$\mathcal{R}^3$: Advertisement Compliance $\mathcal{R}$ectification via Group-$\mathcal{R}$elative Experience Extractor and Curriculum $\mathcal{R}$einforcement

Yuan Chen* **Zhenyu Hu*** **Mengge Xue†** **Te Cao** **Liqun Liu‡**
Peng Shu **Huan Yu** **Jie Jiang**
Tencent
{izayoiychen, mapleshu, berryxue, rosaliecao, liqunliu}@tencent.com
{archershu, huanyu, zeus}@tencent.com

Abstract

Rigorous content moderation is crucial for online advertising but leads to millions of daily rejections. This scale renders manual rectification infeasible, particularly for video advertisements. However, existing safety-driven methods often suffer from aggressive over-editing, which compromises the advertiser's original semantic intent merely to satisfy compliance. In this work, we target the rectification of textual violations in video ads, covering both speech transcripts and on-screen text. We propose $\mathcal{R}^3$, a novel framework designed to harmonize compliance with original semantic intent preservation. Our approach integrates three key innovations: (1) an experience-driven data synthesis framework that bootstraps high-quality supervision via group-**R**elative compliance experience extractor; (2) a curriculum **R**einforcement learning strategy with hierarchical rewards designed to enforce compliance while maximizing semantic consistency; and (3) a comprehensive video **R**ectification framework seamlessly integrating text recognition, rewriting, and re-rendering for industrial deployment. Extensive experiments on industrial datasets and online A/B testing demonstrate that $\mathcal{R}^3$ significantly outperforms state-of-the-art baselines, achieving an optimal trade-off between violation rectification and intent preservation.

1 Introduction

Advertising serves as a cornerstone of the digital economy, acting as the primary engine for revenue and growth across online platforms (Rathee and Milfeld, 2024; Campbell et al., 2025). In pursuit of strict regulatory compliance and user safety, these platforms impose rigorous content moderation policies (Ji et al., 2025b,a; Madio and Quinn, 2025). However, advertisers often struggle to navigate the complexity of these moderation rules, resulting in millions of advertisements being rejected daily. Consequently, it is imperative for online platforms to assist advertisers in automatically rectifying ad material, thereby unlocking ad supply and enhancing the overall advertiser experience (Xia et al., 2025). With the advent of the 5G era, video has emerged as the predominant medium for information consumption, a trend particularly evident in online advertising. Consequently, violations within video advertisements constitute a significant proportion of overall content compliance issues. While these violations span both visual and textual modalities, this work focus specifically on the rectification of textual violations within video ads.

While recent advancements in Large Language Models (LLMs) (Touvron et al., 2023; Yang et al., 2025; OpenAI, 2023) have bolstered capabilities in content moderation and compliance rectification (Pi et al., 2024; Laugier et al., 2021), directly applying them to video ad rectification remains non-trivial. Specifically, deploying general-purpose or naively fine-tuned models for this task faces significant challenges: 1) **Inadequate compliance**: Ad moderation policies are voluminous and highly context-dependent, thus that general-purpose models often fail to grasp the nuanced boundaries of moderation rules via direct prompting, leading to frequent hallucinations or missed detections of subtle violations. This necessitates domain-specific alignment, yet standard Supervised Fine-Tuning (SFT) is hindered by data availability; 2) **Prohibitive Annotation Costs**: The prerequisite data annotation phase for supervised fine-tuning is severely constrained by the complexity and rapid evolution of ad content moderation rules. The substantial cognitive burden involved in comprehending these policies, coupled with the resulting low annotation consistency, renders large-scale manual annotation practically infeasible (Ji et al., 2025b). 3) **Compromised semantic intent**:

*Equal Contribution.
†Project Leader.
‡Corresponding Author.

violation content rewriting is inherently a multi-objective optimization task that demands balancing compliance, semantic intent preservation, and coherence. Naively fine-tuned models, however, tend to over-optimize for compliance, leading to excessive alterations that deviate from the original intent and diminish the advertising effectiveness.

To address the aforementioned challenges, we introduce $\mathcal{R}^3$, a comprehensive video rectification framework designed to rectify textual violations in video advertisements, covering both speech transcripts and on-screen text. By harmonizing strict compliance with the preservation of the advertiser's original semantic intent, $\mathcal{R}^3$ automates the full lifecycle of video rectification—from multi-modal content extraction to final re-rendering. Our main contributions are summarized as follows:

1) Experience-driven Data Synthesis: We propose an experience-driven synthesis framework that bootstraps high-quality supervision from advanced LLMs. By introducing group-relative compliance experience extractor, it effectively addresses data scarcity and covers complex edge cases where naive prompting typically fails.

2) Curriculum RL with Hierarchical Rewards: The task of violation textual content rectification demands the simultaneous satisfaction of multiple, often competing objectives: compliance, semantic intent preservation, and coherence. Optimizing these concurrently poses significant challenges for standard Reinforcement Learning (RL) paradigms. We address this by introducing a curriculum (Bengio et al., 2009; Ko et al., 2022) reinforcement learning framework with hierarchical rewards. This strategy progressively optimizes the model to navigate the intrinsic conflict between enforcing strict compliance and maintaining semantic intent, avoiding the over-editing in existing methods.

3) Holistic Video Rectification Framework: We present a non-compliant video ads rectification system validated within a live online advertising platform. Going beyond violation text rewriting, we engineer a pipeline that seamlessly integrates text recognition, rewriting, and re-rendering for industrial-scale deployment. This enables the automated rectification of non-compliant videos while preserving their original audio-visual fidelity.

We evaluated $\mathcal{R}^3$ using industry datasets and verified its effectiveness on online advertisement platform. Comparative analysis reveals that our system significantly outperforms general-purpose models-ranging from the open-source Qwen3 (Yang et al., 2025) to the advanced Gemini3-Flash (Google, 2025). Crucially, $\mathcal{R}^3$ demonstrates a superior performance to balance compliance with semantic intent preservation, resulting in a significantly improved final successful rectification rate.

2 Related Work

2.1 Alignment for Large Language Models

RL has become the standard for LLM alignment (Christiano et al., 2017), with recent advances like Group Relative Policy Optimization (GRPO) (Shao et al., 2024) improving efficiency by estimating advantages from response groups rather than separate value functions. Beyond general alignment, RL has shown promise in complex decision-making tasks such as content moderation (Ji et al., 2025a,b). Despite these successes, existing methods (Wang et al., 2021; Cai et al., 2023) primarily focus on open-ended or discriminative tasks. In our work, we apply RL to video rectification, which bridges this gap by designing a sophisticated, curriculum reward mechanism.

2.2 Generative AI and Content Moderation in Online Advertising

Generation works leverage generative models to maximize commercial metrics (Deng et al., 2025; Zhang et al., 2025a; Wu et al., 2024), often creating content from scratch rather than preserving existing intent. Meanwhile, Ji et al. (2025a) establish robust baselines for risk detection and localization but stop short of correction. This creates a disjointed industrial pipeline where rejected ads lack automated rectification. Our work bridges this gap. Distinct from over-editing methods like (Zhang et al., 2025b), we employ curriculum RL to rewrite violative content, achieving an optimal trade-off between compliance and intent preservation.

2.3 Training-Free Methods

Existing inference-time methods, including in-context learning (ICL)(Brown et al., 2020), iterative refinement(Madaan et al., 2023; Shinn et al., 2023), and feedback-driven frameworks (Song et al., 2025; Yuksekgonul et al., 2025; Monea et al., 2024), typically target within-sample improvements. Compared to the hierarchical, off-policy design of Agent KB (Tang et al., 2025), Training-Free GRPO (Cai et al., 2025) adopts a global perspective, refining a shared experience library via multi-epoch learning akin to traditional

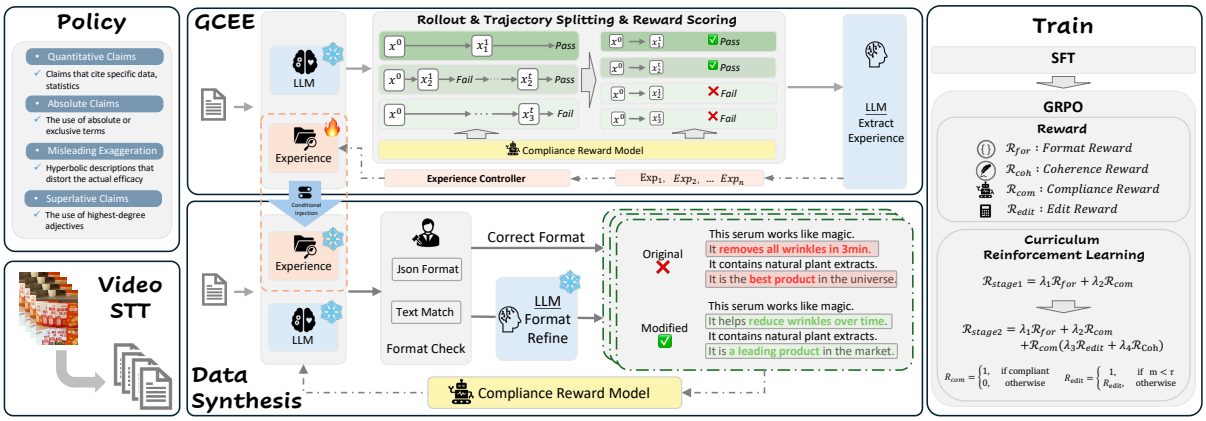


Figure 1: **Overview of $\mathcal{R}^3$.** Taking non-compliant video ads and violation policies as input, the **Experience-driven Data Synthesis** employs the **Group-Relative Compliance Experience Extractor (GCEE)** to extract compliance experience from rectification trajectories for high-quality supervision. The model is initialized via supervised fine-tuning and further optimized using a **Curriculum Reinforcement Learning** strategy with hierarchical rewards.

RL. Inspired by Training-Free GRPO, our approach performs task-aware trajectory split and extracts experience by contrasting successful and failed trajectories.

3 Method

3.1 Problem Formulation

We formulate violation textual content rectification as a constrained rewriting task defined by the tuple $\mathcal{I} = (x, \mathcal{G})$. Here, x represents the non-compliant video textual content, $\mathcal{G}$ denotes the violation policies provided by the online violation detection model. We segment x into sentence units $\mathcal{S}(x) = [u_1, u_2, \ldots, u_n]$ using standard punctuation-based splitting and train a policy $\pi_\theta(y|x, \mathcal{G})$ to generate a structured edit list $y = \{(u_k, v_k) \mid u_k \in \mathcal{N}\}$, where $\mathcal{N} \subset \mathcal{S}(x)$ represents the subset of sentences identified as violations.Specifically, each pair in y indicates that a non-compliant sentence u_k is substituted by its compliant revision v_k. The final rectified text x' is obtained via a deterministic mapping $\mathcal{A}(x, y)$, which applies these substitutions to the original sequence. Our objective is to produce a rectified output x' that ensures compliance, while preserving the semantic intent of x by optimizing for linguistic coherence and minimizing edits.

3.2 Experience-driven Data Synthesis

Manual rectification of non-compliant ad material is labor-intensive, time-consuming, and unscalable. While Advanced LLMs possess strong constrained rewriting capabilities, they often exhibit inadequate compliance, lacking the nuanced understanding of specific moderation policies. To bridge this gap, we propose an automated, experience-driven data synthesis framework that bootstraps high-quality supervision. Specifically, tractable samples are efficiently processed by an advanced LLM Gemini3-Flash, while for intractable samples where direct prompting fails, we incorporate a Group-relative Compliance Experience Extractor.

Group-relative Compliance Experience Extractor. This module is an in-context reinforcement learning paradigm inspired by Training-Free GRPO (Cai et al., 2025). Adopting the core GRPO mechanism, it performs rollouts and computes group advantages, yet distinguishes itself by learning through natural language feedback rather than parameter updates. Operating within an iterative refinement framework (starting from $x^0 = x$ to generate $x^{t+1} = \mathcal{A}(x^t, y^t)$), the extractor pinpoints the pivotal violation-to-compliance transition—where a persistent non-compliant x^t becomes compliant x^{t+1}. It constructs a high-advantage contrastive pair by comparing the failed trajectory $x^0 \to x^t$ with the successful trajectory $x^0 \to x^{t+1}$, prompting the LLM to explicitly articulate the semantic rationale behind the correction. This experience is stored in a dynamic buffer, where an LLM-based controller compares newly extracted experiences with existing entries and decides whether to keep, revise, or discard them to reduce redundancy and conflicts. The refined experience are then injected into subsequent generations. An example can be found in Appendix B.3.

Conditional Experience Injection. However, the extracted experience inherently prioritizes compliance, often biasing the model toward aggressive rewriting that deviates from the original semantic intent. To resolve this trade-off, as shown in Figure 1, we implement a conditional injection mechanism. We first employ a direct prompting strategy using Gemini3-Flash, activating experience injection only if the initial attempt fails. This strategy prioritizes minimal edits on tractable instances and reserves experience injection for intractable ones, thereby producing a high-coverage training corpus that strictly adheres to multi-objectives.

3.3 Training

We train $\mathcal{R}^3$ on top of Qwen3-8B by integrating SFT with GRPO. Specifically, the SFT stage serves to instill moderation policies and align the model with the required structured output format. Subsequently, GRPO further optimizes for compliance and semantic intent preservation using curriculum RL with hierarchical rewards.

3.3.1 Supervised Fine-Tuning

SFT enforce the edit-list schema and condition rectifications. This stage minimizes format violations and injects foundational compliance knowledge, establishing a stable initial state to facilitate efficient exploration in the subsequent RL stage.

3.3.2 Rewards Design

The reward function R optimizes the model's RL performance across four dimensions.

Format Reward (R_{for}). This term enforces structural executability. Beyond JSON parsability, every u_k selected for modification must exactly match the source text (i.e., $u_k \in \mathcal{S}(x)$).

Compliance Reward (R_{com}). We measure compliance using the online violation detection model. The reward is binary: $R_{com} \in \{0, 1\}$. This term acts as the primary driver for satisfying compliance.

Minimal-edit Reward (R_{edit}). We quantify modification magnitude by the count of edited units $m = |y|$ relative to $n = |\mathcal{S}(x)|$. Minimizing m is crucial for: (1) Advertisers prioritize retaining the original narrative structure and minimizing interventions; (2) Limiting text changes reduces duration mismatches in downstream Text-To-Speech (TTS), preventing synchronization failures. However, an unbounded penalty on m creates an optimization imbalance: over-optimizing minimalism on tractable samples penalizes the exploration required for intractable ones. Thus, we introduce a tolerance threshold τ as an edit margin to establish a penalty-free zone for essential modifications:

$$R_{\text{edit}} = \begin{cases} 1, & m \leq \tau, \\ 1 - \frac{m-\tau}{n-\tau}, & m > \tau. \end{cases} \tag{1}$$

Coherence Reward (R_{coh}) To ensure linguistic fluency and semantic preservation, we employ an LLM-as-a-judge to provide dense feedback. For each pair (u_k, v_k), the judge assigns a scalar score $\phi(x, u_k, v_k) \in \{0, 1\}$. The final sentence-level reward is the average of these scores:

$$R_{\text{coh}} = \frac{1}{m} \sum_{k=1}^{m} \phi(x, u_k, v_k). \tag{2}$$

The overall reward R is:

$$R = \lambda_1 R_{for} + \lambda_2 R_{com} + R_{com} \cdot (\lambda_3 R_{edit} + \lambda_4 R_{coh}), \tag{3}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are stage-dependent coefficients.Specifically, in Stage 1 we set $(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (0.3, 5.0, 0, 0)$, and in Stage 2 we use $(0.1, 5, 0.4, 0.5)$. To prevent reward hacking, we gate R_{edit} and R_{coh} using the compliance reward R_{com}. This ensures that auxiliary objectives are only pursued once the primary compliance goal is met.

3.3.3 Curriculum RL with Hierarchical Rewards

While SFT provides a strong initialization, it relies on static supervision, which struggles to balance the often conflicting objectives of strict compliance and semantic preservation. To address this, we employ RL with a hierarchical reward system. However, directly optimizing for these competing goals remains challenging: if edit penalties are introduced too early, the policy tends to collapse into a local optimum of minimal edits to avoid punishment, failing to explore necessary but extensive rewrites. To overcome this exploration hurdle, we design a two-stage curriculum:

Stage 1: Compliance Alignment. In the initial stage, we exclusively prioritize compliance to encourage bold exploration. We set the coefficients λ_3 and λ_4 to zero while assigning a high value to λ_2. This configuration compels the policy model to focus solely on satisfying compliance, incentivizing extensive rewriting. While this may temporarily lead to over-editing, it establishes a critical

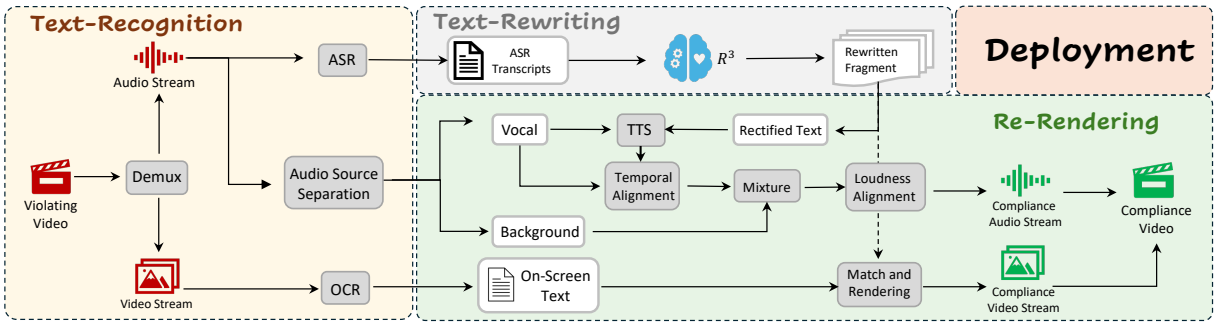


Figure 2: The deployment workflow of $\mathcal{R}^3$. Illustrating the automated pipeline for rectification

high-recall baseline, ensuring the model learns the necessary editing to achieve high compliance rate.

Stage 2: Quality Refinement. Once the model achieves a stable compliance rate, we transition to stage 2 by activating the soft objectives ($\lambda_3, \lambda_4 > 0$) while maintaining λ_2 as the dominant term. This stage acts as a regularization step: within the manifold of compliant solutions, the policy model is guided to select those that minimize text alterations while preserving coherence. This effectively refines the aggressive editing behavior learned in stage 1, converging toward the Pareto frontier of compliance and semantic intent preservation.

4 Deployment

To bridge the gap between algorithmic rewriting and industrial production, we integrate $\mathcal{R}^3$ into an automated video rectification workflow (Figure 2). The pipeline comprises three phases: (1) Text-Recognition, which demultiplexes the video, isolates vocals via audio source separation, and extracts both speech transcripts and on-screen text; (2) Text-Rewriting, where $\mathcal{R}^3$ generates a structured edit list to ensure compliance; and (3) Re-Rendering, which reconstructs both auditory and visual content.

The final Re-Rendering stage is critical for preserving the quality of the rectified video.We employ zero-shot voice cloning to preserve the speaker's timbre. A key challenge is temporal alignment between synthesized speech and the original video timeline. We therefore first adjust silent intervals and, when necessary, apply time-stretching to the generated speech so that the updated segment matches the target duration. For loudness alignment, we employ an iterative dynamic range control procedure: we first estimate the loudness gap between the synthesized speech and the original vocal track, then apply gain adjustment with peak limiting, and iteratively compensate for the attenuation introduced by the limiter until the mixed audio reaches the target loudness. Finally, subtitles are re-rendering using OCR-derived coordinates and the updated audio is merged back into the video, producing a compliant video that preserves the original production layout and audio-visual fidelity.

5 Experiments

5.1 Experimental Setup

5.1.1 Datasets

To assess the efficacy of $\mathcal{R}^3$ within a realistic industrial setting, we constructed a proprietary dataset derived from a production moderation pipeline. The dataset is partitioned into a training set of 11k samples and a held-out test set of 1k samples. To ensure a rigorous evaluation of robustness across challenging compliance scenarios, we focus on four major violation policies: **Quantitative Claims (Qua.C)**, **Absolute Claims (Abs.C)**, **Misleading Exaggeration (Mis.E)**, and **Superlative Claims (Sup.C)**, detailed definitions and examples are provided in the Appendix A.2.

5.1.2 Metrics

To evaluate the performance of our method, we employ four metrics: (1) **Compliance Rate (ComR)**: The percentage of rectified samples which are compliant. (2) **Average Edit (AvgE)**: The average number of modified sentence units per example. (3) **Coherence Rate (CohR)**: The proportion of samples deemed linguistically fluent and semantically consistent. We employ Gemini3-Flash as an evaluator, instructing it to compare the rectified content against the original text. (4) **Qualified Rectification Rate (QRR)**: The fraction of samples that are both compliant and coherent. In practice, advertiser tolerance for text alterations is highly subjective, making a universal threshold for acceptable edits

Model	Qua.C ComR↑	Abs.C ComR↑	Mis.E ComR↑	Sup.C ComR↑	Average			
					ComR↑	AvgE↓	CohR↑	QRR↑
Qwen3-8B	69.32%	67.95%	31.10%	70.88%	50.28%	8.2	45.07%	22.66%
Gemini3-Flash	83.67%	86.32%	65.99%	96.70%	77.05%	**6.24**	**97.98%**	74.43%
Gemini3-Flash with GCEE	90.04%	90.17%	75.58%	98.35%	84.84%	8.63	89.41%	75.21%
Qwen3-8B-SFT	91.63%	94.87%	69.77%	98.90%	82.58%	7.7	87.48%	72.23%
$\mathcal{R}^3$	**93.63%**	**96.58%**	**76.16%**	**98.90%**	**85.50%**	7.49	94.70%	**81.02%**

Table 1: **Performance comparison.** We report ComR on four violation policies. Additionally, we report the average performance on four metrics: ComR, AvgE, CohR, and QRR. $\mathcal{R}^3$ achieves the best performance.

impractical. Thus, QRR indicates whether a rectification is fundamentally usable, while AvgE independently measures the intervention cost required to achieve this qualification.

Model	Online Sample Average	
	ComR↑	AR↑
Qwen3-8B-SFT	83.91%	1.0
$\mathcal{R}^3$	86.53%	1.21

Table 2: Performance comparison on online A/B test. AR denotes the Advertiser Adoption Rate

SFT	R_{com}	R_{edit}	R_{coh}	Curr. RL	Average			
					ComR↑	AvgE↓	CohR↑	QRR↑
✓	-	-	-	-	82.58%	7.70	87.48%	72.23%
✓	✓	-	-	-	**89.94%**	11.28	86.14%	77.47%
✓	✓	✓	-	-	83.43%	**6.31**	87.95%	73.37%
✓	✓	✓	✓	-	83.14%	6.56	92.84%	77.19%
✓	✓	✓	✓	✓	85.50%	7.49	**94.70%**	**81.02%**

Table 3: Ablation study on different reward components and curriculum learning (Curr. RL) strategies.

5.2 Offline Testing

Table 1 presents a comprehensive evaluation of $\mathcal{R}^3$ against baselines, including the Qwen3-8B, Gemini3-Flash (with and without GCEE), and Qwen3-8B-SFT. $\mathcal{R}^3$ achieves state-of-the-art performance with the highest QRR and remarkably competitive AvgE. Specifically, compared to the Qwen3-8B-SFT, $\mathcal{R}^3$ yields a consistent improvement in ComR (+2.92%) and a substantial margin in QRR (+8.79%). This confirms that integrating curriculum GRPO effectively steers the policy toward the optimal Pareto frontier of strict compliance and semantic intent preservation. Furthermore, $\mathcal{R}^3$ ranks first across all four fine-grained violation policies. While Gemini3-Flash exhibits highly conservative editing behavior, $\mathcal{R}^3$ significantly outperforms it in overall quality, surpassing this baseline by +8.45% in ComR and +6.59% in QRR. Although augmenting Gemini3-Flash with GCEE boosts its compliance to 84.84%, it noticeably increases AvgE to 8.63 and limits the overall qualification. $\mathcal{R}^3$ maintains a definitive lead, validating the superiority of a dedicated rectification policy over directly prompting LLMs.

5.3 Online A/B Testing

To assess the practical viability of our approach in a production environment, we conducted a 3-day online A/B test on a live advertisement platform, benchmarking $\mathcal{R}^3$ against Qwen3-8B-SFT. We report adoption rates normalized against the baseline, as absolute figures are commercially sensitive. $\mathcal{R}^3$ yields substantial improvements (Table 2), with a +2.62% ComR gain and a relative adoption increase of +21%. Given that both models exhibit comparable AvgE, the boost in adoption aligns directly with $\mathcal{R}^3$'s higher QRR. This demonstrates that evaluating QRR in conjunction with AvgE effectively predicts real-world advertiser acceptance.

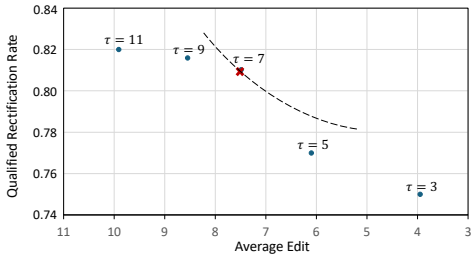


Figure 3: Impact of the tolerance threshold τ.

5.4 Ablation Study

5.4.1 Study on Rewards and Curriculum RL

We evaluate the contributions of distinct reward components and the curriculum strategy in Table 3. Regarding reward design, relying solely on the Compliance Reward guarantees compliance but induces excessive alterations. Integrating the Minimal-edit Reward effectively constrains this over-editing behavior, while the Coherence Reward

is indispensable for preserving linguistic fluency and semantic intent. Regarding the training strategy, comparing single-stage optimization against our two-stage curriculum confirms that the latter yields a superior trade-off.

5.4.2 Study on the Tolerance Threshold

Figure 3 depicts the impact of the tolerance threshold τ . We observe that a low τ overly restricts edits, compromising compliance, whereas an excessively high τ improves compliance but suffers from over-editing. This dynamic constitutes a Pareto optimality problem. As shown in Figure 3, setting $\tau = 7$ yields the most favorable equilibrium between rewriting intensity and compliance quality.

6 Conclusion

We present $\mathcal{R}^3$, a comprehensive video textual rectification framework designed to harmonize strict compliance with semantic intent preservation. By employing experience-driven data synthesis and curriculum reinforcement learning, $\mathcal{R}^3$ achieves state-of-the-art gains in both compliance rate and qualified rectification rate on industrial benchmarks. Furthermore, online A/B testing validates its practical viability and superior advertiser adoption in real-world production environments.

7 Limitations

Our system is optimized against a specific moderation system and a fixed set of policy guidelines used during training, which may reduce flexibility when the moderation system, decision boundary, or rule set changes over time. While our experience-driven data synthesis and curriculum reinforcement learning improve robustness within the covered policy scope, adapting to newly introduced or rapidly evolving rules may still require additional data regeneration and re-alignment. As future work, we plan to explore continual and modular policy alignment strategies that can rapidly incorporate rule updates with minimal re-training.

8 Ethical Statement

This research was conducted in strict adherence to ethical guidelines and data privacy regulations. The industrial dataset utilized in this study was derived from a production environment and was rigorously desensitized to ensure the anonymity of advertisers and users. The non-compliant advertisement examples presented herein are strictly for illustrative scientific analysis and do not reflect the views or values of the authors or the affiliated platform. All resources and methodologies are intended solely for academic research.

References

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Qingpeng Cai, Will Shiao, Jilong Xue, Li He, Kun Gai, Li Chen, and Peng Jiang. 2023. Constrained reinforcement learning for short video recommender systems. In *Proceedings of the ACM Web Conference 2023*, pages 1037–1047.

Yuzheng Cai, Siqi Cai, Yuchen Shi, Zihan Xu, Lichao Chen, Yulei Qin, Xiaoyu Tan, Gang Li, Zongyi Li, Haojia Lin, and 1 others. 2025. Training-free group relative policy optimization. *arXiv preprint arXiv:2510.08191*.

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

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30.

Jiaxin Deng, Shiyao Wang, Kuo Cai, Lejian Ren, Qigen Hu, Weifeng Ding, Qiang Luo, and Guorui Zhou. 2025. Onerec: Unifying retrieve and rank with generative recommender and iterative preference alignment. *arXiv preprint arXiv:2502.18965*.

Google. 2025. A new era of intelligence with gemini 3. https://blog.google/products/gemini/gemini-3.

Deyi Ji, Yuekui Yang, Liqun Liu, Peng Shu, Haiyang Wu, Shaogang Tang, Xudong Chen, Shaoping Ma, Tianrun Chen, and Lanyun Zhu. 2025a. RAVEN++: Pinpointing fine-grained violations in advertisement videos with active reinforcement reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1–10.

Deyi Ji, Yuekui Yang, Haiyang Wu, Shaoping Ma, Tianrun Chen, and Lanyun Zhu. 2025b. RAVEN: Robust advertisement video violation temporal grounding

via reinforcement reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 22–31.

Keonwoo Ko, Gu Jin, and Youngchul Sung. 2022. Curriculum offline reinforcement learning. In *International Conference on Learning Representations*.

Léo Laugier, John Pavlopoulos, Jeffrey Sorensen, and Lucas Dixon. 2021. Civil rephrases of toxic texts with self-supervised transformers. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1442–1461.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

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

Giovanni Monea, Antoine Bosselut, Kianté Brantley, and Yoav Artzi. 2024. Llms are in-context reinforcement learners. *arXiv e-prints*, pages arXiv–2410.

OpenAI. 2023. Gpt-4 technical report. Technical report, OpenAI.

Renjie Pi, Tianyang Han, Yueqi Xie, Rui Pan, Qing Lian, Hanze Dong, Jipeng Zhang, and Tong Zhang. 2024. Mllm-protector: Ensuring mllm's safety without hurting performance. *Preprint*, arXiv:2401.02906.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.

Kefan Song, Amir Moeini, Peng Wang, Lei Gong, Rohan Chandra, Shangtong Zhang, and Yanjun Qi. 2025. Reward is enough: Llms are in-context reinforcement learners. *arXiv preprint arXiv:2506.06303*.

Xiangru Tang, Tianrui Qin, Tianhao Peng, Ziyang Zhou, Daniel Shao, Tingting Du, Xinming Wei, Peng Xia, Fang Wu, He Zhu, and 1 others. 2025. Agent kb: Leveraging cross-domain experience for agentic problem solving. *arXiv preprint arXiv:2507.06229*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Kai Wang, Zhene Zou, Qilin Deng, Runze Wu, Jianrong Tao, Changjie Fan, Liang Chen, and Peng Cui. 2021. Rl4rs: A real-world benchmark for reinforcement learning based recommender system. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 5660–5669. IEEE.

Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, and 1 others. 2024. A survey on large language models for recommendation. *World Wide Web*, 27(5):60.

Yifan Xia, Guorui Chen, Wenqian Yu, Zhijiang Li, Philip Torr, and Jindong Gu. 2025. Reimagining safety alignment with an image. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9589–9603.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Pan Lu, Zhi Huang, Carlos Guestrin, and James Zou. 2025. Optimizing generative ai by backpropagating language model feedback. *Nature*, 639(8055):609–616.

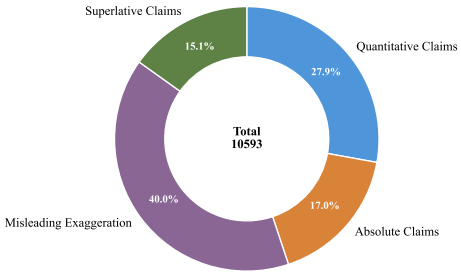
Jun Zhang, Yi Li, Yue Liu, Changping Wang, Yuan Wang, Yuling Xiong, Xun Liu, Haiyang Wu, Qian Li, Enming Zhang, and 1 others. 2025a. Gpr: Towards a generative pre-trained one-model paradigm for large-scale advertising recommendation. *arXiv preprint arXiv:2511.10138*.

Ruiyang Zhang, Jiahao Luo, Xiaoru Feng, Qiufan Pang, Yaodong Yang, and Juntao Dai. 2025b. Safeeditor: Unified mllm for efficient post-hoc t2i safety editing. *arXiv preprint arXiv:2510.24820*.

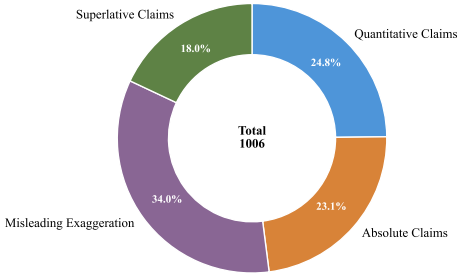
A Dataset Statistics and Violation Policy Definitions

A.1 Data Distribution

Figure 4 illustrates the distribution of the violation policies across our industrial dataset.



(a) Training Set Distribution



(b) Test Set Distribution

Figure 4: Distribution of violation policies in the training and test datasets.

A.2 Violation Policy Definitions

We align our violation policies with the production moderation standards of the online platform. Table 4 presents desensitized definitions and illustrative examples, intended solely to capture the general linguistic characteristics of each violation policy.

B Details of Experience-Driven Data Synthesis

B.1 Prompt for Standard Rectification (SFT & Data Synthesis)

Figure 5 illustrates the standard prompt template used to guide the LLM in rectifying non-compliant text and generating structured outputs.

B.2 Prompt for Group-Relative Compliance Experience Extractor

Figure 6 presents the prompt utilized by the Group-Relative Compliance Experience Extractor.

B.3 Case Study: Trajectory Splitting and Experience Extraction

Figure 7 demonstrates a complete workflow of trajectory splitting and experience extraction, where the LLM performs three consecutive rounds of rewriting, incrementally resolving non-compliance based on the previous output. This continuous trajectory is split into contrastive pairs. The extractor then analyzes the semantic gap between these paired responses to extract specific compliance experiences.

B.4 Algorithmic Details of GCEE

The concrete hyperparameters for the GCEE data-synthesis stage are as follows: we run 3 epochs with batch size 10; for each sample, we generate 5 rollouts; and we extract at most 2 experiences per sample.

C Details of Coherence Evaluation

To ensure the rectified text maintains linguistic fluency and semantic intent, we employ Gemini3-Flash as an LLM-as-a-judge evaluator. We designed a rigorous three-level evaluation rubric to assess the quality of modifications, mapping each quality level to a specific scalar reward. To align the LLM's judgment with human preferences, we include representative few-shot examples in the prompt to guide the evaluation. Table 5 details the specific criteria used by the evaluator.

D Implementation Details

We build $\mathcal{R}^3$ on top of Qwen3-8B. We perform SFT with LoRA (r=64) using a learning rate of 1×10^{-4}, batch size 64, and train for 2 epochs. We then apply GRPO with the two-phase curriculum; each phase is trained for 1 epoch with a learning rate of 1×10^{-6}, group size 8, and a KL coefficient of 0.01. All experiments are conducted on 8 NVIDIA H20 GPUs.

Policy	Definition	Examples
Quantitative Claims	Claims that cite specific data, statistics, or probability rates (e.g., success rates, reduction percentages).	Our course has successfully helped 5,000 students pass the exam.
Absolute Claims	The use of absolute or exclusive terms that imply a product is indispensable, universal, or has no alternatives.	This product is a cosmetic that is essential for every woman.
Superlative Claims	The use of highest-degree adjectives (e.g., "Best," "No.1," "Top") to describe a product's quality or status.	We are the No.1 education app in the market.
Misleading Exaggeration	Hyperbolic descriptions that distort the product's actual efficacy.	Use this cream and look 20 years younger instantly.

Table 4: Definitions and examples of the four violation policies targeted in our experiments.

Label (Score)	Fluency (Linguistic Quality)	Semantic Intent Preservation
Improvement (1.0)	**Repairs Flaws.** The edit fixes pre-existing grammatical truncations or logical gaps, resulting in a flow that is significantly more natural than the original.	**Clarifies Intent.** The modification removes ambiguity or awkward phrasing in the original text, making the marketing message clearer and more impactful.
Neutral (1.0)	**Maintains Status Quo.** The sentence structure remains unchanged. Pre-existing typos or colloquialisms are tolerated as long as it is not exacerbated.	**Consistent Intent.** The core marketing message is fully preserved. Modifications are strictly limited to replacing non-compliant terms with compliant synonyms.
Degradation (0.0)	**Introduces Errors.** The edit creates new grammatical faults, awkward collocations, or disjointed connections that did not exist in the source text.	**Distorts Meaning.** The edit reverses the original meaning (e.g., positive to negative), introduces logical contradictions, or loses key product information.

Table 5: The evaluation rubric for Coherence Reward. The judge assesses both linguistic fluency and semantic intent preservation to determine the quality of the rectification, assigning a binary-style score

<system prompt>

Role:
You are an expert content moderator specializing in advertising laws and platform policies. Your objective is to analyze advertisement copy with precision and provide compliant rectification solutions.
You will be provided with a violation reason (derived from Policy Details) and the transcript of a video advertisement that failed review due to non-compliant content.

Task Instructions:
1. Analyze Rules: Based on the provided [Policy Details], understand the definitions, prohibited content, and exemptions behind the violation labels.
2. Locate Issues: Pinpoint the specific sentences in the transcript that caused the violation.
3. Identify Key Points: Determine which specific Key Point within the <Rule Block> of the [Policy Details] the sentence violates.
4. Rectify & Rewrite: Generate a compliant revised version.
5. Output Format: Return a strict JSON object containing a list where each element includes "original" and "modified".

Strict Constraints:
1. Compliance: The modification must completely eliminate any suspicion of violation.
2. Semantic Preservation: Maintain the original commercial intent; remove only the violating parts.
3. Full Sentence Citation: The "original" field must contain the complete, punctuation-delimited independent clause from the source text.

Policy Details:
{policy details}

Output Format:
```
[
   {
        "original": "The original sentence containing the violation",
        "modified": "The compliant sentence of similar length"
   },
 ...
]
```

Examples:
{examples}

<user prompt>

Violation Policy:
{policy}

Transcript:
{transcript}

Figure 5: Standard Prompt for Rectification

<system prompt>

You are a Senior Advertising Compliance and Risk Control Expert. Your task is to perform "Attribution Analysis" and "Rule Tuning."
<Agent Goal>
Rewrite non-compliant ad copy into compliant copy. Input includes the violation analysis report and the original transcript. Output is a JSON list of modifications, where each item contains ‘original’, ‘modified. The rewriting must: 1) Eliminate violating content; 2) Not contradict the original semantics; 3) Not modify non-violating content.
</Agent Goal>
<Learning Goal>
By analyzing the success and failure of rewrite attempts, discover blind spots in the execution of existing rules and learn how to effectively rewrite violating copy to pass compliance checks. Focus on the following:
1. Find failure patterns where synonym replacement (e.g., changing 'No.1' to 'First Choice') is attempted but still ruled as a violation.
2. Analyze cases that fail across multiple attempts to identify stubborn mindset fixations causing the impasse.
3. Capture 'Implicit Violation Terms': Analyze terms initially ignored by the model but eventually forced to change to pass.
4. Establish Keyword & Pattern Mappings: Summarize specific violation vocabularies and sentence patterns.
</Learning Goal>
<Current Context Background>
The rewriting model already knows and is required to adhere to the following basic policy details during task execution:
{policy_details}
</Current Context Background>

<Core Task>
You will see multiple rewrite attempts (including both successful and failed cases) for the same violating copy. Note that these attempts were generated under the guidance of the basic rules above. Your goal is NOT to repeat the above rules, but to identify deviations, blind spots, or execution loopholes in the model's understanding, and summarize specific strategies to correct these issues.
Please execute the following analysis steps:
1. For failed rewrites, analyze why they failed despite knowing the basic rules. Was it a misunderstanding of the rule?
2. For successful rewrites, what specific phrasing or sentence structure was adopted to meet compliance requirements?
3. Based on the comparison, extract incremental, actionable rewriting experiences.

<Experience Extraction Principles>
1. DO NOT repeat basic rules (e.g., "Do not exaggerate").
2. BE specific and actionable (e.g., "Directly delete words like 'Ultimate' instead of finding synonyms").
3. FOCUS on "Negative Constraints" (tell the model what NOT to do).
4. Prioritize summarizing specific keywords or patterns often ignored by the model but causing violations.
</Experience Extraction Principles>

<Output Format>
Please output in the following format, ensuring corresponding tags are included:

<Success Analysis>
[Analyze how the successful case cleverly avoided the violation point, e.g., what alternative expressions were used.]
</Success Analysis>
<Failure Analysis>
[Analyze common issues in failed cases, especially "pseudo-compliance" features that seem to follow rules but are actually still non-compliant.]
</Failure Analysis>
<Comparative Analysis>
[Compare the key differences between successful and failed rewrites to identify the deciding factors for success.]
</Comparative Analysis>

<Experience Summary>
Extract up to { num_experiences } reusable rewriting experiences. Each experience should be a clear rewriting strategy or principle.
1. [Experience 1]
2. ...
</Experience Summary>
</Output Format>

<user prompt>
Original Input: { transcript }
Violation Policys: { policy }
<Multiple Rewrite Attempts>
{ attempts }
</Multiple Rewrite Attempts>

Figure 6: Prompt for Group-Relative Compliance Experience Extractor

Violating Content

Neighbors are going crazy for these! 100% of users agree: these shoes cure cold feet instantly. You don't need to spend a single penny—click the link to grab a pair for $0 cost! Today only, everyone watching is guaranteed to receive a pair directly from the factory. It combines moxibustion heat to permanently solve your health problems. Limited supply! No tricks, just click and take it home for free!

Rectification Trajectory

<LLM>:

```
[
  {
      "original": " Neighbors are going crazy for these !",
      "modified": " Many neighbors are rushing to buy these! "
  }
]
```

Non-compliance ❌

<LLM>: (Continue Rewriting)

```
[
  {
      "original": "100% of users agree: these shoes cure cold feet instantly.",
      "modified": " User feedback suggests these shoes help warm your feet. "
  }
]
```

Non-compliance ❌

<LLM>: (Continue Rewriting)

```
[
  {
      "original": "click the link to grab a pair for $0 cost!",
      "modified": "click the link to check our exclusive trial offer!"
  }
]
```

Compliance ✅

Trajectory Splitting

<Response1>:

```
[
  {
      "original": " Neighbors are going crazy for these !",
      "modified": " Many neighbors are rushing to buy these! "
  },
  {
      "original": "100% of users agree: these shoes cure cold feet instantly.",
      "modified": " User feedback suggests these shoes help warm your feet. "
  }
]
```

Non-compliance ❌

<Response2>:

```
[
  {
      "original": " Neighbors are going crazy for these !",
      "modified": " Many neighbors are rushing to buy these! "
  },
  {
      "original": "100% of users agree: these shoes cure cold feet instantly.",
      "modified": " User feedback suggests these shoes help warm your feet. "
  },
  {
      "original": "click the link to grab a pair for $0 cost!",
      "modified": "click the link to check our exclusive trial offer!"
  }
]
```

Compliance ✅

Extracted Experience

When detecting absolute zero-cost terms (e.g., '$0 cost', 'Free'), strictly verify if conditions apply. If so, replace with 'Trial Offer' or 'Campaign Price' to avoid misleading price violations.

Figure 7: A complete workflow of trajectory splitting and experience extraction