired by recall-ranking architectures commonly used in recommendation systems, our approach employs a lightweight router as a first-stage filter. The router selectively passes only high-risk content, allowing the MLLM to focus on a small subset of potentially violating videos. The cascade design greatly reduces computational costs compared to direct full-scale deployment.

Second, as generative models, MLLMs are not inherently suited for discriminative classification tasks. Effectively converting a generative model into a classifier remains an open research problem. Some prior works (Chen et al., 2024; Mitra et al., 2025) have explored innovative approaches to this transformation, yet no existing study has specifi-

*Equal contribution.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 873–880
July 28-30, 2025

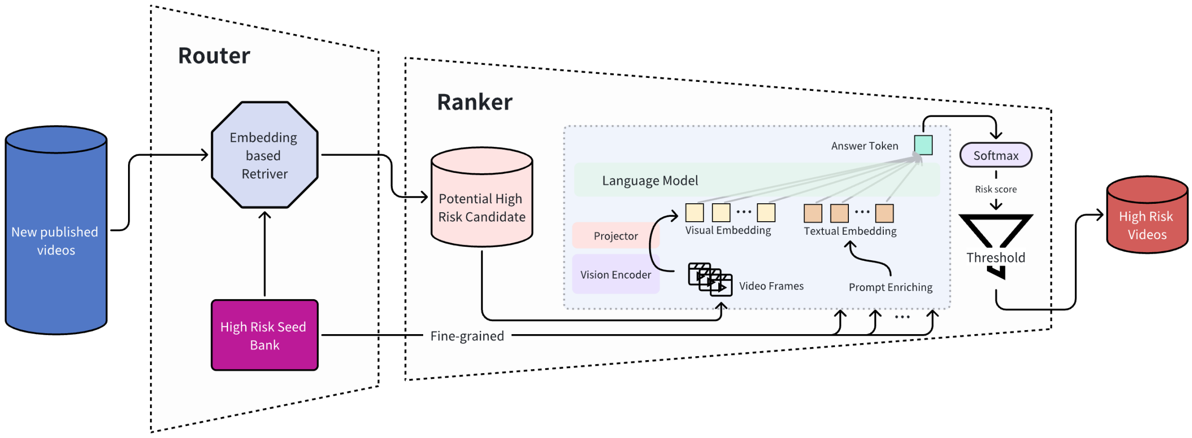


Figure 1: Overview of the cascade system design. The system consists of two stages: a Router and a Ranker. The Router filters and selects potentially high-risk content, while the Ranker performs fine-grained classification to refine the final decision.

cally focused on adapting generative MLLMs for content moderation. In this paper, we address this gap with a straightforward yet effective transformation method that requires only minimal fine-tuning data while demonstrating strong performance in real-world content moderation applications.

We summarize our contribution as follows:

- To the best of our knowledge, this work represents one of the first successful applications of MLLMs in a large-scale content moderation system.
- We introduce a novel router-ranking cascade architecture that enables full-traffic deployment while significantly reducing computational costs.
- We propose a straightforward yet effective method to adapt generative MLLMs for classification tasks, requiring only minimal fine-tuning data.
- We validate the approach through comprehensive offline and online experiments on production data and enable deployment of the model in a real-world production environment.

2 Related Work

2.1 ML-based Content Moderation

As social media platforms continue to expand, efficient content moderation becomes increasingly critical. Over the past few years, significant strides have been made in identifying harmful content such as hate speech (Das et al., 2023), explicit material (notAI.tech, 2024), and toxic language (Yuan et al., 2024) across multiple modalities. Given that social media content naturally integrates video, images, and text, multimodal frameworks, for example (Yuan et al., 2024; Binh et al., 2022), have become a standard approach. Despite some relying on user feedback(Yu et al., 2025), relatively few studies (Ye et al., 2023; Mullick et al., 2023) focus solely on moderating images or text. With the rapid advancement of multimodal large language models (MLLMs), these techniques are increasingly being applied to content moderation, demonstrating strong performance (Ma et al., 2023; Wu et al., 2024).

2.2 MLLM and Supervised Fine-tuning

Although Multimodal Large Language Models, such as LLaVA series (Liu et al., 2023, 2024b), GPT-4 (OpenAI, 2024) and DeepSeek series (DeepSeek-AI, 2025b,a), have shown versatility across diverse tasks, fine-tuning remains essential to achieve optimal performance for specific applications. InstructGPT (Ouyang et al., 2022) has demonstrated that with the help of human feedback, fine-tuning LLMs using reinforcement learning from human feedback (Stiennon et al. (2020); Christiano et al. (2017)) is able to outperform larger models. Furthermore, there are other parameter-efficient ways to leverage multimodal data, such as PEFT(Zhou et al., 2024) and FedMLLM(Xu et al., 2025). The composition and quantity of data also significantly affect the capabilities of LLMs. Dong et al. (2024); Pareja et al. (2024); Pang et al. (2024)

highlight the need for strategic data selection and stages in the fine-tuning process to balance and optimize various model capabilities.
With the nature of generative models, MLLM does not demonstrate a strong capability in multimodal classification (Zhang et al., 2024). Chen et al. (2025); Liu et al. (2024a) explores the application in anomaly detection with different prompt formats.

3 Cascade System Design

Deploying Multimodal Large Language Models (MLLMs) at an industrial scale presents computational challenges, particularly for high-traffic platforms, where hundreds of millions of new videos are uploaded daily. Directly applying MLLMs to full traffic is prohibitively expensive and inefficient, which makes a scalable and resource-efficient moderation pipeline important. Inspired by recall-ranking architectures in recommendation systems, we introduce a two-stage router-ranking cascade system in Figure 1 to optimize moderation efficiency. This framework includes:

`Lightweight Router (Recall Stage).` A computationally efficient model acts as a first-stage filter, quickly identifying suspicious content while discarding low-risk videos.

`MLLM-Based Ranker (Ranking Stage).` The more powerful yet costly MLLM then analyzes only the high-risk subset, performing fine-grained reasoning to accurately detect harmful content.
This hierarchical filtering approach significantly reduces unnecessary MLLM processing, improving scalability while preserving high moderation accuracy on the real-time video platform.

3.1 Router

The router model serves as the first-stage filter in our cascade system(Liang et al., 2025). It can be implemented using any feasible architecture, such as classification models or embedding-based retrieval systems.
In our implementation, we leverage an embedding retrieval system as the router due to its effectiveness and efficiency. This system operates by maintaining a pre-selected bank of high-risk representative videos, called seed videos. The newly published videos are then filtered based on semantic similarity with the seed videos to pick high-risk candidates.
We designed several strategies to ensure high-quality seed selection, such as Centroid-Proximity Seed Selection, which uses clustering algorithms to identify good seeds, and Manual Seed Selection, which relies on annotators to identify "golden seeds". Our retrieval-based router offers several key advantages: Unlike classification models, our approach does not require labeled data and is trained in an unsupervised manner. The seed bank architecture offers the system rapid adaptation and great flexibility. By efficiently filtering content before MLLM processing, our router significantly reduces computational costs while maintaining high recall for potentially violating videos.

3.2 Ranker

The MLLM serves as the ranker, refining the Router's output by predicting a more precise moderation decision. It takes both the extracted visual features from the video and a task-specific prompt corresponding to the target class. The model outputs a single token representing the predicted label and token probabilities as the confidence score. Unlike conventional classifiers with fixed output structures, MLLMs offer greater flexibility through prompt engineering, enabling adaptation to various moderation tasks without retraining. Their advanced reasoning and contextual understanding further enhance ranking performance, allowing the model to act as a strong refiner in the cascade system. Additionally, the extensive pretraining on open-domain knowledge provides a strong initialization for the ranking stage. For details on the MLLM-based ranker, refer to the next section.

4 Finetune MLLM as Ranker

In this section, we first introduce the multimodal large language model (MLLM) architecture. We then describe our continuous supervised fine-tuning process, covering the construction of the fine-tuning dataset and two fine-tuning strategies explored to optimize the model's performance. Next, we outline how the model's output is calibrated into probabilistic scores for online serving. Finally, we discuss further improvements such as prompt engineering and result ensembling. All together, they enable the generative MLLM to function effectively as a discriminative ranker within our system.

4.1 Model Backbone

We adopt LLaVA (Liu et al., 2024b) as the MLLM architecture, leveraging its strong performance and flexibility. It consists of three main components:

Models	Prompt	PR-AUC	ROC-AUC	Max-F1
Multi-Modal Classification (X-VLM)	-	30.79	65.31	36.81
LLaVA	-	23.17	58.59	31.32
LLaVA w/ Caption	-	28.85	65.88	36.71
Mixed Sequential Phased Learning	P1	66.96	87.01	60.64
	P2	<u>68.10</u>	<u>87.47</u>	<u>60.98</u>
	P3	62.43	84.90	57.06
	P4	66.97	87.05	60.51
Multi-task Learning	P1	66.33	86.90	59.94
	P2	**68.73**	**87.68**	**61.29**
	P3	65.11	86.05	58.54
	P4	67.60	87.32	60.84

Table 1: Performance results (%) across models with different training strategies and prompt designs. The top section presents results from traditional multimodal classification models. The middle section includes two zero-shot models: the first is the original LLaVA model, while the second is further fine-tuned on a captioning task. The bottom section reports results from different models fine-tuned on the classification dataset.

`LLM (Large Language Model)`: We use Mistral-7B(Jiang et al., 2023), chosen for its compatibility with industry-serving environments.
`Vision Encoder`: We employ ViT-Large, which provides robust visual feature extraction.
`Projector`: A two-layer MLP is used to align vision and language representations.
The training process begins with Mistral-7B, pre-trained by the LLaVA team, as the initialization. During fine-tuning, we follow standard next-token prediction for captioning and VQA datasets. Given a sentence that is segmented as a sequence of tokens $x = (x_1, x_2, ..., x_n)$, where x_i belongs to V, which is the vocabulary dictionary. The joint probability of the sequence x is modeled as:

$$p(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} P(x_i | x_1, x_2, ..., x_{i-1})$$

While for the finetuning on classification dataset, the task reduces to single-token prediction, where only one token represents the final classification label. The extraction of the predicted token probability is elaborated in Section 4.4.

4.2 Training Dataset

The training dataset consists of three parts:
VQA Dataset. A randomly subsampled dataset from LLaVA-Mix665k (Liu et al., 2024b) that is used for fine-tuning. It includes COCO, GQA, OCR-VQA, TextVQA, and VG, providing a strong foundation for visual comprehension and question-answering capabilities.
Video Caption Dataset. A high-quality caption dataset designed to provide rich contextual summaries of videos. Captions cover key aspects such as subjects, attributes, actions, and scenes. For inappropriate videos, the captions highlight potential violations based on these aspects.
Classification Dataset. This dataset is customized for moderation tasks, with each video labeled with a fine-grained issue tag and an overall label indicating whether or not action should be taken. We selected representative moderation issues and sampled the dataset according to the online traffic distribution. The dataset exactly aligns with the online data distribution after the Router.
In total, the dataset contains 300k samples, with a 1:1:1 ratio across the three subsets.

4.3 Training Strategy

We explored two different Supervised Fine-Tuning (SFT) strategies as mentioned in the paper (Dong et al., 2024). Let D_1, D_2, and D_3 represent the three datasets used in training.
Multi-task learning. Directly mix different fine-tuning data sources $D = \cup_{1 \leq i \leq 3} D_i$ and then train on the mixed dataset. For multi-task learning, the overall training procedure is about 20 hours using 8×A100 GPUs.
Mixed Sequential phased learning. The first stage is Visual Instruction Tuning. We first mix D_1 and D_2 and train to get the best epoch. Then, in the second stage, called Moderation-Oriented Supervised Fine-Tuning. We fine-tune on D_3 specifically for Moderation. For sequential phased training, the first phase of sequential training is about 10 hours, and the continuous training is about 10.5 hours using 8×A100 GPUs.

4.4 Transform Model Output

To make the model's output fit for actual online deployment service and more flexible to adjustments, we applied a transformation to the single token output to the actual probability. This adjustment also facilitates easier evaluation and comparison against classification models. By setting specific thresholds, we can also tune the model's behavior. Below, we present the Algorithm 1 illustrating this.

Algorithm 1 Modified Output Pseudocode

Input: Prompt P, Model M, Tokenizer T
Output: Output Score $S = [p_Y, p_N]$

Step 1: Model Inference
  $input\_ids \leftarrow T.tokenize(P)$
  $output\_ids \leftarrow M.generate(input\_ids)$
  $logits \leftarrow output\_ids.scores$
Step 2: Compute Probabilities for Answers
  $\ell_Y \leftarrow logits[Y]$
  $\ell_N \leftarrow logits[N]$
Compute softmax probabilities:
  $p_Y \leftarrow \frac{e^{\ell_Y}}{e^{\ell_Y}+e^{\ell_N}}$
  $p_N \leftarrow \frac{e^{\ell_N}}{e^{\ell_Y}+e^{\ell_N}}$
Step 3: Generate Output Score
  $S \leftarrow [p_Y, p_N]$ {Final probability list}
return S

5 Experiments

In this section, we introduce our experimental setup, including prompt design and adjustments to MLLM output for ranking probability. We then briefly describe the baseline models used for comparison. Finally, we present our experiments and provide a detailed analysis of the results.

5.1 Prompt Design

Prompt engineering plays a crucial role in optimizing MLLM performance. For our content moderation application, we designed two straightforward prompt questions, each targeting a different level of labels in the dataset. These prompts can be used independently or combined in various ways. In total, we designed four different prompt templates. (see Figure 2 for details).

To simulate classification, we restrict the model's output to a single-token response (Yes/No) by controlling the answer format in the training dataset. This ensures that the MLLM operates in a structured classification framework while retaining the adaptability of prompt-based reasoning.

5.2 Baseline models

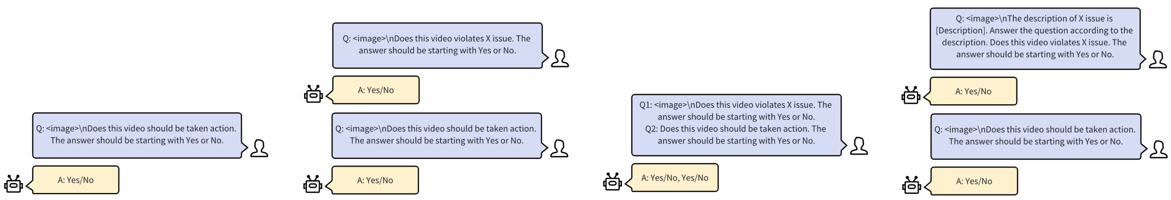
We compare our models with two types of models: *Traditional Multimodal Classification Model (Zeng et al., 2022).* This kind of model is widely used in modern content moderation systems. Comparison against it highlights whether our MLLM-based approach provides a performance advantage over conventional methods.
Zero-Shot MLLMs. This comparison evaluates the impact of our supervised fine-tuning pipeline, demonstrating whether fine-tuned MLLMs outperform their zero-shot counterparts.

5.3 Evaluation Data and Metrics

To ensure alignment with online data distribution, we randomly sample cases from the Router's output and use high-quality annotators as ground truth. The final evaluation dataset consists of 50K samples. For a comprehensive performance assessment, we report PR-AUC, ROC-AUC, and Max-F1 scores.

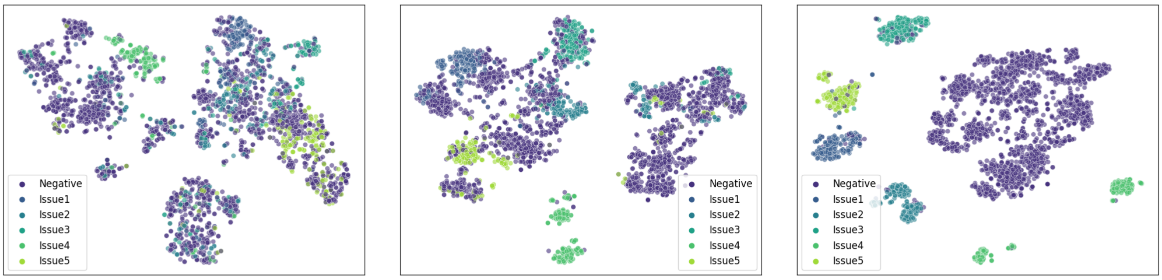
5.4 Offline Evaluation Results

From Table 1, we may conclude the following aspects.
Model Architect. MLLM significantly outperforms traditional multimodal classification models on F1 score by 66.50%, demonstrating its superior ability in content moderation.



(a) Prompt Template 1 (b) Prompt Template 2 (c) Prompt Template 3 (d) Prompt Template 4

Figure 2: Illustration of the four prompt templates: (a) Directly ask about the overall label, (b) Ask the fine-grained label and overall label separately, (c) Ask the fine-grained and overall labels sequentially to emphasize their relationship, and (d) Provide a definition of the fine-grained issue before asking both questions separately.



(a) LLaVA (b) LLaVA w/ Caption (c) Best fine-tuned model

Figure 3: Visualization of the embeddings extracted from the last hidden layer of each model.

Supervised Fine-tuning. Fine-tuned MLLMs outperform zero-shot models by 45.55% in PR-AUC, confirming the effectiveness of our supervised fine-tuning pipeline.

Training Strategy. Multi-task training models consistently outperform the alternative approach across all prompts, demonstrating greater robustness. In contrast, the sequential phased training strategy is more time-efficient and flexible. It allows us to achieve nearly the same performance in significantly less time, as fine-tuning is only required in the second stage with the content moderation dataset.

5.4.1 Ablation Study

Prompt Design. Prompt design matters: Separately asking two questions yields the best performance. Single-question prompts like P1 and P3 do not provide as much information as multiple questions do. As for P2 being better than P4, it is likely because combining both labels in a single prompt introduces additional noise, confusing the final prediction of the model.

Label Assemble. We compared several widely used assembling methods to aggregate fine-grained label predictions and overall label predictions: *Union Probability*, *Maximum Probability*, *Weighted Sum Probability*, and *Bayesian Fusion Probability*(Chen et al., 2022). As shown in Table 2, the *Weighted Sum* method achieves the highest PR-AUC, while the *Union Probability* method performs best in ROC-AUC.

Temperature Tuning. We experimented with different temperature values ranging from 0.2 to 0.8 to thoroughly investigate the impact of randomness on the final outcome. However, the results show that temperature does not have a big impact on model performance.

Method	PR-AUC	ROC-AUC	Max-F1
Original	68.73	87.68	**61.29**
Union	68.78	**87.83**	61.28
Maximum	68.79	87.78	**61.29**
Weighted Sum	**68.83**	87.78	61.28
Bayesian Fusion	68.67	87.79	61.22

Table 2: Result(%) of different label assemble methods.

5.4.2 Visualization

To more intuitively demonstrate the model output distribution, we extracted the final hidden layer of three models and visualized the embeddings. It is obvious that our best model draws a better decision boundary, as shown in Figure 3.

5.5 Online Experiment

We deploy our cascade system online and conduct A/B experiments on 12 representative issues. We evaluate the final result on the following metrics.

5.5.1 Action Volume and Precision Increment

The online experiment shows an average increase of 41.27% in action volume. Furthermore, with the addition of ranker, system-wise precision saw an improvement of 19.16%. For a detailed breakdown of each issue, see Appendix A.

5.5.2 Resources Saving

We observed that the router has eliminated traffic flow by 97.5% without increasing latency in serving, which means filtering numerous compliance videos and saving resources for the ranker to better distinguish potential high-risk videos. Furthermore, compared to the traditional multimodality classification model, our MLLM uses only 2% of the human-annotated data, significantly saving human resources.

6 Conclusion

In this paper, we introduced an MLLM-based cascade system for industrial-scale content moderation. Our approach demonstrated strong performance in both offline and real-world online experiments. Furthermore, our system design enables the efficient deployment of MLLMs at production scale while maintaining affordable computational costs. This solution has been successfully integrated into production systems, driving actual downstream business applications and setting a new benchmark for scalable AI-driven content moderation.

Limitations

The current model still relies on a small amount of human-annotated data, which may introduce additional noise. Moreover, due to the limitations of the router component, the system still carries a risk of missed detection.

References

Le Binh, Rajat Tandon, Chingis Oinar, Jeffrey Liu, Uma Durairaj, Jiani Guo, Spencer Zahabizadeh, Sanjana Ilango, Jeremy Tang, Fred Morstatter, et al. 2022. Samba: Identifying inappropriate videos for young children on youtube. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 88–97.

Yi-Ting Chen, Jinghao Shi, Zelin Ye, Christoph Mertz, Deva Ramanan, and Shu Kong. 2022. Multimodal object detection via probabilistic ensembling. In *European Conference on Computer Vision*, pages 139–158. Springer.

Zhanpeng Chen, Chengjin Xu, Yiyan Qi, and Jian Guo. 2024. MLLM is a strong reranker: Advancing multimodal retrieval-augmented generation via knowledge-enhanced reranking and noise-injected training. *arXiv preprint arXiv:2407.21439*.

Zhiling Chen, Hanning Chen, Mohsen Imani, and Farhad Imani. 2025. Can multimodal large language models be guided to improve industrial anomaly detection? *arXiv preprint arXiv:2501.15795*.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4302–4310, Red Hook, NY, USA. Curran Associates Inc.

Mithun Das, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta, and Animesh Mukherjee. 2023. Hatemm: A multi-modal dataset for hate video classification. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 1014–1023.

DeepSeek-AI. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI. 2025b. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2024. How abilities in large language models are affected by supervised fine-tuning data composition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 177–198, Bangkok, Thailand. Association for Computational Linguistics.

Albert Q. 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. *Preprint*, arXiv:2310.06825.

Hanzhong Liang, Jinghao Shi, et al. 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 (SIGIR)*, Padua, Italy. To appear.

Gorden Liu, Yu Sun, Ruixiao Sun, Xin Dong, and Hongyu Xiong. 2024a. Agentps: Agentic process supervision for multi-modal content quality assurance through multi-round qa. *arXiv preprint arXiv:2412.15251*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Huan Ma, Changqing Zhang, Huazhu Fu, Peilin Zhao, and Bingzhe Wu. 2023. Adapting large language models for content moderation: Pitfalls in data engineering and supervised fine-tuning. *arXiv preprint arXiv:2310.03400*.

Chancharik Mitra, Brandon Huang, Tianning Chai, Zhiqiu Lin, Assaf Arbelle, Rogerio Feris, Leonid Karlinsky, Trevor Darrell, Deva Ramanan, and Roei Herzig. 2025. Sparse attention vectors: Generative multimodal model features are discriminative vision-language classifiers. *Preprint*, arXiv:2412.00142.

Sankha Subhra Mullick, Mohan Bhambhani, Suhit Sinha, Akshat Mathur, Somya Gupta, and Jidnya Shah. 2023. Content moderation for evolving policies using binary question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 561–573.

notAI.tech. 2024. Nudenet: lightweight nudity detection.

OpenAI. 2024. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Wei Pang, Chuan Zhou, Xiao-Hua Zhou, and Xiaojie Wang. 2024. Phased instruction fine-tuning for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5735–5748, Bangkok, Thailand. Association for Computational Linguistics.

Aldo Pareja, Nikhil Shivakumar Nayak, Hao Wang, Krishnateja Killamsetty, Shivchander Sudalairaj, Wenlong Zhao, Seungwook Han, Abhishek Bhandwaldar, Guangxuan Xu, Kai Xu, Ligong Han, Luke Inglis, and Akash Srivastava. 2024. Unveiling the secret recipe: A guide for supervised fine-tuning small llms. *Preprint*, arXiv:2412.13337.

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.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Mengyang Wu, Yuzhi Zhao, Jialun Cao, Mingjie Xu, Zhongming Jiang, Xuehui Wang, Qinbin Li, Guangneng Hu, Shengchao Qin, and Chi-Wing Fu. 2024. ICM-Assistant: instruction-tuning multimodal large language models for rule-based explainable image content moderation. *arXiv preprint arXiv:2412.18216*.

Binqian Xu, Xiangbo Shu, Haiyang Mei, Guosen Xie, Basura Fernando, and Jinhui Tang. 2025. Fedmllm: Federated fine-tuning mllm on multimodal heterogeneity data. *Preprint*, arXiv:2411.14717.

Meng Ye, Karan Sikka, Katherine Atwell, Sabit Hassan, Ajay Divakaran, and Malihe Alikhani. 2023. Multilingual content moderation: A case study on Reddit. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3828–3844, Dubrovnik, Croatia. Association for Computational Linguistics.

Chenghui Yu, Peiyi Li, Haoze Wu, Yiri Wen, Bingfeng Deng, and Hongyu Xiong. 2025. Usm: Unbiased survey modeling for limiting negative user experiences in recommendation systems. *Preprint*, arXiv:2412.10674.

Jialin Yuan, Ye Yu, Gaurav Mittal, Matthew Hall, Sandra Sajeev, and Mei Chen. 2024. Rethinking multimodal content moderation from an asymmetric angle with mixed-modality. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8532–8542.

Yan Zeng, Xinsong Zhang, and Hang Li. 2022. Multi-grained vision language pre-training: Aligning texts with visual concepts. *Preprint*, arXiv:2111.08276.

Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. 2024. Why are visually-grounded language models bad at image classification? *arXiv preprint arXiv:2405.18415*.

Xiongtao Zhou, Jie He, Yuhua Ke, Guangyao Zhu, Victor Gutierrez Basulto, and Jeff Pan. 2024. An empirical study on parameter-efficient fine-tuning for MultiModal large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10057–10084, Bangkok, Thailand. Association for Computational Linguistics.

A Detailed Experiment Result

This is a detailed breakdown of the volume increase for each issue.

Issue	Action Volume Increase (%)
1	47.07
2	59.96
3	45.18
4	27.64
5	22.03
6	36.04
7	41.78
8	65.62
9	26.11
10	63.31
11	29.66
12	30.88
Average	41.27

Table 3: Action Volume Increase for Each Issue