RADAR: Risk-Aware Distilled Adaptive Routing for Efficient Short-Form Video Platform Ecosystem Governance

Baoyu Jing[1], **Zixuan Wang**[2], **Junwen Chen**[2], **Xin Dong**[3] and **Bingfeng Deng**[2]
[1]University of Illinois at Urbana-Champaign
[2]Independent Researcher
[3]Rutgers University

Abstract

Large-scale integrity enforcement on short-form video platforms typically relies on multiple specialized vertical modules, each dedicated to a specific risk category. However, exhaustively executing these computationally intensive modules over massive content streams leads to substantial inference overhead, despite the fact that most content is benign and violations are usually confined to limited policy domains. To address this inefficiency, we propose RADAR, a lightweight risk-aware routing framework that selectively releases low-risk content while dispatching high-risk instances to appropriate vertical modules. Industrial deployment of such routing systems presents two major challenges: (1) systematic label sparsity caused by disjoint annotation pipelines across risk categories, and (2) the capacity-efficiency tradeoff inherent to compact routing architectures. To overcome these challenges, RADAR incorporates Validity-Aware Masking to handle fragmented supervision and Expert-Guided Knowledge Distillation to transfer knowledge from heavyweight expert models into the lightweight router. Experiments on large-scale real-world datasets demonstrate that the proposed masking strategy effectively mitigates disjoint annotation issues, while distillation substantially enhances routing accuracy, enabling the lightweight router to achieve competitive or superior performance compared to specialized expert models.

1 Introduction

In recent years, Multimodal-Large Language Models (MLLMs) have evolved from academic prototypes (Team et al., 2025; Bai et al., 2025; Li et al., 2024; Liu et al., 2023) to robust industrial applications. Empowered by their exceptional multimodal reasoning and generalization capabilities, these models have been successfully deployed in critical fields such as healthcare (Tu et al., 2024), autonomous driving (Tian et al., 2024) and embodied robotics (Zitkovich et al., 2023). In short-form video platforms such as YouTube Shorts and Instagram Reels, MLLMs have become indispensable for processing massive streams of multimodal content, serving as the backbone for personalized recommendation (Nawara and Kashef, 2025), and ecosystem governance (Zeng et al., 2024).

To uphold ecosystem integrity, platforms ensure policy compliance by deploying a comprehensive suite of vertical modules. Within each module, specialized MLLMs are fine-tuned to detect fine-grained integrity risks, ranging from specific subcategories to different severity levels for tiered enforcement. Conventionally, to ensure strict compliance, every uploaded post is subjected to inference by the full suite of vertical modules. However, this exhaustive execution incurs prohibitive computational costs and latency, hindering scalability at the industrial level (Ong et al., 2024; Ding et al., 2024). This inefficiency arises from the intrinsic nature of content distribution: (1) *Benign Dominance*: the vast majority of user-generated content adheres to integrity policies; and (2) *Violation Boundedness*: integrity breaches typically do not span the entire policy taxonomy simultaneously. Consequently, indiscriminately executing the full suite of verticals inevitably incurs computational waste.

To mitigate these inefficiencies, a promising strategy for industrial deployment is to design a lightweight yet powerful MLLMs-based router. Recent advances in general query routing for Large Language Models (LLMs) (Ong et al., 2024; Feng et al., 2024; Zhang et al., 2025) function as intelligent dispatchers, dynamically mapping diverse incoming queries to specific experts to leverage their specialized capabilities. However, unlike general query routing, risk routing in ecosystem governance introduces two distinct imperatives; (1) *Execution Mechanism*: instead of routing all inputs to downstream models, it incorporates an *early-exit*

mechanism to preemptively release benign content, restricting expert dispatching to a smalls subset of potential risks. (2) *Risk Sensitivity*: operating in a safety-centric environment, the system generally prioritizes high recall.

However, training such a risk router for production faces non-trivial challenges: (1) **Label Sparsity from Disjoint Annotations**: industrial training data is inherently fragmented, originating from independent annotation pipelines tailored to specific risks. Consequently, a sample confirmed as "fake engagement" typically remains unverified for "sexualized behaviors". Naively treating these missing labels as negatives induces systematic label bias, creating supervision gaps that bias the router to overlook potential violations. (2) **The Capacity-Efficiency Dilemma**: industrial deployment demands high throughput, necessitating a lightweight router. However, integrity violations often rely on subtle, context-dependent nuances (e.g., distinguishing context-appropriate beachwear from provocative domestic exposure), that typically require complex vertical models. Compressing such deep reasoning capabilities into a compact architecture creates a severe semantic gap.

To address these challenges, we propose RADAR, a risk-aware routing framework tailored for short-form video platform ecosystem governance. (1) First, to tackle the label sparsity from disjoint annotations, we use a **Validity-Aware Masking** strategy. By masking unverified risk categories during loss computation, this mechanism prevents the model from treating missing labels as negatives, effectively neutralizing systematic supervision bias. (2) Second, to bridge the capacity-efficiency gap, we employ **Expert-Guided Knowledge Distillation**. We first train expert models for each risk category, whose deep semantic knowledge is then distilled into the lightweight router to ensure high precision with minimal parameter overhead.

We evaluate RADAR on industrial datasets. Experimental results demonstrate that incorporating validity-aware masking is crucial for handling disjoint labels. Furthermore, RADAR not only effectively captures the expertise of the teacher expert models but, notably, even outperforms the experts.

The major contributions of the paper are:

- We introduce a VLM-based risk-aware router RADAR, for short-form video platform ecosystem governance, which dynamically dispatch posts to downstream vertical modules.
- We adopt a *Validity-Aware Masking* strategy to tackle disjoint annotations, effectively neutralizing systematic label bias by filtering unverified supervision signals.
- We introduce an *Expert-Guided Knowledge Distillation* approach to bridge the capacity-efficiency gap, enabling the router to inherit strong reasoning capabilities of expert models through distillation.
- Experiments on real-world industrial datasets demonstrate that the proposed RADAR has a strong risk detection capability and could reduce more than 80% traffic. Notably, RADAR surpasses the performance of specialized expert models in their respective risk domains.

2 Related Work

2.1 LLM Routers

LLM Routing optimizes the trade-off between performance and cost by dynamically dispatching queries. Established approaches employ predictive classifiers to estimate model win rates (Ong et al., 2024; Feng et al., 2024; Song et al., 2025) or cascading strategies based on sequential thresholds (Chen et al., 2023; Aggarwal et al., 2024). Recent advancements enhance robustness by integrating uncertainty estimation (Ding et al., 2024), while reinforcement learning frameworks like Router-R1 (Zhang et al., 2025) further model routing as a sequential decision process for complex reasoning tasks. Unlike general query routing, risk routing in ecosystem governance employs an early-exit mechanism to target only potential integrity violations, prioritizing high recall for safety over general performance balance.

2.2 Ecosystem Governance Models

As social media posts naturally integrate imagery and text, multimodal models have emerged as the prevailing paradigm for content understanding, particularly in the identification of integrity risks, such as "Child Protection Content" (Wu et al., 2025), and "Fake Engagement" (Sun et al., 2025). Although user feedback signal models (Yu et al., 2024, 2025) and traditional neural network models (Momo et al., 2023) remain components of ecosystem governance, architectures dedicated to content understanding based on Multimodal Large Language Models (MLLMs) have seen rapid development (AlDahoul et al., 2024; Zeng et al., 2024;

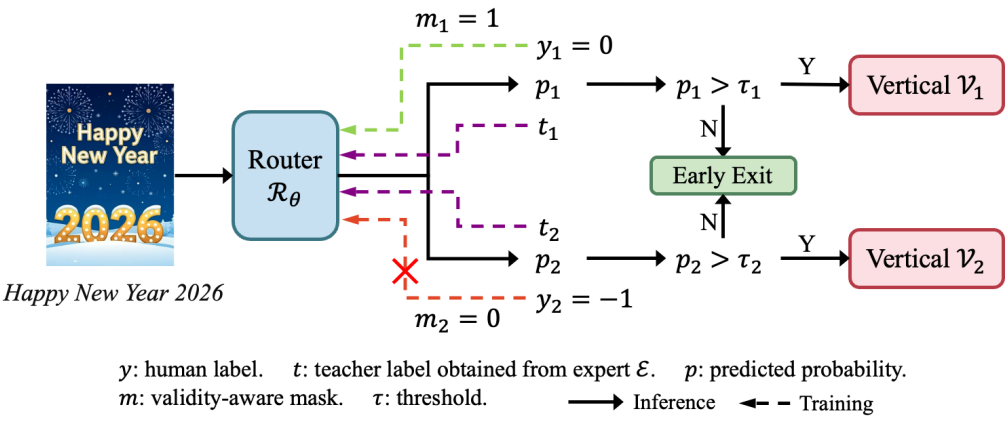


Figure 1: Overview of RADAR.

Wang et al., 2025a,b). Mainstream approaches for applying MLLMs include zero-shot inference (Chen et al., 2024), supervised fine-tuning (Ma et al., 2024; Liu et al., 2025), and reinforcement learning (Firooz et al., 2025). While effective, the prohibitive cost of full MLLM inference across massive content streams demands more efficient, risk-aware routing architectures.

3 Methodology

3.1 Problem Formulation

Short-form video platforms enforce ecosystem integrity through a set of specialized *vertical modules*, where each module comprises multiple models dedicated to a specific risk category, such as fake engagement. To mitigate the prohibitive computational overhead of indiscriminately executing this full suite of vertical modules, incorporating a lightweight risk-aware router is essential. The risk-aware router functions as an intelligent dispatcher, preemptively releasing low-risk content to bypass redundant computations, while directing suspicious instances to the vertical modules for further analysis.

Formally, given an input $x \in \mathcal{X}$, let $\mathcal{V} = \{\mathcal{V}_k\}_{k=1}^{K}$ denote a set of downstream vertical modules, where K is total number of risk categories, each $\mathcal{V}_k$ performs a complex fine-grained assessment to determine the specific violation subcategories and assign severity levels under the k-th integrity policy.

We define the risk-aware router as a lightweight model, parameterized by θ:

$$\mathcal{R}_\theta : \mathcal{X} \to [0,1]^K. \tag{1}$$

For an input x, the router predicts a probability vector $\mathbf{p} = \mathcal{R}_\theta(x)$, where p_k represents the estimated likelihood of the k-th violation. Based on $\mathbf{p}$, the system makes a binary routing decision $\mathbf{d} \in \{0,1\}^K$ via a thresholding mechanism:

$$d_k = \mathbb{I}(p_k > \tau_k), \tag{2}$$

where $\mathbb{I}(\cdot)$ is the indicator function and τ_k is a category-specific threshold. If $d_k = 1$, x is dispatched to the vertical module $\mathcal{V}_k$; otherwise, the system performs an early exit, skipping the processing stage of $\mathcal{V}_k$ for x.

3.2 Evaluation

To comprehensively assess the trade-off between risk detection capability and traffic reduction, we adopt two groups of metrics. For *risk detection capability*, we use standard metrics such as **Area Under the Precision-Recall Curve (PR-AUC)** and **Recall at Fixed Precision (R@P)**. By fixing the guardrail precision to a target level P, R@P measures the maximum achievable recall, indicating the model's capacity to identify violations without excessive false alarms. For *traffic reduction*, we propose **Auto-Pass Rate at Recall at Fixed Recall (APR@R)** to explicitly quantify the efficiency gain brought by the router. Given a recall target R, this metric calculates the proportion of inputs that are preemptively released by the router, thereby successfully bypassing the downstream vertical modules. Formally, let $\tau_k \in [0,1]$ be the threshold satisfying the recall constraint for the k-th vertical module $\mathcal{V}_k$. The auto-pass rate is defined as:

$$\text{APR@R} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(p_k^{(i)} \le \tau_k), \tag{3}$$

where $p_k^{(i)}$ is the router's estimated likelihood that the i-th input regarding the k-th risk category, N is

total number of samples. A higher APR@R implies greater computational savings while adhering to integrity constraints.

3.3 Inference

Figure 1 illustrates the overall inference flow (solid arrows) of the proposed RADAR. Given an input x, the router $\mathcal{R}_\theta$ performs a multi-label classification to yield risk probabilities $\mathbf{p} \in [0,1]^K$. Specifically, given an input x, we extract the final hidden state from a MLLM backbone, e.g., Qwen3-VL-4B (Bai et al., 2025). This representation is then projected via an Multi-Layer Perceptron (MLP) to a multi-label classifier with category-specific heads, where each head is dedicated to a distinct risk category.

To balance safety coverage and computational efficiency, we employ a per-policy triggering mechanism: x is dispatched to the vertical module $\mathcal{V}_k$ only when the predicted score p_k exceeds a predefined confidence threshold τ_k. This allows the system to bypass expensive inferences for low-risk content while ensuring that potential violations are processed by the appropriate vertical models.

3.4 Training

The training of routers faces two main challenges: (1) **Label Sparsity from Disjoint Annotations** arising from independent annotation pipelines in industrial production, and (2) **The Capacity-Efficiency Dilemma** between the lightweight router and heavy models. We address these via Validity-Aware Masking and Expert-Guided Knowledge Distillation.

Validity-Aware Masking. In industrial production, each vertical module typically follows an independent development lifecycle, resulting in isolated annotation pipelines. Consequently, we construct our training set by aggregating samples from these distinct sources. Since a sample is verified only within its specific pipeline, it lacks annotations for other risk categories. We denote this using a unified label vector $\mathbf{y} \in \{-1, 0, 1\}^K$, where $y_k \in \{0, 1\}$ means that the sample has been verified as positive (1) or negative (0) by human labelers for the k-th category, and $y_k = -1$ indicates a lack of verification, i.e., missing annotation. Instead of naively imputing unverified labels (-1) as safe (0) due to the sparsity of real-world risks, we employ a validity mask $\mathbf{m}$ to strictly isolate valid supervision:

$$m_k = \mathbb{I}(y_k \neq -1). \tag{4}$$

We define the per-sample masked loss as:

$$\mathcal{L}_{\text{mask}} = \sum_{k=1}^{K} m_k \cdot \ell_{\text{bce}}(p_k, y_k), \tag{5}$$

where $\ell_{\text{bce}}(p, y) = -[y \log p + (1-y)\log(1-p)]$ denotes the binary cross-entropy loss. By filtering unverified signals, this formulation eliminates the systematic bias caused by sparse annotations.

Expert-Guided Knowledge Distillation. Directly optimizing the lightweight router often yields sub-optimal performance due to its limited parameter budget and representational capacity (Hinton et al., 2015; Gou et al., 2021). To address this, we adopt a teacher-student distillation paradigm (Hinton et al., 2015). We first train high-capacity expert models $\mathcal{E} = \{\mathcal{E}_k\}_{k=1}^K$, e.g., Qwen3-VL-8B, independently for each risk category to serve as strong teachers. Subsequently, we distill their superior risk-detection capabilities into the router $\mathcal{R}_\theta$.

This strategy offers two distinct advantages: (1) *Decision Boundary Alignment*: it enables the lightweight router to approximate the complex decision boundaries of the heavy experts, maintaining high accuracy alongside inference efficiency. (2) *Mitigation of Label Sparsity*: crucially, this approach addresses the label sparsity issue caused by disjoint annotation pipelines. The experts provide dense supervision signals, i.e., soft pseudo-labels, across all categories for every sample, effectively filling the supervision gaps.

Formally, let $t_k = \mathcal{E}_k(x)$ denote the soft probability output by the k-th expert. We minimize the Kullback-Leibler (KL) divergence between the teacher distribution t_k and student distribution p_k (Hinton et al., 2015):

$$\mathcal{L}_{\text{distill}} = \sum_{k=1}^{K} \text{KL}(t_k || p_k). \tag{6}$$

Overall Objective. The overall training objective is formulated as a convex combination of the masking and distillation losses:

$$\mathcal{L} = (1-\lambda)\mathcal{L}_{\text{mask}} + \lambda \mathcal{L}_{\text{distill}}, \tag{7}$$

where $\lambda \in [0, 1]$ is a tunable hyperparameter balancing the trade-off between direct supervision and expert guidance.

Loss	Method	LQC		RCS		FE		Average	
		PR-AUC	R@P60	PR-AUC	R@P60	PR-AUC	R@P60	PR-AUC	R@P60
BCE	Individual Experts	82.57	81.06	70.90	72.53	61.10	**57.46**	71.52	70.35
BCE w/o Mask	SigLIP2	58.27	54.41	41.77	22.89	33.94	19.88	44.66	32.39
	LLaVA-OV-0.5B	68.13	64.89	51.39	35.11	43.82	31.31	54.45	43.77
	Qwen3-VL-2B	72.50	71.56	56.22	46.40	51.57	42.25	60.10	53.40
Masked BCE	SigLIP2	78.16	76.18	53.57	46.74	37.42	15.36	56.38	46.09
	LLaVA-OV-0.5B	79.97	78.56	65.91	64.53	52.37	43.48	66.08	62.19
	Qwen3-VL-2B	79.80	78.56	68.23	70.41	56.46	46.97	68.16	65.31
Masked BCE	Gemma3-4B	81.10	79.38	69.55	71.91	56.92	49.85	69.19	67.05
	LLaVA-OV-1.5-4B	80.53	79.18	67.41	69.27	53.50	46.36	67.15	64.94
	Qwen3-VL-4B	80.27	79.18	70.34	72.08	60.53	53.48	70.38	68.25
Masked BCE + Distill	Gemma3-4B	81.15	80.06	70.12	71.98	56.95	49.86	69.41	67.30
	LLaVA-OV-1.5-4B	80.74	80.00	67.90	69.71	53.53	48.03	67.39	65.91
	Qwen3-VL-4B	**83.43**	**82.47**	**73.00**	**74.89**	**62.78**	54.55	**73.07**	**70.64**

Table 1: Risk detection capability. The best and second-best results are highlighted in bold and underlined.

Loss	Method	LQC		RCS		FE		Average	
		APR@R90	APR@R95	APR@R90	APR@R95	APR@R90	APR@R95	APR@R90	APR@R95
BCE	Individual Experts	89.02	78.92	**88.87**	83.96	86.31	75.25	88.07	79.38
BCE w/o Mask	SigLIP2	65.07	46.75	62.56	45.39	58.97	41.09	62.20	44.41
	LLaVA-OV-0.5B	78.75	67.12	77.63	67.74	72.78	56.41	76.39	63.76
	Qwen3-VL-2B	83.99	76.17	78.22	65.95	79.82	62.61	80.68	68.24
Masked BCE	SigLIP2	79.31	68.02	71.75	57.78	68.61	48.96	73.22	58.25
	LLaVA-OV-0.5B	85.13	78.00	84.67	77.52	76.57	65.58	82.12	73.70
	Qwen3-VL-2B	87.16	77.00	87.13	81.02	82.51	72.55	85.60	76.86
Masked BCE	Gemma3-4B	86.96	78.65	86.15	79.61	84.66	73.63	85.92	77.30
	LLaVA-OV-1.5-4B	87.73	78.02	85.12	77.45	78.19	67.54	83.68	74.34
	Qwen3-VL-4B	87.09	78.82	86.42	79.48	85.68	74.44	86.40	77.58
Masked BCE + Distill	Gemma3-4B	87.03	78.82	86.20	79.65	84.58	73.56	85.94	77.34
	LLaVA-OV-1.5-4B	86.38	77.79	85.14	78.31	79.45	65.89	83.66	74.00
	Qwen3-VL-4B	**90.30**	**80.96**	88.34	**84.21**	**87.49**	**77.16**	**88.71**	**80.78**

Table 2: Traffic reduction. The best and second-best results are highlighted in bold and underlined.

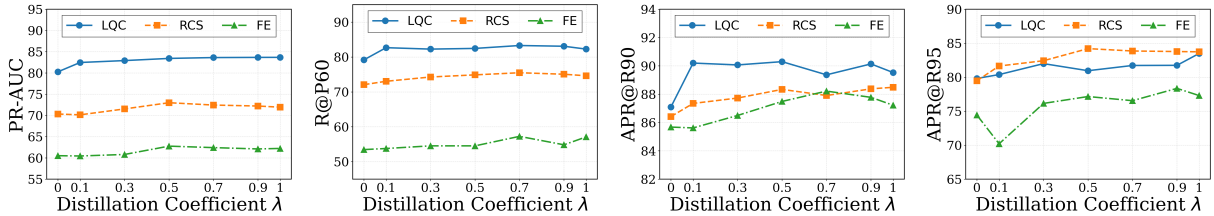
4 Experiments

4.1 Experimental Setup

Datasets. Each sample consists of a list of images and associated text, such as title and Optical Character Recognition (OCR) text. We utilize a large-scale *training* dataset of 1.7M real-world human-labeled samples across three risk categories: **Low Quality Content (LQC)** with 800k samples (25.5% positive), **Regulated Commercial Services (RCS)** with 300k samples (17.7% positive), and **Fake Engagement (FE)** with 600k samples (9.0% positive). For *evaluation*, we use a separate hold-out test set containing 30k samples for FE (3.3% positive), 30k for RCS (4.8% positive), and 15k for LQC (4.4% positive).

Evaluation Metrics. We evaluate the performance of the router from two aspects: risk detection capability and inference cost reduction. For the *risk detection capability*, we use **Recall at Precision 60% (R@P60)**, which measures the fraction of actual risks identified when the detection precision is anchored at 60%, and **Area Under the Precision-Recall Curve (PR-AUC)** which provides a comprehensive summary of the detection performance across all possible decision thresholds. For *traffic reduction*, we use **Auto-Pass Rate at Recall 90% (APR@R90)** and **95% (APR@R95)**, which quantifies the traffic savings achievable while maintaining 90% and 95% recall of positive samples.

Models. We use Qwen3-VL-8B (Bai et al., 2025) as the backbone for experts (or teachers). We implement routers based on the following groups of MLLMs: (1) *Tiny MLLMs*: SigLIP2 (Tschannen et al., 2025), LLaVA-OneVision-0.5B (Li et al., 2024), Qwen3-VL-2B (Bai et al., 2025). (2) *Small MLLMs*: Qwen3-VL-4B (Bai et al., 2025), Gemma3-4B (Team et al., 2025), LLaVA-OneVision-1.5-4B (An et al., 2025). These MLLMs are trained with different loss functions.



(a) PR-AUC (b) R@P60 (c) APR@R90 (d) APR@R95

Figure 2: Sensitivity study for distillation coefficient λ.

Implementation Details. We instantiate routers based on the aforementioned tiny and small MLLMs. The router consists of a MLLM backbone followed by a multi-label classifier. Specifically, we utilize the final hidden embedding as the representation of the input, which is then processed by an MLP followed by separate output heads that serve as independent binary classifiers for each risk category. During training, we apply LoRA (Hu et al., 2022) with rank=128 and $\alpha = 256$. We employ the AdamW optimizer (Loshchilov and Hutter, 2017), and set the learning rate to 1×10^{-5} under a cosine schedule with 20% warm-up.

4.2 Main Results

We present the results of risk detection capability and traffic reduction in Table 1 and Table 2.

Validity-Aware Masking. A comparison of tiny MLLMs trained with and without the validity mask consistently demonstrates the significant benefits of our masking strategy. Notably, in Table 1, for the LLaVA-OV-0.5B backbone, the mask yields a substantial gain of 18.42 points in R@P60 ($43.77 \rightarrow 62.19$). Similarly, in Table 2, for the SigLIP2 backbone, it delivers an average improvement of 13.84 points in APR@95 ($44.41 \rightarrow 58.25$). These results indicate that validity-aware masking could effectively reduce the bias of labels.

Model Size. Comparing tiny (0.5B$\sim$2B) and small ($\sim$4B) MLLMs trained with the masked BCE, we observe that increased model capacity translates to better performance. For instance, scaling the backbone from Qwen3-VL-2B to 4B yields clear gains, boosting PR-AUC from 68.16 to 70.38 and APR@90 from 85.60 to 86.40.

Expert-Guided Knowledge Distillation. The last group of rows corresponds to models trained by distilling the experts from the first group. Remarkably, after distillation, Qwen3-VL-4B not only outperforms its non-distilled counterpart but also surpasses the expert models $\mathcal{E}$ in terms of both risk detection capability and traffic reduction. This suggests that the proposed distillation strategy not only effectively reduces the semantic gap between lightweight router and experts, but also effectively serves as a remedy for label sparsity, filling the supervision gaps left by disjoint annotation pipelines.

4.3 Sensitivity Experiments

Figure 2 illustrates the sensitivity analysis of the distillation coefficient λ on the Qwen3-VL-4B model. As observed in Figure 2a-2b, the introduction of the distillation term ($\lambda > 0$) leads to marked improvements in both PR-AUC and R@P60. Specifically, models trained with $\lambda \geq 0.5$ significantly outperform the baseline trained without distillation ($\lambda = 0$) for all LQC, RCS and FE. Regarding APR@R90 and APR@R95 (Figures 2c-2d), performance stabilizes at a high level when $\lambda \geq 0.1$ for LQC and $\lambda \geq 0.3$ for RCS. For FE, optimal performance for APR@R90 and APR@R95 is achieved when $\lambda \geq 0.5$ and $\lambda \geq 0.3$, respectively. Overall, these results demonstrate the effectiveness of our expert-guided knowledge distillation and robustness of our framework to hyperparameter variations.

5 Conclusion

In this paper, we propose RADAR, a lightweight risk-aware routing framework to resolve the computational bottlenecks in short-form video platform ecosystem governance. To address disjoint annotations and the capacity-efficiency dilemma, we introduced Validity-Aware Masking to neutralize label bias and Expert-Guided Knowledge Distillation to distill heavy experts' capability to the compact router. Experimental results show that RADAR reduces downstream traffic by over 80% while maintaining strong risk detection capability. Remarkably, RADAR not only inherits experts' ability but even surpasses the performance of experts.

References

Pranjal Aggarwal, Aman Madaan, Ankit Anand, Srividya Pranavi Potharaju, Swaroop Mishra, Pei Zhou, Aditya Gupta, Dheeraj Rajagopal, Karthik Kappaganthu, Yiming Yang, and 1 others. 2024. Automix: Automatically mixing language models. *Advances in Neural Information Processing Systems*, 37:131000–131034.

Nouar AlDahoul, Myles Joshua Toledo Tan, Harishwar Reddy Kasireddy, and Yasir Zaki. 2024. Advancing content moderation: Evaluating large language models for detecting sensitive content across text, images, and videos. *Preprint*, arXiv:2411.17123.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, and 1 others. 2025. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. *Preprint*, arXiv:2402.04788.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. Frugalgpt: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*.

Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Ruhle, Laks VS Lakshmanan, and Ahmed Hassan Awadallah. 2024. Hybrid llm: Cost-efficient and quality-aware query routing. *arXiv preprint arXiv:2404.14618*.

Tao Feng, Yanzhen Shen, and Jiaxuan You. 2024. Graphrouter: A graph-based router for llm selections. *arXiv preprint arXiv:2410.03834*.

Hamed Firooz, Rui Liu, Yuchen Lu, Zhenyu Hou, Fangzhou Xiong, Xiaoyang Zhang, Changshu Jian, Zhicheng Zhu, Jiayuan Ma, Jacob Tao, Chaitali Gupta, Xiaochang Peng, Shike Mei, Hang Cui, Yang Qin, Shuo Tang, Jason Gaedtke, and Arpit Mittal. 2025. Scaling reinforcement learning for content moderation with large language models. *Preprint*, arXiv:2512.20061.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International journal of computer vision*, 129(6):1789–1819.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

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

Mingchao Liu, Yu Sun, Ruixiao Sun, Xin Dong, Xiang Shen, and Hongyu Xiong. 2025. Agentps: Agentic process supervision for content moderation with multimodal llms. *Preprint*, arXiv:2412.15251.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Huan Ma, Changqing Zhang, Huazhu Fu, Peilin Zhao, and Bingzhe Wu. 2024. Adapting large language models for content moderation: Pitfalls in data engineering and supervised fine-tuning. *Preprint*, arXiv:2310.03400.

Mhd Adel Momo, Hezerul Bin Abdul Karim, Michael Aaron G. Sy, Ahmad Albunni, Myles Joshua Toledo Tan, and Nouar Aldahoul. 2023. Evaluation of convolution and attention networks for nudity and pornography detection in sketch images. *2023 IEEE Symposium on Computers & Informatics (ISCI)*, pages 7–12.

Dina Nawara and Rasha Kashef. 2025. A comprehensive survey on llm-powered recommender systems: from discriminative, generative to multi-modal paradigms. *IEEE Access*.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E Gonzalez, M Waleed Kadous, and Ion Stoica. 2024. Routellm: Learning to route llms with preference data. *arXiv preprint arXiv:2406.18665*.

Wei Song, Zhenya Huang, Cheng Cheng, Weibo Gao, Bihan Xu, GuanHao Zhao, Fei Wang, and Runze Wu. 2025. Irt-router: Effective and interpretable multi-llm routing via item response theory. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15629–15644.

Yu Sun, Yin Li, Ruixiao Sun, Chunhui Liu, Fangming Zhou, Ze Jin, Linjie Wang, Xiang Shen, Zhuolin Hao, and Hongyu Xiong. 2025. Audio-enhanced vision-language modeling with latent space broadening for

high quality data expansion. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 4872–4881.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. 2024. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, and 1 others. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, and 1 others. 2024. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138.

Zixuan Wang, Jinghao Shi, Hanzhong Liang, Xiang Shen, Vera Wen, Zhiqian Chen, Yifan Wu, Zhixin Zhang, and Hongyu Xiong. 2025a. 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. 2025b. 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.

Mengyang Wu, Yuzhi Zhao, Jialun Cao, Mingjie Xu, Zhongming Jiang, Xuehui Wang, Qinbin Li, Guangneng Hu, Shengchao Qin, and Chi-Wing Fu. 2025. Icm-assistant: Instruction-tuning multimodal large language models for rule-based explainable image content moderation. *Preprint*, arXiv:2412.18216.

Chenghui Yu, Peiyi Li, Haoze Wu, Yiri Wen, Bingfeng Deng, and Hongyu Xiong. 2024. Usm: Unbiased survey modeling for limiting negative user experiences in recommendation systems. *arXiv preprint arXiv:2412.10674*.

Chenghui Yu, Haoze Wu, Jian Ding, Bingfeng Deng, and Hongyu Xiong. 2025. Unified survey modeling to limit negative user experiences in recommendation systems. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, pages 1104–1107.

Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, and 1 others. 2024. Shieldgemma: Generative ai content moderation based on gemma. *arXiv preprint arXiv:2407.21772*.

Haozhen Zhang, Tao Feng, and Jiaxuan You. 2025. Router-r1: Teaching llms multi-round routing and aggregation via reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, and 1 others. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR.