RAVEN++: Pinpointing Fine-Grained Violations in Advertisement Videos with Active Reinforcement Reasoning

Deyi Ji[1*] **Yuekui Yang**[1,2*] **Liqun Liu**[1†] **Peng Shu**[1] **Haiyang Wu**[1] **Shaogang Tang**[1]
Xudong Chen[1] **Shaoping Ma**[2] **Tianrun Chen**[3] **Lanyun Zhu**[4†]
[1]Tencent [2]Department of Computer Science and Technology, Tsinghua University
[3]Zhejiang University [4]Nanyang Technological University
{deyiji,yuekuiyang,liqunliu,archershu,gavinwu,teetang,seadenchen}@tencent.com,
msp@tsinghua.edu.cn,tianrun.chen@zju.edu.cn,lanyun.zhu@ntu.edu.sg

Abstract

Advertising (Ad) is a cornerstone of the digital economy, yet the moderation of video advertisements remains a significant challenge due to their complexity and the need for precise violation localization. While recent advancements, such as the RAVEN model, have improved coarse-grained violation detection, critical gaps persist in fine-grained understanding, explainability, and generalization. To address these limitations, we propose RAVEN++, a novel framework that introduces three key innovations: 1) Active Reinforcement Learning (RL), which dynamically adapts training to samples of varying difficulty; 2) Fine-Grained Violation Understanding, achieved through hierarchical reward functions and reasoning distillation; and 3) Progressive Multi-Stage Training, which systematically combines knowledge injection, curriculum-based passive RL, and active RL. Extensive experiments on both public and proprietary datasets, on both offline scenarios and online deployed A/B Testing, demonstrate that RAVEN++ outperforms general-purpose LLMs and specialized models like RAVEN in terms of fine-grained violation understanding, reasoning capabilities, and generalization ability.

1 Introduction

Advertising (Ad) plays a pivotal role in the digital economy, serving as a primary driver of revenue and growth for online platforms (Rathee and Milfeld, 2024; Campbell et al., 2025; Ji et al., 2025a). To maintain legal compliance, support sustainable development, and create a positive user experience, platforms implement strict content moderation policies for advertisers. Despite these efforts, violations of Ad guidelines remain prevalent, presenting ongoing challenges for effective moderation. While recent advancements in large language models (LLMs) (Liu et al., 2023; Bai et al., 2023a,b; Zhu et al., 2024a,b, 2025a) have improved the detection of non-compliant content, critical gaps persist, especially in the moderation (Qiao et al., 2024; Madio and Quinn, 2025; Hasan and Juhannis, 2024; Al Kurdi and Alshurideh, 2025; Baek et al., 2024) of video advertisements.

Among all content types, video advertisements are the most challenging to moderate due to their complexity and the need for precise localization (Huang et al., 2024; Ji et al., 2025b, 2023; Chen et al., 2024; Gu et al., 2024; Ji et al., 2022) of violations. Recent work, such as RAVEN (Ji et al., 2025c), has made significant progress in identifying coarse-grained violation labels and subscenes by leveraging large-scale coarsely annotated datasets and fine-tuning with smaller, precisely annotated datasets. However, in practical industrial applications, RAVEN's capabilities still face certain limitations that highlight opportunities for further enhancement: 1) **Fine-Grained Understanding**: While RAVEN performs well in detecting major violation categories, its ability to identify finer-grained sub-categories and localize sub-scenes with high precision remains an area for improvement; 2) **Explainability**: RAVEN's reasoning process occasionally produces explanations that are verbose or misaligned with label rules, and in some cases, its chain-of-thought (CoT) reasoning may contradict the final output, reducing its interpretability; 3) **Generalization**: RAVEN's performance in out-of-domain (OOD) scenarios is limited, which restricts its applicability in diverse industrial settings.

These limitations arise from three key factors: 1) RAVEN's reliance on fixed rule-based rewards in its GRPO-based (Guo et al., 2025) passive reinforcement learning framework, which limits its adaptability to samples of varying difficulty; 2) insufficient constraints on the model's reasoning and attribution processes, which can lead to inconsistent explanations; 3) a lack of generalized understanding of Ad knowledge and moderation rules,

[*]The first two authors contribute equally to this work.
[†]Corresponding Authors.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 1–10
November 4-9, 2025

which can result in overfitting to specific labels and reduced OOD performance.

To the end, we propose RAVEN++ with three key innovations: 1) **Active Reinforcement Learning**: RAVEN++ dynamically rolls out samples of varying difficulty during training and adaptively interleaves supervised fine-tuning (SFT) with reinforcement learning (RL) to optimize performance, leveraging the complementary strengths of SFT for tasks beyond the model's current capabilities and RL for refining existing skills; 2) **Fine-Grained Violation Understanding**: By designing a hierarchical set of several reward functions that incorporate Tversky Distance and distill reasoning ability from larger models, RAVEN++ significantly improves the accuracy of fine-grained violation categories, the precision of sub-scene localization, and the reliability and logical coherence of its explanations; 3) **Progressive Multi-Stage Training**: RAVEN++ introduces a systematic training framework that combines progressive knowledge injection of Ad and moderation rules, curriculum-based passive RL, and active RL, effectively leveraging both large-scale noisy data and small-scale precisely annotated data to enhance violation granularity, generalization, and interpretability.

We conduct extensive experiments to validate RAVEN++ using both publicly available datasets and proprietary industrial data, on both offline and online deployment scenarios. Our results demonstrate that the RAVEN++ model outperforms models of the same scale, including general-purpose LLMs like LLaVa (Liu et al., 2023) and Qwen(Bai et al., 2023a), as well as the specialized video moderation model RAVEN(Ji et al., 2025c), in terms of fine-grained violation understanding, reasoning capabilities, and generalization ability.

2 Related Work

2.1 Advancements in Reinforcement Learning for Multimodal Systems

Recent breakthroughs in Multimodal Large Language Models (MLLMs) have revolutionized how machines interpret (Liu et al., 2023; Bai et al., 2023a,b; Maity et al., 2024) and generate content across visual and textual domains (Yin et al., 2023; Ji et al., 2024a; Xu et al., 2024a,b; Zhu et al.; Ji et al., 2024b). While foundational architectures like CLIP (Radford et al., 2021) and BLIP (Li et al., 2022) established strong cross-modal alignment, and subsequent models like Flamingo (Alayrac et al., 2022) incorporated temporal awareness, these systems predominantly rely on supervised learning paradigms. This dependence often leads to catastrophic forgetting when adapting to new tasks and limited generalization to unseen data distributions. Emerging research has begun exploring reinforcement learning (Zhu et al., 2025b; Yu et al., 2024; Amini et al., 2024; Song et al., 2024; Liu et al.) with curriculum learning (Narvekar et al., 2020; Bengio et al., 2009; Graves et al., 2017; Hacohen and Weinshall, 2019; Soviany et al., 2022; Pentina et al., 2015) as a complementary framework to address these limitations. Rather than treating reinforcement learning as a separate optimization mechanism, contemporary approaches integrate it as a core component of the multimodal reasoning pipeline. Techniques such as reward-weighted policy optimization (Christiano et al., 2017) and preference-based learning (Rafailov et al., 2024) have demonstrated remarkable success in aligning model outputs with human preferences, particularly in domains requiring nuanced temporal understanding. The integration of chain-of-thought reasoning (Wei et al., 2022; Ji et al., 2024c) with reinforcement learning has further enabled models to decompose complex multimodal tasks into structured decision-making processes, creating more interpretable and controllable systems.

2.2 Active Learning Paradigms for Content Understanding

The paradigm of active reinforcement learning represents a significant shift from traditional passive learning approaches. While conventional systems process whatever data they are given, active reinforcement learning empowers models to strategically acquire the most valuable perceptual information through dynamic interaction with their environment. This approach fundamentally transforms the model from a passive observer to an active participant in the learning process. Modern implementations of active reinforcement learning build upon early theoretical foundations (Epshteyn et al., 2008) but incorporate sophisticated perceptual capabilities. Contemporary systems (Shang and Ryoo, 2023) demonstrate that active viewpoint selection and strategic information gathering can dramatically improve sample efficiency and generalization performance. In the context of content moderation (Al Kurdi and Alshurideh, 2025; Kolla et al., 2024; Kumar et al., 2024; Blackwell, 2025), this

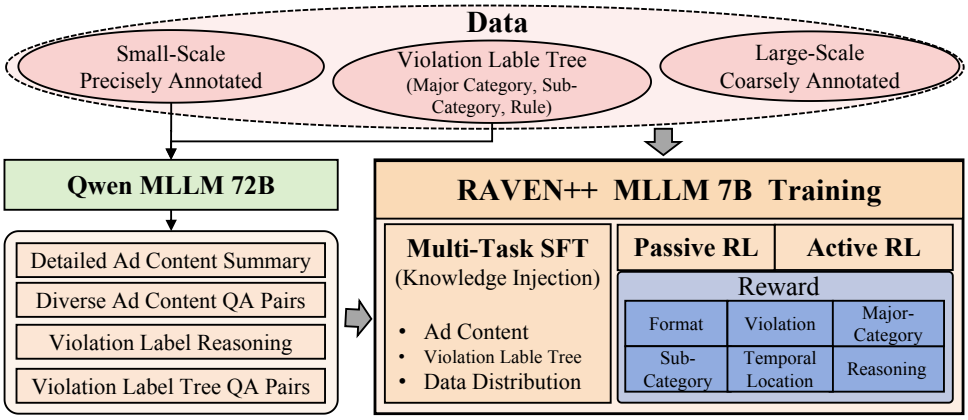

Figure 1: The Overview of RAVEN++ Training.

active approach enables models to proactively identify subtle violations that might be missed through passive observation alone. The framework allows models to dynamically adjust their perceptual focus, prioritizing regions of interest and temporal segments that are most likely to contain policy violations, thereby optimizing both computational resources and detection accuracy.

3 Method

3.1 Overview

3.1.1 Problem Formulation

Given an input video V, a predefined hierarchical violation labels tree T, and a prompt P, the RAVEN++ framework outputs: 1) the violation labels $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$ associated with V, 2) the sub-categories $\mathcal{S}_c = \{s_1, s_2, \ldots, s_m\}$ for each violation label c, 3) the temporal intervals $\mathcal{X}_{c,s} = (t^l_{c,s}, t^r_{c,s})$ of the sub-scenes corresponding to each sub-category s of violation c, and 4) the reasoning $\mathcal{R}_{c,s}$ behind each sub-category s, providing interpretable explanations for the detected violations. Here, $t^l_{c,s}$ and $t^r_{c,s}$ denote the start and end times of the sub-scene for sub-category s of violation c, respectively. The hierarchical structure of T is defined as *"major category → sub-category → rule"*. This extends RAVEN's (Ji et al., 2025c) capabilities by incorporating fine-grained reasoning and improved generalization. Specifically, RAVEN++ outputs fine-grained sub-labels for violations and their corresponding reasoning, enabling more detailed and actionable insights.

3.1.2 Training

As shown in Fig. 1, the manually annotated results $\mathcal{Y}_{c,s} = (y^l_{c,s}, y^r_{c,s})$ often deviate from the ground truth $\mathcal{Z}_{c,s} = (z^l_{c,s}, z^r_{c,s})$ due to annotation errors or ambiguities. While RAVEN achieved significant improvements by leveraging GRPO-based passive reinforcement learning (RL) to handle noisy annotations, we aim to further enhance the framework for: 1) fine-grained identification of sub-labels $\mathcal{S}_c$ for each violation c, 2) precise localization of the temporal intervals $\mathcal{X}_{c,s} = (t^l_{c,s}, t^r_{c,s})$ of sub-scenes corresponding to each sub-category s, and 3) interpretable and reliable reasoning $\mathcal{R}_{c,s}$ for each violation.

To achieve these goals, RAVEN++ introduces progressive multi-stage training: 1) Knowledge Injection with Augmented Data: the base model is initialized with Ad content and label rule knowledge with augmented data by lightweight SFT; 2) Passive RL with Large-Scale Noisy Data: the model is trained on large-scale coarsely annotated data and a small set of precisely annotated data, guided by a hierarchical set of 6 reward functions to learn global data distributions; 3) Active RL with Small-Scale Precise Data: the model actively identifies valuable samples through designed sampling strategies, enabling comprehensive improvements with minimal precisely annotated data.

3.2 Stage 1: Knowledge Injection with Augmented Data

The knowledge injection process integrates both Ad content and moderation rule knowledge into the base model through a unified SFT framework. This process consists of the following steps:

1) Video Summarization and Ad Knowledge Extraction: The input video V is processed by a large model (e.g., Qwen-72B) to generate a detailed caption or summary, including the product name, product attributes (e.g., medicine, cosmetics, clothing),

target audience, and key messaging. Based on the summary, Qwen-72B generates question-answer (QA) pairs, such as:

> *Question: What product is advertised in this video?*
> *Answer: The product is a sunscreen named XXX.*
> *Question: Who is the target audience for this advertisement?*
> *Answer: The target audience is young adults aged about 18–60.*

2) Hierarchical Rule Knowledge Extraction: The moderation rules are structured hierarchically. QA pairs are generated to capture the relationships between major categories, sub-categories, and rules, such as:

> *Question: What are the sub-categories and rules for the main category 'Discomforting Content'?*
> *Answer: Sub-labels: 'Gory Content,' '...', Rules: ...*
> *Question: What constitutes a violation under the sub-category 'Misleading Claims'?*
> *Answer: Claims that exaggerate product efficacy without evidence.*

3) Joint SFT: All the generated QA pairs are combined with a small portion of precisely annotated data to fine-tune the base model.

3.3 Reward Design for RL

The reward function R optimizes the model's RL performance across six dimensions,

3.3.1 Format Reward R_{format}

It ensures the model's output adheres to the predefined structure, which is critical for downstream processing and interpretation, The format requires the output to include the following components:

$$\begin{aligned} R_{\text{format}} = \mathbb{I}(&\langle\text{think}\rangle\langle/\text{think}\rangle\langle\text{reason}\rangle\text{content summarization: ... risk analysis: ... conclusion: ...}\langle/\text{reason}\rangle\langle\text{violation}\rangle\text{Y/N}\langle/\text{violation}\rangle \\ &\langle\text{result}\rangle\{\text{major: ..., sub: ..., ground: ...}\}\langle/\text{result}\rangle \\ &\langle\text{result}\rangle\{\text{major: ..., sub: ..., ground: ...}\}\langle/\text{result}\rangle\ldots), \end{aligned} \tag{1}$$

where $\mathbb{I}(\cdot)$ is an indicator function.

3.3.2 Violation Reward $R_{\text{violation}}$

This reward evaluates whether the model's prediction of a violation matches the ground truth (GT):

$$R_{\text{violation}} = \mathbb{I}(P_{\text{violation}} = GT_{\text{violation}}). \tag{2}$$

3.3.3 Reasoning Consistency Reward R_{reason}

This reward encourages logical and coherent reasoning chains $\mathcal{R}_{c,s}$, structured into summarization, risk analysis, and conclusion. It compares the model's CoT reasoning with a high-quality structured CoT from Qwen-72B (Eq. 1), using a semantic cosine similarity measure with BERT (Devlin, 2018) to enhance interpretability and reliability.

$$R_{\text{reason}} = \text{Sim}(\text{CoT}_{\text{model}}, \text{CoT}_{\text{Qwen-72B}}). \tag{3}$$

3.3.4 Major Category Reward R_{major}

This reward measures the accuracy of the model's prediction for the major violation category $c \in \mathcal{C}$. It uses the Tversky distance (Tversky, 1977), a generalization of the Dice coefficient, to evaluate the similarity between the predicted and ground truth major categories:

$$R_{\text{major}} = 1 - \text{Tversky}(P_{\text{major}}, GT_{\text{major}}), \tag{4}$$

the Tversky distance is particularly effective for imbalanced datasets, as it allows for flexible weighting of false positives and false negatives.

3.3.5 Sub-Category Reward R_{sub}

Similar to the major category reward, this component evaluates the model's performance in identifying the specific violation sub-category $s \in \mathcal{S}_c$. It also employs the Tversky distance for robust evaluation:

$$R_{\text{sub}} = 1 - \text{Tversky}(P_{\text{sub}}, GT_{\text{sub}}). \tag{5}$$

3.3.6 Temporal Grounding Reward R_{ground}

It evaluates the model's ability to localize the violation temporal sub-scenes $\mathcal{X}_{c,s} = (t^l_{c,s}, t^r_{c,s})$ within the video. Following RAVEN, it combines the Intersection over Union (IoU) metric, which measures the overlap between predicted and ground truth temporal segments, with a boundary alignment reward that penalizes deviations in the start and end times to ensure precise localization:

$$\begin{aligned} R_{\text{ground}} = &\text{IoU}(P_{\text{ground}}, GT_{\text{ground}}) + \\ &\text{Boundary}(P_{\text{ground}}, GT_{\text{ground}}). \end{aligned} \tag{6}$$

The overall reward R is

$$\begin{aligned} R = \lambda_1 R_{\text{format}} + \lambda_2 R_{\text{violation}} + \lambda_3 R_{\text{major}} + \\ \lambda_4 R_{\text{sub}} + \lambda_5 R_{\text{ground}} + \lambda_6 R_{\text{reason}}, \end{aligned} \tag{7}$$

$\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6$ are dynamically adjustable hyperparameters that control the relative importance of each reward component.

3.4 Stage 2: Curriculum-Based Passive RL with Large-Scale Noisy Data

To effectively train the model, we adopt a curriculum-based passive reinforcement learning strategy, which dynamically adjusts the weights of the reward components across three phases. This approach ensures the model progressively learns from simpler to more complex tasks, aligning with its evolving capabilities.

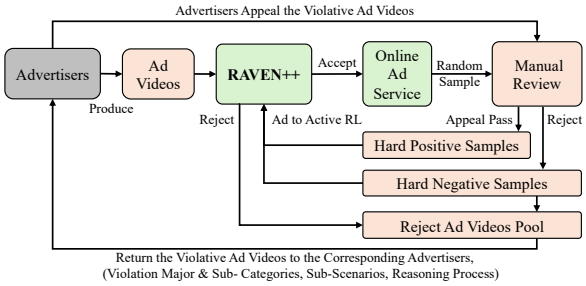

Figure 2: The deployment of RAVEN++.

3.4.1 Phase 1: Format and Violation Learning

In the early phase, the model focuses on mastering the output format and accurately identifying whether a violation exists. This is critical because the task format is complex, and errors in format or violation detection can propagate to downstream tasks. To prioritize these aspects, the weights λ_1 (for R_{format}) and λ_2 (for $R_{\text{violation}}$) are set to relatively large values, while the weights for other reward components are kept small (the reward weights are [1, 1, 0.5, 0.3, 0, 0.1]). This stage ensures the model builds a strong foundation in adhering to the required structure and correctly classifying violations.

3.4.2 Phase 2: Fine-Grained Violation Detection and Reasoning

Once the model demonstrates proficiency in format adherence and violation detection, the focus shifts to more challenging tasks, including identifying major categories $\mathcal{C}$, sub-categories $\mathcal{S}_c$, and generating interpretable reasoning $\mathcal{R}_{c,s}$. In this stage, the weights λ_1 and λ_2 are gradually reduced, while the weights λ_3 (for R_{major}), λ_4 (for R_{sub}), and λ_6 (for R_{reason}) are increased (the reward weights are [0.5, 0.5, 1, 1, 0, 0.5]). This transition enables the model to refine its ability to perform fine-grained violation detection and provide coherent explanations.

3.4.3 Phase 3: Temporal Grounding

In the final phase, the model tackles the most challenging task: precise temporal localization of violations $\mathcal{X}_{c,s} = (t^l_{c,s}, t^r_{c,s})$. At this stage, the weight λ_5 (for R_{ground}) is significantly increased, while the weights for other components are maintained or slightly reduced (the reward weights are [0.2, 0.2, 1, 1, 1, 0.5]). This ensures the model dedicates sufficient resources to learning the temporal boundaries of violations, which is critical for accurate sub-scene localization.

3.5 Stage 3: Active RL with Small-Scale Precise Data

In the final training phase of RAVEN++, we introduce an active reinforcement learning (RL) stage, which leverages a small amount of high-quality, finely annotated data to perform targeted replay training on valuable samples. This stage is designed to maximize data efficiency while significantly improving the model's performance. During training, we interleave supervised fine-tuning (SFT) and RL, maintaining two separate data buffers for each approach. This strategy is motivated by our observation that RL excels at maintaining and improving performance on tasks within the model's current capabilities, while SFT is more effective at enabling progress on tasks beyond the model's current scope, which is also articled in (Ma et al., 2025; Chu et al.).

Operation: To operationalize this, we dynamically sample and group training examples based on their complexity and the model's performance, assigning them to either the SFT or RL buffer as follows:

- **Fundamental Knowledge Gaps**: Samples where the model fails to identify violations or misclassifies the major category are assigned to the SFT buffer. These cases represent fundamental knowledge gaps that require targeted supplementation to address the model's inability to recognize violations or major categories.
- **Refinement of Existing Capabilities**: Samples where the model correctly identifies the major category but misclassifies the sub-category, exhibits low IoU for sub-scene localization, or deviates significantly from Qwen-72B's reasoning are assigned to the RL buffer with larger weights. These cases represent tasks within the model's current scope but require refinement in sub-category classification, temporal grounding, or reasoning consistency.
- **Standard Cases**: Samples that do not fall into the above categories are processed using normal RL execution but are assigned a very small weight in the reward function. This ensures that the model maintains its performance on tasks it already handles well, without overfitting to less challenging examples. This dynamic sampling strategy ensures that SFT is used for knowledge supplementation on tasks

Method	Marketing Exaggerate		Discomforting Content		Vulgar Content		Requiring Credential Review		Prohibited Goods/Services		Average Major Category	
	Cate.(P/R)	Gro.	Cate.(P/R)	Gro.	Cate.(P/R)	Gro.	Cate.(P/R)	Gro.	Cate.(P/R)	Gro.	Cate.(P/R)	Gro.
Small Models	0.681/0.532	-	0.707/0.679	-	0.667/0.654	-	0.711/0.687	-	0.721/0.734	-	0.697/0.657	-
LLaVA -v1.5-SFT	0.796/0.756	0.398	0.798/0.772	0.385	0.771/0.799	0.400	0.754/0.701	0.432	0.789/0.761	0.567	0.782/0.758	0.436
Qwen2.5-VL -7B-SFT	0.832/0.787	0.424	0.821/0.798	0.402	0.800/0.810	0.411	0.773/0.702	0.461	0.797/0.771	0.580	0.805/0.774	0.456
RAVEN	0.851/0.801	0.521	0.843/0.812	0.477	0.810/0.831	0.565	0.802/0.713	0.541	0.825/0.784	0.669	0.826/0.788	0.555
RAVEN++	0.905/0.840	0.601	0.889/0.859	0.562	0.862/0.870	0.650	0.873/0.738	0.671	0.859/0.832	0.741	**0.878/0.828**	**0.647**

Table 1: Performance of Violation Major Category (Precision/Recall) and Violation Temporal Grounding (mIoU) on Industrial Dataset. “Cate." indicates “Category”, and “Gro." indicates “Grounding”.

Method	Marketing Exaggerate	Discomforting Content	Vulgar Content	Requiring Credential Review	Prohibited Goods/Services	Average Sub-Category (P/R)
RAVEN	0.638/0.585	0.601/0.590	0.588/0.611	0.586/0.567	0.596/0.579	0.602/0.586
RAVEN++	0.735/0.654	0.705/0.622	0.668/0.652	0.659/0.607	0.682/0.704	**0.690/0.650**

Table 2: Performance of Violation Sub-Category (Precision/Recall) on Industrial Dataset.

that exceed the model’s current capabilities, while RL is applied to refine performance on tasks within the model’s current scope.

Training Details and Data Efficiency: The active RL stage places a high emphasis on data quality, as it relies on a small number of finely annotated samples. To avoid overfitting and ensure stable learning, we use a small learning rate during this phase. When the buffer accumulates enough challenging questions to form a training batch, we perform an SFT or RL step using these examples. This ensures that the model is trained on a diverse and representative set of challenging cases, enabling it to address its weaknesses effectively. Remarkably, even with this limited data, we observe rapid performance improvements, demonstrating the data efficiency of this approach. This efficiency is attributed to the targeted replay training, which focuses on the most valuable and challenging samples, enabling the model to quickly address its weaknesses and enhance its strengths.

4 Deployment

As shown in Fig. 2, based on the deployment of RAVEN, RAVEN++ integrates difficult positive and negative samples identified by online reviewers into the Active RL stage, enabling incremental learning and continuous model improvement. This enhancement allows the system to adapt dynamically to emerging violation patterns and improve its precision over time.

Method	Average	
	Cate. (P/R)	Gro.
LLaVA-v1.5-SFT	0.509/0.501	0.370
Qwen2.5-VL-7B-SFT	0.537/0.517	0.384
RAVEN	0.551/0.530	0.435
RAVEN++	**0.584/0.551**	**0.487**

Table 3: Performance of Violation Category (Precision/Recall) and Violation Temporal Grounding (mIoU) on Public MultiHateClip Dataset.

5 Experiments and Results

We conduct extensive experiments from both offline testing and online testing, utilizing both public dataset and practical industrial dataset. For fair comparison, we follow the dataset settings in RAVEN.

5.1 Datasets

Our experimental setup follows the dataset settings in RAVEN (Ji et al., 2025c). The industrial dataset consists of 38,000 training videos (with mixed precise and coarse annotations) and 5,000 precisely annotated test videos. All annotations adhere to the same six major violation categories and temporal interval labels defined in RAVEN. For public testing, we use the available subset of the MultiHateClip (Wang et al., 2024) dataset, consistent with the RAVEN benchmark.

Model	Online Sample Average	
	Cate.(P/R)	Gro.
Small Models	0.711/0.668	-
Qwen2.5-VL-7B-SFT	0.800/0.787	0.478
RAVEN	0.821/0.803	0.563
RAVEN++	**0.873/0.847**	**0.662**

Table 4: A/B Test on the Online Serving.

Method	In-Domain (Average Gro.)	Out-of-Domain (Average Gro.)
Qwen2.5-VL-7B-SFT	0.433	0.246
RAVEN	0.546	0.408
RAVEN++	0.631	0.463

Table 5: Study on Generalization Capabilities.

Multi-Task SFT	Passive RL	Active RL	Gro.
✓			0.461
✓	✓		0.587
✓	✓	✓	0.647

Table 6: Study on Progressive Training Stages.

Tversky Distance Reward	Reasoning Reward	Average Major Category (P/R)
✓		0.869/0.820
	✓	0.860/0.812
✓	✓	0.878/0.828

Table 7: Study on Reward Functions.

5.2 Offline Testing

We evaluate RAVEN++ against several baselines, including LLaVA-v1.5 (Liu et al., 2023), Qwen2-VL-7B (Bai et al., 2023b), Qwen2.5-VL-7B (Bai et al., 2023b), and RAVEN. The results in Table 1, 2 and 3 show that RAVEN++ outperforms all baselines in category accuracy and grounding precision, achieving significant improvements over RAVEN in sub-scene interval localization and reasoning consistency. These gains highlight the effectiveness of RAVEN++'s active RL stage. Furthermore, we also perform independent-samples t-tests, which confirm that the performance differences between RAVEN and RAVEN++ are statistically significant ($p < 0.02$).

5.3 Online A/B Testing

To evaluate real-world applicability, we perform a day-long online A/B test on a live business platform, adhering to the framework established by RAVEN. By dedicating 20% of the total traffic for assessment, we compare RAVEN++ with RAVEN, a compact legacy model, and Qwen2.5-VL-7B-SFT. As shown in Table 4, RAVEN++ delivers substantial gains in identifying violative videos, achieving higher precision and recall in category detection compared to the legacy model. Moreover, RAVEN++ achieves an 9.9% improvement over RAVEN in temporal interval localization accuracy.

5.4 Study on Generalization Capabilities

To further evaluate the generalization ability of RAVEN++, we follow the experimental design of RAVEN, training the model on three in-domain categories (Discomforting Content, Marketing Exaggeration, Requiring Credential Review) and testing it on two out-of-domain categories (Vulgar Content, Prohibited Goods/Services). The results, presented in Table 5, reveal that RAVEN++ consistently achieves higher accuracy and better generalization compared to both the Qwen2.5-VL SFT model and RAVEN.

5.5 Study on Progressive Training Stages

Table 6 presents ablation studies on the three-stage training process used in this work. The results show that RAVEN++ achieves steady and significant performance improvements with each progressive training stage. Notably, the final active RL stage boosts grounding performance by 6% compared to the previous stage.

5.6 Study on Reward Functions

RAVEN++ introduces two novel reward functions, Tversky Distance based and Reasoning Reward, which are evaluated through ablation studies on the Industrial dataset. As shown in Table 7, both functions significantly enhance performance, validating their effectiveness compared to RAVEN.

6 Conclusion

We address the critical challenges of video Ad moderation by introducing RAVEN++, a novel framework that significantly enhances fine-grained violation understanding, reasoning capabilities, and generalization ability. Through the integration of Active RL, Fine-Grained Violation Understanding, and Progressive Multi-Stage Training, RAVEN++ achieves superior performance on both offline and online deployment scenarios.

7 Limitations

A key limitation of this work is its exclusive focus on the video modality for Ad content moderation. Real-world advertising is inherently multi-modal, combining text, images, and video. Our current approach, which analyzes video in isolation, cannot capture cross-modal inconsistencies—such as misleading text overlaid on benign visuals. This restricts its applicability to video-dominant scenarios. However, this constraint outlines a clear path for future work. Our method provides a foundation for extension towards a unified multi-modal framework. Future efforts will focus on integrating text and image analysis to build a holistic system for robust, comprehensive ad moderation.

8 Ethical Statement

Our research adheres to ethical principles and prioritizes user rights. The dataset samples are for scientific analysis only and do not reflect the authors' views. All resources are intended for scientific research purposes only, contributing to the development of more secure and reliable digital platforms.

References

Shelly Rathee and Tyler Milfeld. Sustainability advertising: literature review and framework for future research. *International Journal of Advertising*, 43(1): 7–35, 2024.

Colin Campbell, Sean Sands, Brent McFerran, and Alexis Mavrommatis. Diversity representation in advertising. *Journal of the Academy of Marketing Science*, 53(2):588–616, 2025.

Enkai Ji, Yihan Wang, Suchuan Xing, and Jianian Jin. Hierarchical reinforcement learning for energy-efficient api traffic optimization in large-scale advertising systems. *IEEE Access*, 2025a.

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

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023a.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023b.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. *arXiv preprint arXiv:2402.18476*, 2024a.

Lanyun Zhu, Tianrun Chen, Deyi Ji, Jieping Ye, and Jun Liu. Llafs: When large language models meet few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3065–3075, 2024b.

Lanyun Zhu, Tianrun Chen, Deyi Ji, Peng Xu, Jieping Ye, and Jun Liu. Llafs++: Few-shot image segmentation with large language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025a.

Wei Qiao, Tushar Dogra, Otilia Stretcu, Yu-Han Lyu, Tiantian Fang, Dongjin Kwon, Chun-Ta Lu, Enming Luo, Yuan Wang, Chih-Chun Chia, et al. Scaling up llm reviews for google ads content moderation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 1174–1175, 2024.

Leonardo Madio and Martin Quinn. Content moderation and advertising in social media platforms. *Journal of Economics & Management Strategy*, 34(2): 342–369, 2025.

Kamaruddin Hasan and Hamdan Juhannis. Religious education and moderation: A bibliometric analysis. *Cogent Education*, 11(1):2292885, 2024.

Barween Al Kurdi and Muhammad Turki Alshurideh. The effect of social media influencer traits on consumer purchasing decisions for keto products: examining the moderating influence of advertising repetition. *Journal of Marketing Communications*, 31(4): 422–443, 2025.

Tae Hyun Baek, Jungkeun Kim, and Jeong Hyun Kim. Effect of disclosing ai-generated content on prosocial advertising evaluation. *International Journal of Advertising*, pages 1–22, 2024.

De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant. In *European Conference on Computer Vision*, pages 202–218. Springer, 2024.

Deyi Ji, Feng Zhao, Hongtao Lu, Feng Wu, and Jieping Ye. Structural and statistical texture knowledge distillation and learning for segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025b.

Deyi Ji, Feng Zhao, Hongtao Lu, Mingyuan Tao, and Jieping Ye. Ultra-high resolution segmentation with ultra-rich context: A novel benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23621–23630, 2023.

Shimin Chen, Xiaohan Lan, Yitian Yuan, Zequn Jie, and Lin Ma. Timemarker: A versatile video-llm for long and short video understanding with superior temporal localization ability. *arXiv preprint arXiv:2411.18211*, 2024.

Xin Gu, Heng Fan, Yan Huang, Tiejian Luo, and Libo Zhang. Context-guided spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18330–18339, 2024.

Deyi Ji, Haoran Wang, Mingyuan Tao, Jianqiang Huang, Xian-Sheng Hua, and Hongtao Lu. Structural and statistical texture knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16876–16885, 2022.

Deyi Ji, Yuekui Yang, Haiyang Wu, Shaoping Ma, Tianrun Chen, and Lanyun Zhu. RAVEN: Robust advertisement video violation temporal grounding via reinforcement reasoning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Industry Track)*, 2025c.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Krishanu Maity, Poornash Sangeetha, Sriparna Saha, and Pushpak Bhattacharyya. Toxvidlm: A multimodal framework for toxicity detection in code-mixed videos. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11130–11142, 2024.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.

Deyi Ji, Wenwei Jin, Hongtao Lu, and Feng Zhao. Pptformer: pseudo multi-perspective transformer for uav segmentation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 893–901, 2024a.

Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024a.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. Large language models for generative information extraction: A survey. *Frontiers of Computer Science*, 18(6):186357, 2024b.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Haiyang Wu, De Wen Soh, and Jun Liu. Cpcf: A cross-prompt contrastive framework for referring multimodal large language models. In *Forty-second International Conference on Machine Learning*.

Deyi Ji, Feng Zhao, Lanyun Zhu, Wenwei Jin, Hongtao Lu, and Jieping Ye. Discrete latent perspective learning for segmentation and detection. In *International Conference on Machine Learning*, pages 21719–21730. PMLR, 2024b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Lanyun Zhu, Tianrun Chen, Qianxiong Xu, Xuanyi Liu, Deyi Ji, Haiyang Wu, De Wen Soh, and Jun Liu. Popen: Preference-based optimization and ensemble for lvlm-based reasoning segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30231–30240, 2025b.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816, 2024.

Afra Amini, Tim Vieira, and Ryan Cotterell. Direct preference optimization with an offset. *arXiv preprint arXiv:2402.10571*, 2024.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. Preference ranking optimization for human alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18990–18998, 2024.

Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu Liu. Statistical rejection sampling improves preference optimization. In *The Twelfth International Conference on Learning Representations*.

Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E Taylor, and Peter Stone. Curriculum learning for reinforcement learning domains: A framework and survey. *Journal of Machine Learning Research*, 21(181):1–50, 2020.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

Alex Graves, Marc G Bellemare, Jacob Menick, Remi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. In *international conference on machine learning*, pages 1311–1320. Pmlr, 2017.

Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. In *International conference on machine learning*, pages 2535–2544. PMLR, 2019.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum learning: A survey. *International Journal of Computer Vision*, 130(6):1526–1565, 2022.

Anastasia Pentina, Viktoriia Sharmanska, and Christoph H Lampert. Curriculum learning of multiple tasks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5492–5500, 2015.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Deyi Ji, Lanyun Zhu, Siqi Gao, Peng Xu, Hongtao Lu, Jieping Ye, and Feng Zhao. Tree-of-table: Unleashing the power of llms for enhanced large-scale table understanding. *arXiv preprint arXiv:2411.08516*, 2024c.

Arkady Epshteyn, Adam Vogel, and Gerald DeJong. Active reinforcement learning. In *Proceedings of the 25th international conference on Machine learning*, pages 296–303, 2008.

Jinghuan Shang and Michael S Ryoo. Active vision reinforcement learning under limited visual observability. *Advances in Neural Information Processing Systems*, 36:10316–10338, 2023.

Mahi Kolla, Siddharth Salunkhe, Eshwar Chandrasekharan, and Koustuv Saha. Llm-mod: Can large language models assist content moderation? In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2024.

Deepak Kumar, Yousef Anees AbuHashem, and Zakir Durumeric. Watch your language: Investigating content moderation with large language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 865–878, 2024.

Lindsay Blackwell. Content moderation futures. *arXiv preprint arXiv:2509.09076*, 2025.

Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Amos Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977.

Lu Ma, Hao Liang, Meiyi Qiang, Lexiang Tang, Xiaochen Ma, Zhen Hao Wong, Junbo Niu, Chengyu Shen, Runming He, Bin Cui, et al. Learning what reinforcement learning can't: Interleaved online fine-tuning for hardest questions. *arXiv preprint arXiv:2506.07527*, 2025.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. In *Forty-second International Conference on Machine Learning*.

Han Wang, Tan Rui Yang, Usman Naseem, and Roy Ka-Wei Lee. Multihateclip: A multilingual benchmark dataset for hateful video detection on youtube and bilibili. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7493–7502, 2024.