ARGUS: Policy-Adaptive Ad Governance via Evolving Reinforcement with Adversarial Umpiring

Deyi Ji[1] **Junyu Lu**[2*] **Xuanyi Liu**[3*] **Liqun Liu**[1] **Hailong Zhang**[1] **Peng Shu**[1]
Huan Yu[1] **Jie Jiang**[1] **Tianru Chen**[4] **Lanyun Zhu**[5†]
[1]Tencent [2]Dalian University of Technology [3]Peking University
[4]Zhejiang University [5]Tongji University
deyiji@tencent.com dutljy@mail.dlut.edu.cn xuanyi@stu.pku.edu.cn
{liqunliu,lericzhang,archershu,huanyu,zeus}@tencent.com
tianrun.chen@zju.edu.cn zhulanyun1999@gmail.com

Abstract

Online advertising governance faces significant challenges due to the non-stationary nature of regulatory policies, where emerging mandates (e.g., restrictions on education or aesthetic anxiety) create severe label inconsistencies and reasoning ambiguities in historical datasets. In this paper, we propose ARGUS, a policy-adaptive governance system that enables evolving reinforcement through multi-agent adversarial umpiring. ARGUS addresses the sparsity of new policy data by employing a three-stage framework: (1) Policy Seeding for initial perception; (2) Adversarial Label Rectification, which utilizes a "Prosecutor-Defender-Umpire" architecture to resolve conflicts between stale labels and new mandates; and (3) Latent Knowledge Discovery, which employs a tripartite dialectical discussion to unearth sophisticated, "gray-area" violations. By leveraging RAG-enhanced policy knowledge and Chain-of-Thought synthesis as dynamic rewards for reinforcement learning, ARGUS synchronizes its reasoning pathways with evolving regulations. Extensive experiments on both industrial and public datasets demonstrate that ARGUS significantly outperforms traditional fine-tuning baselines, achieving superior policy-adaptive learning with minimal gold data.

1 Introduction

Online advertising serves as the economic cornerstone of the modern internet, acting as a critical bridge between brands and global consumers. Given the proactive nature of algorithmic dissemination, maintaining a rigorous governance framework is essential for ensuring legal compliance and user safety. Unlike general content moderation (Gillespie, 2020; Gorwa et al., 2020; Langvardt, 2017), a single non-compliant advertisement can achieve massive reach through algorithmic pushing, necessitating a governance system that is not only precise but also highly responsive to an evolving regulatory landscape.

However, a fundamental challenge in ad governance is that policies are never static. Driven by seasonal trends and emerging social hotspots, new mandates, such as restrictions on K12 Achievement-Driven Tutoring or Body & Aesthetic Anxiety (as shown in Table 5), frequently emerge to address novel violations. Adapting existing models to these newly-added policies is a mission-critical yet formidable task. While platforms can collect sparse "gold data" reflecting new mandates, these samples are often insufficient to recalibrate large-scale models without compromising performance on historical data.

Designing such a policy-adaptive system entails three significant hurdles. 1) Label Inconsistency: historical data, though vast, was labeled under outdated policies ($\mathcal{P}_{old}$). Many samples may technically violate new dimensions ($\Delta\mathcal{P}$) but remain marked as compliant, leading to severe gradient conflicts. 2) Reasoning Ambiguity: new policies often involve subtle, "gray-area" interpretations where binary labels are insufficient for the model to internalize the underlying logic. 3) Recovery of Hard Samples: while models can identify overt violations, they often fail to recall sophisticated or deceptive non-compliant ads hidden within massive historical streams.

In this paper, we propose **ARGUS**, a Policy-Adaptive Ad Governance system that enables Evolving Reinforcement through Multi-Agent Adversarial Umpiring. To address label inconsistency, ARGUS employs a multi-stage alignment process. After an initial *Policy Seeding* phase, we introduce a **Prosecutor-Defender-Umpire** architecture. For historical samples, ARGUS utilizes an adversarial Vision-Language Model (VLM) to

[*]Junyu Lu and Xuanyi Liu participated in this work while interning at Tencent as part of the Tencent Rhino-Bird Research Elite Program, with Deyi Ji as the program leader.

[†]Corresponding Author.

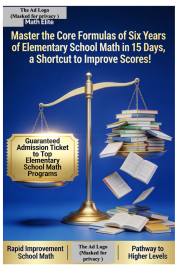
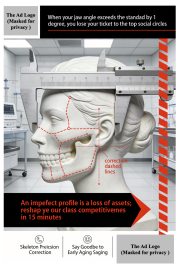
Case Image	Agent Role	Reasoning Process
 **Case 1: K12-T**	**Prosecutor**	The ad promises to "Master six years of math in 15 days", which is a typical false claim violating educational laws. The "Admission Ticket" metaphor implies a shortcut to elite schools, exploiting parental anxiety for enrollment. **Verdict: Violate**.
	Defender	The visual depicts a highly efficient learning method using a scale to emphasize the "weight" of knowledge. It is a rhetorical device for academic excellence and does not explicitly guarantee illegal admission outcomes. **Verdict: Comply**.
	Umpire	The conflict lies in whether "Rapid Improvement" constitutes inducement. Per **P33 (K12-T)**, the explicit use of "Shortcut" and "Guaranteed Admission" creates extreme anxiety and promises social mobility, meeting the definition of achievement-driven violation. **Verdict: Violate**.
 **Case 2: Aest-A**	**Skeptic**	Although the ad uses a plaster model instead of a human, its logic that "losing 1 degree means losing your ticket to top social circles" is extreme. Does this precise quantification of the jawline fall under "pathological aesthetic standards" prohibited by policy?
	Prosecutor	The ad uses calipers to measure the face and defines natural features as "imperfect profiles" and "loss of assets". It explicitly links appearance to social class, inducing self-loathing and urgency. **Verdict: Violate**.
	Defender	The visual is a professional simulation for medical education. Mentioning "social circles" is merely a rhetorical way to describe professional competitiveness and does not force a mandatory standard on all individuals. **Verdict: Comply**.
	Umpire	The Skeptic's point on "quantified aesthetics" is decisive. According to **P34 (Aest-A)**, inducing anxiety via quantified metrics is forbidden. Linking surgical reshaping to "class competitiveness" bypasses rational decision-making through psychological pressure. **Verdict: Violate**.

Table 1: Multi-agent dialectical reasoning analysis for representative violations.

act as a defender, generating counter-arguments for compliance against the model's prosecution. A neutral Umpire VLM then adjudicates these conflicting Chains-of-Thought (CoT), incorporating RAG-enhanced policy knowledge to rectify labels and provide high-fidelity rewards for reinforcement learning (RL). Furthermore, to unearth latent violations, ARGUS identifies latent non-compliant candidates, samples exhibiting high probabilistic affinity for new policies despite compliant labels. By subjecting these candidates to a tripartite dialectical discussion, ARGUS iteratively refines its reward model, synchronizing its reasoning pathways with the evolving regulatory mandates. Table 1 shows two cases of multi-agent dialectical reasoning analysis for representative violations.

Our contributions are: 1) We propose the ARGUS, which achieves autonomous policy adaptation by resolving label inconsistencies between historical data and newly-emerging multi-dimensional policies. 2) We introduce an Adversarial Umpiring mechanism that utilizes multi-agent dialectical reasoning and CoT synthesis as a dynamic reward function for reinforcement learning. 3) Extensive experiments on both industrial and public datasets demonstrate that ARGUS significantly outperforms traditional fine-tuning baselines, achieving superior policy-adaptive learning with minimal gold data.

2 Related Work

2.1 Advertisement Governance

Unlike general content moderation tasks such as hate speech or toxicity detection (Maity et al., 2024; Wang et al., 2026a; Lu et al., 2024; Xiao et al., 2024; Lu et al., 2024; Wang et al., 2024, 2026b; Lu et al., 2026), advertisement governance presents unique challenges due to the highly adversarial nature of malicious actors and the subtle semantic ambiguity of compliance rules. In recent years, a growing body of research has emerged to address these complexities. For instance, RAVEN (Ji et al., 2025a) first introduced a robust reinforcement learning (RL) framework for temporal localization in video ads, utilizing a multi-stage training scheme that combines coarse and fine-grained annotations. Building upon this, RAVEN++ (Ji et al., 2025b) focused on finer granularity and proposed an active reinforcement learning approach to proactively identify high-value samples during training, thereby optimizing localization boundaries. Fur-

thermore, Hi-Guard (Li et al., 2025) established a trustworthy multimodal moderation framework through hierarchical labeling and rule-based knowledge injection, while BLM-Guard (Yang et al., 2026) leveraged RL to score the model's reasoning process based on dynamic principles. However, all aforementioned works operate under the assumption of static governance rules. They do not explore a framework capable of adapting to significant policy shifts and evolving mandates. To the best of our knowledge, this paper is the first to propose an evolutionary multi-agent dialectic framework specifically designed to handle the continuous shift and expansion of advertising policies in an industrial setting.

2.2 Reinforcement Learning in Large Models

The landscape of Multimodal Large Language Models (MLLMs) (Yin et al., 2023; Amini et al., 2024; Chen et al., 2024; Zhang et al., 2024; Ji et al., 2024; Xu et al., 2024; Jiang et al., 2024) has been fundamentally reshaped, and establishes the bedrock for cross-modal alignment and temporal reasoning (Yin et al., 2023; Pentina et al., 2015; Ji et al., 2024; Liu et al., 2023; Zhu et al., 2025; Bai et al., 2023; Zhu et al., 2024b; Nye et al.; Wei et al., 2022; Zhu et al.). Beyond general cross-modal understanding, complex visual perception scenarios demand fine-grained feature modeling and robust spatial reasoning capabilities (Soviany et al., 2022; Zhu et al., 2025; Shazeer et al., 2017; Zhu et al., 2025; Ji et al., 2025; Graves et al., 2017; Zhu et al., 2021; McKinzie et al., 2024; Ji et al., 2022; Hacohen and Weinshall, 2019; Ma et al., 2024; Narvekar et al., 2020; Chiang et al., 2023). Such visual-oriented tasks present unique optimization challenges that vanilla pre-training fails to resolve, calling for dedicated alignment strategies to unify low-level visual sensing and high-level semantic comprehension (Ji et al., 2023; Zhu et al., 2024a; Ji et al., 2023). Reinforcement Learning (RL) has emerged as a critical paradigm for aligning model outputs with complex, high-dimensional human preferences (Yu et al., 2024; Rafailov et al., 2024; Liu et al.).

However, applying RL to the specific domain of advertisement governance introduces unique structural challenges that traditional RLHF frameworks (Christiano et al., 2017) rarely encounter. First, semantic policy ambiguity creates a reward-shaping dilemma; unlike objective tasks such as coding or mathematics, advertising compliance relies on nuanced interpretations of "harmfulness" or "deception", leading to high-variance reward signals. Second, the adversarial evolution of violations necessitates that RL agents go beyond simple pattern recognition to develop robust counter-reasoning against intentional obfuscation. Third, the dynamic policy shift in industrial environments requires RL frameworks to maintain high semantic plasticity, integrating emerging regulations while strictly preserving historical governance logic. ARGUS addresses these gaps by shifting from a monolithic reward model to a tri-party evolutionary game, where the reward is derived from a structured linguistic debate rather than a scalar preference score.

3 Methodology

3.1 Problem Formulation

We define the ad governance task as a multi-dimensional mapping function $f : \mathcal{X} \to (\mathbf{y}, \mathcal{C})$. Here, $\mathcal{X}$ is the space of multi-modal ads, and $\mathbf{y} = \{y^{(k)}\}_{k \in \mathcal{K}}$ is a vector of compliance labels, where $y^{(k)} \in \{0, 1\}$ indicates whether the ad violates the k-th specific policy category in the set $\mathcal{K}$. $\mathcal{C}$ denotes the corresponding Chain-of-Thought (CoT) reasoning that justifies the labels.

In a dynamic regulatory environment, the policy set expands from $\mathcal{P}_{old}$ to $\mathcal{P}_{new}$, where $\Delta\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$ represents n newly-emerging policy mandates (e.g., K12-T, Aest-A, as shown in Table 5). This evolution implies that the label space $\mathcal{K}$ is non-stationary, effectively adding new dimensions to the compliance vector $\mathbf{y}$.

Given a massive historical dataset $\mathcal{D}_{hist} = \{(x_i, \mathbf{y}_i^{old})\}_{i=1}^{N}$ labeled under $\mathcal{P}_{old}$, and a sparse "gold" dataset $\mathcal{D}_{gold} = \{(x_j, \mathbf{y}_j^{new})\}_{j=1}^{M}$ $(M \ll N)$ specifically annotated for the emerging categories in $\Delta\mathcal{P}$, the objective is to optimize a model f_θ to align with $\mathcal{P}_{new}$. The core challenge is label multi-dimensionality and inconsistency: a historical sample $x \in \mathcal{D}_{hist}$ marked as compliant under $\mathcal{P}_{old}$ may now possess a positive violation label $y^{(k)} = 1$ for some $k \in \Delta\mathcal{P}$. The model must therefore learn to infer these latent labels across multiple new policy dimensions **without exhaustive manual re-annotation** of $\mathcal{D}_{hist}$.

3.2 Evolving Reinforcement Framework

To bridge the gap between static training and dynamic mandates, we propose **Evolving Reinforcement**, a three-stage framework powered by Group Relative Policy Optimization (GRPO) (Shao et al.,

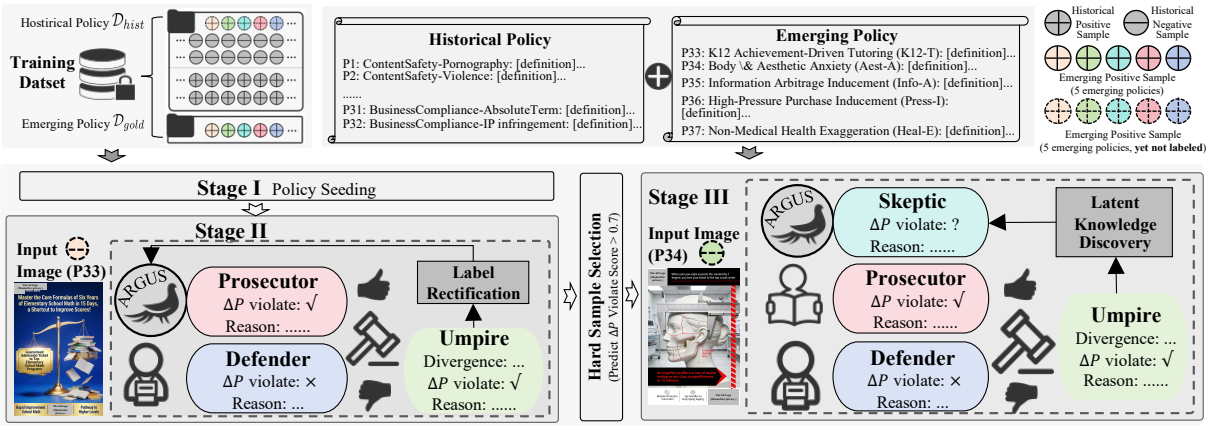


Figure 1: Overview of the ARGUS. ARGUS transitions through three stages: **Stage I (Policy Seeding)** for foundational policy perception; **Stage II (Adversarial Label Rectification)** which utilizes a bilateral debate between a **Prosecutor** and a **Defender** to rectify stale historical labels; and **Stage III (Latent Knowledge Discovery)** which employs a tripartite architecture including a **Skeptic** to unearth hard-to-detect violations in the uncertainty zone through logical triangulation by the **Umpire**. Here, "Positive" indicates non-compliant (violative), "Negative" indicates compliant (non-violative). The detailed case study is in Table 1 in the Appendix.

2024). Unlike traditional RL training (Christiano et al., 2017), ARGUS *evolves the policy and reward signals in tandem*, transitioning from initial perception to adversarial rectification and finally to latent discovery.

Stage I: Policy Seeding. This is a warm-up perception stage, we initialize the model via supervised alignment by blending $\mathcal{D}_{gold}$ with a curated subset of $\mathcal{D}_{hist}$. This phase establishes the foundational "policy sense" ($f_{\theta_{base}}$), ensuring the model internalizes the basic semantic boundaries of the new dimensions in $\Delta\mathcal{P}$ before entering high-variance reinforcement stages.

Stage II: Adversarial Label Rectification. This stage resolves *label inconsistencies* in historical data through active reinforcement. For samples in $\mathcal{D}_{hist}$, the model acts as a Prosecutor to challenge outdated labels. A *Prosecutor-Defender-Auditor* game generates dynamic rewards R_{rect} by adjudicating multi-agent rationales. This process effectively "overwrites" stale historical noise with high-fidelity rewards aligned with $\mathcal{P}_{new}$.

Stage III: Latent Knowledge Discovery. The final stage targets *hard samples* within the "uncertainty zone" (cases where violation intent is deeply latent or deceptive). By subjecting these candidates to a tripartite dialectic, the Auditor synthesizes conflicting views to refine decision boundaries in complex territories. This completes the evolution from surface-level pattern matching to deep, policy-driven reasoning.

3.3 Stage I: Policy Seeding

The goal of Stage I is *Incremental Policy Calibration*: enabling the production-grade model f_θ (pre-aligned with $\mathcal{P}_{old}$) to perceive the boundaries of $\Delta\mathcal{P}$ without degrading historical performance.

Strategic Data Blending. To incorporate new knowledge while avoiding *Gradient Overwhelming* from stale labels, we construct a hybrid training set $\mathcal{D}_{SFT} = \mathcal{D}_{gold} \cup \mathcal{D}'_{hist}$. Here, $\mathcal{D}'_{hist}$ is a representative subset ($\approx 40\%$) of historical data. This ratio ensures that the sparse but precise signal from $\mathcal{D}_{gold}$ is not drowned out by historical samples that might technically violate $\Delta\mathcal{P}$ but are marked as compliant in the legacy logs.

Optimization Objective. We optimize f_θ to maximize the joint likelihood of multi-dimensional compliance labels $\mathbf{y}$ and reasoning paths $\mathcal{C}$ under the expanded policy $\mathcal{P}_{new}$:

$$\mathcal{L}_{stage1}(\theta) = -\sum_{(x,\mathbf{y},\mathcal{C})\in\mathcal{D}_{SFT}} \log P(\mathbf{y},\mathcal{C}|x,\mathcal{P}_{new};\theta) \tag{1}$$

This yields a seeded model $f_{\theta_{base}}$ that possesses a preliminary "policy sense" for the new categories in $\Delta\mathcal{P}$.

Transition to Evolution. While Stage I enables the identification of overt violations, $f_{\theta_{base}}$ inevitably encounters **logical dissonance** where its emerging perception of $\Delta\mathcal{P}$ contradicts stale labels in $\mathcal{D}_{hist}$. This dissonance is not treated as noise but as the necessary impetus for the adversarial dialectic in Stage II.

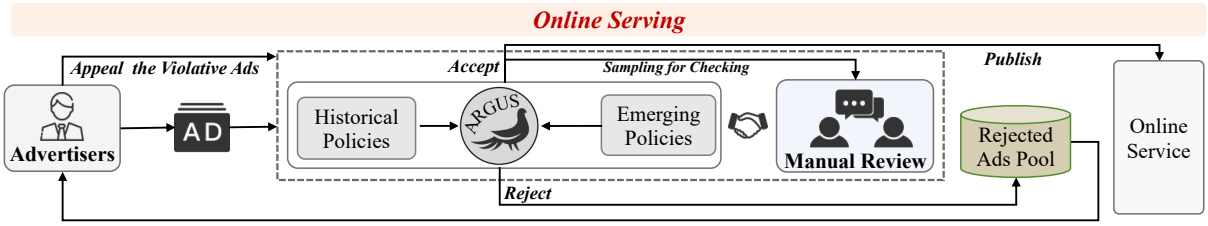


Figure 2: The online deployment of ARGUS.

3.4 Stage II: Adversarial Label Rectification

Following the seeding stage, we initiate Adversarial Label Rectification to resolve explicit conflicts where historical labels in $\mathcal{D}_{hist}$ contradict the emerging logic of $\Delta\mathcal{P}$.

Dialectical Debate. We establish a competitive reasoning environment to "stress-test" historical labels. The current policy model acts as the **Prosecutor**, generating a CoT that identifies violations within $\Delta\mathcal{P}$. Conversely, a strong VLM serves as the **Adversarial Defender**, tasked with justifying compliance through benign interpretations. This bilateral debate ensures that the model does not become over-sensitized to new polices while ignoring valid creative nuances.

Umpire Adjudication. The conflicting rationales are submitted to a neutral **Umpire VLM**. To ensure grounded decision-making, the Umpire utilizes a RAG-enhanced mechanism to retrieve specific clauses from $\Delta\mathcal{P}$ and reference samples from $\mathcal{D}_{gold}$. The Umpire adjudicates the debate by evaluating logical rigor and policy adherence:

$$(y^*, \mathcal{C}^*) = \text{Umpire}(CoT_{pros}, CoT_{def} \mid \mathcal{P}_{new}) \tag{2}$$

The output consists of a rectified label y^* and a high-fidelity reasoning chain $\mathcal{C}^*$.

In this stage, we redefine the reward function to guide the policy away from historical bias. The reward is composed of two components: the *Historical Consistency Reward* (R_{hist}) and the *Dialectic Rectification Reward* (R_{rect}). For R_{hist}, initially, the reward is derived from the legacy labels $y_{old} \in \mathcal{D}_{hist}$. However, as discussed, R_{hist} contains "stale noise" that contradicts $\Delta\mathcal{P}$. For R_{rect}, the Umpire adjudicates the debate to produce a gold-standard tuple $(y^*, \mathcal{C}^*)$. The reward for a model-generated sample $(y, \mathcal{C})$ is then computed as:

$$R_{rect}(y, \mathcal{C}) = \mathbf{1}(y = y^*) + \text{sim}(\mathcal{C}, \mathcal{C}^*) \tag{3}$$

where $\text{sim}(\cdot)$ measures the semantic alignment between the model's reasoning and the Umpire's adjudicated CoT.

By utilizing R_{rect} as the primary optimization target in GRPO, f_θ is forced to resolve the "logical dissonance" encountered in Stage I. The model learns to prioritize the umpire's adjudicated logic over the noisy y_{old}, achieving a high-quality self-corrected training stream. This ensures that both the final verdict $\mathbf{y}$ and the underlying reasoning $\mathcal{C}$ are synchronized with the emerging policy $\mathcal{P}_{new}$.

3.5 Stage III: Latent Knowledge Discovery

While Stage II resolves explicit label inconsistencies, sophisticated violations, such as deceptive creatives employing subtle evasion techniques, often persist in the decision boundary's "gray area." In these cases, the model f_θ may exhibit a high probabilistic affinity toward $\Delta\mathcal{P}$ yet still output a compliant label due to cautious bias. Stage III aims to evolve the model's discriminative depth by unearthing these hard samples.

Latent Candidate Selection. To identify hard samples within the vast $\mathcal{D}_{hist}$, we leverage the model's internal confidence. We define latent non-compliant candidates as samples predicted as compliant ($y^{(k)} = 0$) but possessing a posterior probability for an emerging policy $k \in \Delta\mathcal{P}$ that exceeds an empirical threshold τ:

$$\begin{aligned}\mathcal{D}_{latent} = \{x \in \mathcal{D}_{hist} \mid y^{(k)} = 0 \text{ and} \\ P(y^{(k)} = 1|x) > \tau\}.\end{aligned} \tag{4}$$

These candidates represent the "uncertainty zone" where the model's latent representation contradicts its categorical prediction.

Tripartite Dialectical Reasoning. To resolve high-entropy cases, we upgrade the adversarial mechanism, by introducing the current model f_θ as the **Skeptic**, which provides a "doubt-based" CoT that articulates the suspicious features triggering its uncertainty. This is combined with polarized perspectives from the agents introduced in Stage II,

Method	Historical Overall		Emerging Policies ($\Delta\mathcal{P}$)										Avg. $\Delta\mathcal{P}$	
	Prec.	Rec.	K12-T		Aest-A		Info-A		Press-I		Heal-E		Prec.	Rec.
			Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.		
Historical Expert (Qwen3-VL-8B)														
SFT (on $\mathcal{D}_{hist}$ only)	**0.842**	**0.858**	0.352	0.421	0.312	0.385	0.344	0.412	0.451	0.512	0.412	0.485	0.374	0.443
Zero-shot (Large Models)														
GPT-4o	0.485	0.612	0.421	0.582	0.384	0.521	0.442	0.594	0.521	0.655	0.481	0.615	0.450	0.593
Gemini 2.0 Flash	0.462	0.594	0.402	0.551	0.365	0.504	0.425	0.572	0.501	0.632	0.462	0.581	0.431	0.568
Qwen2.5-VL-72B	0.455	0.582	0.412	0.542	0.375	0.495	0.415	0.561	0.492	0.622	0.455	0.574	0.430	0.559
Qwen3-235B-A22B	0.512	0.635	0.454	0.612	0.412	0.554	0.478	0.642	0.572	0.704	0.521	0.641	0.487	0.631
Zero-shot (Base Models)														
Qwen2.5-VL-7B	0.312	0.451	0.312	0.484	0.285	0.454	0.344	0.511	0.421	0.582	0.374	0.522	0.347	0.511
Qwen3-VL-8B	0.342	0.482	0.342	0.512	0.312	0.485	0.375	0.542	0.454	0.612	0.402	0.552	0.377	0.541
Fine-tuning (Qwen2.5-VL-7B)														
Vanilla SFT (on $\mathcal{D}_{gold}$)	0.421	$0.402^{\dagger}$	0.762	0.745	0.714	0.682	0.735	0.712	0.782	0.754	0.751	0.732	0.749	0.725
SFT + Replay (40%)	0.772	0.765	0.735	0.722	0.692	0.665	0.715	0.692	0.782	0.751	0.741	0.725	0.733	0.711
EWC (Continual Learning)	0.792	0.784	0.744	0.732	0.702	0.682	0.724	0.705	0.792	0.765	0.752	0.741	0.743	0.725
Fine-tuning (Qwen3-VL-8B)														
Vanilla SFT (on $\mathcal{D}_{gold}$)	0.454	$0.432^{\dagger}$	0.782	0.761	0.735	0.702	0.754	0.731	0.815	0.782	0.784	0.764	0.774	0.748
SFT + Replay (40%)	0.791	0.785	0.752	0.745	0.708	0.685	0.735	0.718	0.805	0.775	0.765	0.741	0.753	0.733
EWC (Continual Learning)	0.802	0.794	0.761	0.746	0.715	0.698	0.742	0.725	0.811	0.782	0.772	0.754	0.760	0.741
ARGUS (Qwen2.5-VL-7B)	0.815	0.822	0.785	0.804	0.702	0.765	0.741	0.792	0.801	0.854	0.784	0.812	0.763	0.805
ARGUS (Qwen3-VL-8B)	0.828	0.841	**0.812**	**0.835**	**0.734**	**0.792**	**0.782**	**0.824**	**0.834**	**0.885**	**0.815**	**0.842**	**0.795**	**0.836**

Table 2: Results on Industrial Dataset. Prec. and Rec. denote Precision and Recall. Zero-shot models exhibit low precision and higher recall on $\Delta\mathcal{P}$ due to unfamiliarity with specific boundary nuances. † indicates catastrophic forgetting, vanilla SFT shows strong adaptation to new rules but suffers from severe historical knowledge collapse. ARGUS-8B maintains a minimal performance gap ($< 2\%$) compared to the Historical Expert while outperforming all baselines on average $\Delta\mathcal{P}$.

1) Prosecutor Agent: Constructs a rigorous argument for violation based on the latest policy $\Delta\mathcal{P}$; **2) Defender Agent:** Provides benign interpretations of the ad to prevent over-sensitization; **3) Skeptic (f_θ):** Highlighting internal conflicts and the reasoning behind its high-probability latent state.

Umpire Synthesis and RL Evolution. The neutral Umpire VLM gathers the dialectical triplet $\{CoT_{pros}, CoT_{def}, CoT_{skeptic}\}$. Through logical triangulation, the Umpire synthesizes a rectified conclusion y^* and a standardized reasoning chain $\mathcal{C}^*$. We define the **Evolution Reward** R_{evolve} to align the model with this refined logic:

$$R_{latent}(y, \mathcal{C}) = \mathbf{1}(y = y^*) + \text{sim}(\mathcal{C}, \mathcal{C}^*) \quad (5)$$

By optimizing against R_{latent} via GRPO, the model's decision boundary is pushed into hard-negative territories. This completes the policy evolution from surface-level perception to deep, intent-driven policy deduction.

4 Online Deployment

Fig. 2 shows the online deployment of ARGUS, we include the detailed illustration in Sec. A in Appendix.

5 Experiments

5.1 Offline Testing on Industrial Datasets

As illustrated in Table 2, a comprehensive comparison between ARGUS and diverse baselines yields the following insights:

Mitigating Catastrophic Forgetting. Maintaining stability on historical mandates during policy adaptation is a core challenge. The Vanilla SFT baseline exhibits severe knowledge erosion, with historical recall collapsing to 0.432 (†). In contrast, ARGUS-8B retains a historical recall of 0.841, matching the specialized Historical Expert (0.858) within a marginal 1.7% gap. This confirms that our evolutionary framework effectively preserves foundational governance logic while incorporating new mandates.

Domain Expertise vs. Model Scale. General-purpose giants like GPT-4o and Qwen3-235B exhibit a "low-precision, high-recall" bias, failing to outperform the specialized ARGUS-8B on emerging policies ($\Delta\mathcal{P}$). ARGUS achieves a precision improvement of over 30% compared to these models, demonstrating that targeted adversarial evolution provides finer boundary sensitivity for industrial compliance than raw parameter scaling.

Effectiveness of Adversarial Dialectic. While

Method	Historical Overall		Emerging: Dispirited Culture (Avg. $\Delta\mathcal{P}$)	
	Prec.	Rec.	Prec.	Rec.
Historical Expert (Qwen3-VL-8B)				
SFT (on $\mathcal{D}_{hist}$ only)	**0.634**	**0.658**	0.124	0.152
Zero-shot (Large Models)				
GPT-4o	0.531	0.557	0.311	0.365
Qwen3-VL-8B	0.304	0.402	0.221	0.262
Fine-tuning (Qwen3-VL-8B)				
Vanilla SFT (on $\mathcal{D}_{gold}$)	0.332	$0.384^{\dagger}$	0.415	0.435
SFT + Replay (40%)	0.492	0.512	0.395	0.417
EWC (Continual Learning)	0.512	0.525	0.410	0.413
ARGUS (Qwen3-VL-8B)	0.621	0.655	**0.454**	**0.482**

Table 3: Results on the public ToxiCN MM dataset. "Dispirited Culture" serves as the emerging policy to test the model's adaptation to subtle linguistic metaphors. † indicates catastrophic forgetting where vanilla SFT fails to retain historical knowledge. ARGUS maintains stability on historical domains while achieving best results on emerging toxic memes.

continual learning methods like EWC effectively mitigate forgetting, their adaptation to new policies is suboptimal. ARGUS-8B outperforms EWC (Kirkpatrick et al., 2017) in average $\Delta\mathcal{P}$ recall by 9.5% (0.836 vs. 0.741), with the most significant gains in high-ambiguity domains. This confirms that the tri-party game, specifically the Prosecutor's ability to synthesize "gray-area" cases, forces the model to internalize intent-level non-compliance rather than surface-level pattern matching.

Foundation Model Impact. Comparing ARGUS variants, the Qwen3-8B backbone consistently yields a 2%–3% improvement over the Qwen2.5-7B version. While the framework is model-agnostic, a stronger visual-linguistic foundation provides superior "logical intuition", facilitating a more sophisticated equilibrium during the policy evolution process.

5.2 Offline Testing on Public Datasets

To validate cross-domain generalization, we evaluated ARGUS on the public ToxiCN MM dataset. Results in Table 3 mirror industrial trends. On the one hand, the *Historical Expert* fails on the emerging *Dispirited Culture* policy (0.152 recall), as standard safety models are "blind" to subtle cultural metaphors. Conversely, while *Vanilla SFT* adapts to the new category, it suffers from catastrophic forgetting, with historical recall plunging to 0.384. ARGUS achieves a superior balance, maintaining near-optimal historical recall (0.655) while effectively integrating new policy knowledge. On the other hand, general models like GPT-4o lack the policy-sensitivity required for specialized Chinese memes, with ARGUS outperforming GPT-4o by 11.7% in recall on the emerging category. Compared to EWC and SFT+Replay, ARGUS demonstrates higher plasticity; its 6.9% absolute improvement over EWC highlights that the tri-party game is not mere regularization, but a reasoning-driven adaptation process that resolves logical contradictions between evolving mandates.

Online A/B Testing Group	VLR ↓	AAR ↑	FPR ↓
Control Group (Production)	1.42%	68.5%	0.35%
Treatment Group (**ARGUS**)	0.92%	76.2%	0.32%
Relative Improvement	+35.2%	+11.2%	+8.5%

Table 4: Online A/B testing results. ARGUS significantly reduces violation leakage while improving audit automation.

5.3 Online A/B Testing

To evaluate the practical utility of ARGUS, we conduct the online A/B experiment on our advertising platform. We allocate a small amount of the production traffic to the Treatment group, while the remaining traffic was handled by the Control group. Both the groups are based on Qwen3-VL 8B model. We evaluated three primary industrial metrics: (1) Violation Leakage Rate (VLR), determined by human expert back-checking; (2) Audit Automation Rate (AAR), the ratio of ads processed without human intervention; and (3) False Positive Rate (FPR), verified by professional auditors. As shown in Table 4, ARGUS achieves a 35.2% reduction in VLR, and the 11.2% increase in AAR significantly reduced human labor costs by moving high-confidence "gray-area" traffic into automated pipelines. Crucially, these gains are achieved alongside a lower FPR (0.32%) as verified by auditors, ensuring that the platform's efficiency is improved without compromising the experience of compliant advertisers.

6 Conclusion

In this paper, we present ARGUS, an evolutionary framework designed for ad governance under non-stationary policy environments. ARGUS employs a multi-agent tri-party dialectic to transform passive policy compliance into an active process. Extensive evaluations on both industrial and public datasets demonstrate that ARGUS achieves superior adaptation to emerging policies.

7 Ethical Considerations

Our research strictly adheres to established ethical guidelines and prioritizes data integrity and user privacy. The datasets utilized in this study consist of publicly accessible or de-identified industrial advertising samples. All labeling and evaluation processes are conducted solely for scientific analysis and do not reflect the institutional stances or personal opinions of the authors. Furthermore, the ARGUS framework is designed to enhance digital safety and platform reliability; however, we emphasize that its deployment should remain under human-in-the-loop oversight to prevent unintended biases. All resources and methodologies described herein are intended for research purposes, contributing to the broader goal of fostering a secure, compliant, and transparent digital advertising ecosystem.

Acknowledgment

This research was supported in part by the National Natural Science Foundation of China (Nos. 625B2033). We acknowledge Baoyan Zhuang and Pengda Qin from Tencent for collaborating on data resources and application scenarios to validate and improve algorithm performance.

References

Lanyun Zhu, Deyi Ji, Tianrun Chen, Haiyang Wu, and Shiqi Wang. Retrv-r1: A reasoning-driven mllm framework for universal and efficient multimodal retrieval. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. Show your work: Scratchpads for intermediate computation with language models. In *Deep Learning for Code Workshop*.

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*.

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

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

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

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Kyle Langvardt. 2017. Regulating online content moderation. *Geo. LJ*, 106:1353.

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

Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1):2053951719897945.

Tarleton Gillespie. 2020. Content moderation, ai, and the question of scale. *Big Data & Society*, 7(2):2053951720943234.

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

Lanyun Zhu, Deyi Ji, Shiping Zhu, Weihao Gan, Wei Wu, and Junjie Yan. 2021. Learning statistical texture for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12537–12546.

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

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

Deyi Ji, Haoran Wang, Mingyuan Tao, Jianqiang Huang, Xian-Sheng Hua, and Hongtao Lu. 2022. 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.

Deyi Ji, Feng Zhao, and Hongtao Lu. 2023. Guided patch-grouping wavelet transformer with spatial congruence for ultra-high resolution segmentation. *International Joint Conference on Artificial Intelligence*, pages 920–928.

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

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

Deyi Ji, Feng Zhao, Hongtao Lu, Mingyuan Tao, and Jieping Ye. 2023. 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.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

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

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

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

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

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

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024. Fine-tuning llama for multi-stage text retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2421–2425.

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

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

Lanyun Zhu, Tianrun Chen, Deyi Ji, Jieping Ye, and Jun Liu. 2024b. 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.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Zhenting Wang, Shuming Hu, Shiyu Zhao, Xiaowen Lin, Felix Juefei-Xu, Zhuowei Li, Ligong Han, Harihar Subramanyam, Li Chen, Jianfa Chen, et al. 2024. Mllm-as-a-judge for image safety without human labeling. *arXiv preprint arXiv:2501.00192*.

B McKinzie, Z Gan, J Fauconnier, S Dodge, B Zhang, P Dufter, D Shah, X Du, F Peng, F Weers, et al. 2024. Mm1: methods, analysis & insights from multimodal llm pre-training. arxiv. *Preprint posted online on April*, 18.

Chao Zhang, Haoxin Zhang, Shiwei Wu, Di Wu, Tong Xu, Yan Gao, Yao Hu, and Enhong Chen. 2024. Notellm-2: Multimodal large representation models for recommendation. *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2024. 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.

Yunkai Chen, Qimeng Wang, Shiwei Wu, Yan Gao, Tong Xu, and Yao Hu. 2024. Tomgpt: Reliable text-only training approach for cost-effective multi-modal large language model. *ACM Transactions on Knowledge Discovery from Data*.

Junyu Lu, Bo Xu, Xiaokun Zhang, Hongbo Wang, Haohao Zhu, Dongyu Zhang, Liang Yang, and Hongfei

Lin. 2024. Towards comprehensive detection of chinese harmful memes. *Advances in Neural Information Processing Systems*, 37:13302–13320.

Junyu Lu, Bo Xu, Xiaokun Zhang, WangHongbo, Haohao Zhu, Dongyu Zhang, Liang Yang, and Hongfei Lin. 2024. Towards comprehensive detection of chinese harmful memes. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yunze Xiao, Yujia Hu, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2024. ToxiCloakCN: Evaluating robustness of offensive language detection in Chinese with cloaking perturbations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6012–6025.

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

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

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

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

Deyi Ji, Yuekui Yang, Liqun Liu, Peng Shu, Haiyang Wu, Shaogang Tang, Xudong Chen, Shaoping Ma, Tianrun Chen, and Lanyun Zhu. 2025b. 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. 2025a. 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.

Qi Zhu, Jiangwei Lao, Deyi Ji, Junwei Luo, Kang Wu, Yingying Zhang, Lixiang Ru, Jian Wang, Jingdong Chen, Ming Yang, Dong Liu, and Feng Zhao. 2025. Skysense-o: Towards open-world remote sensing interpretation with vision-centric visual-language modeling. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 14733–14744.

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

Anqi Li, Wenwei Jin, Jintao Tong, Pengda Qin, Weijia Li, and Guo Lu. 2025. Towards trustworthy multimodal moderation via policy-aligned reasoning and hierarchical labeling. *arXiv preprint arXiv:2508.03296*.

Yiran Yang, Zhaowei Liu, Yuan Yuan, Yukun Song, Xiong Ma, Yinghao Song, Xiangji Zeng, Lu Sun, Yulu Wang, Hai Zhou, et al. 2026. Blm-guard: Explainable multimodal ad moderation with chain-of-thought and policy-aligned rewards. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 35985–35993.

Junyu Lu, Deyi Ji, Liqun Liu, Xiaokun Zhang, Youlin Wu, Roy Ka-Wei Lee, Peng Shu, Huan Yu, Jie Jiang, Bo Xu, Liang Yang, and Hongfei Lin. 2026. Decoding multimodal cues: Unveiling the implicit meaning behind hateful videos. In *ACM SIGIR Conference on Research and Development in Information Retrieval*.

Han Wang, Deyi Ji, Junyu Lu, Lanyun Zhu, Hailong Zhang, Haiyang Wu, Liqun Liu, Peng Shu, and Roy Ka-Wei Lee. 2026a. Multi-agent vlms guided self-training with pnu loss for low-resource offensive content detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 39387–39396.

Han Wang, Deyi Ji, Lanyun Zhu, Jiebo Luo, and Roy Ka-Wei Lee. 2026b. Streamsense: Streaming social task detection with selective vision–language model routing. In *Proceedings of the ACM Web Conference 2026*, pages 8897–8906.

Appendix

A Online Deployment

As shown in Fig. 2, the online infrastructure employs a cascaded filtering mechanism to maintain low-latency response: 1) Initial Screening: Incoming ads from advertisers undergo a initial check. Ads triggered by this process are immediately rejected and returned to the advertiser. 2) ARGUS Adaptive Governance: Ads passing the initial filter are processed by the ARGUS engine, which specifically evaluates compliance against both Historical and Emerging Policies ($\Delta\mathcal{P}$). By leveraging the reasoning pathways evolved during offline stages, the engine can accurately identify novel violations that legacy systems might miss. 3) Manual Review & Feedback: To ensure the highest fidelity, a specialized Manual Review module performs sampling for checking. This human-in-the-loop component serves as a final quality gate, and the resulting high-quality labels are fed back to the offline modeling pipeline to facilitate continuous model evolution. 4 Final Decision: Ads verified as compliant are published to the Online Service, while non-compliant samples are diverted to the Rejected Ads Pool.

B Datasets

B.1 Industrial Dataset

To evaluate ARGUS in a rigorous production-level environment, we construct a large-scale multimodal (image & text) dataset derived from our advertising platform. The dataset is specifically designed to reflect the non-stationary nature of advertising governance, categorizing policies into two primary sets: **Historical Policies** ($\mathcal{P}_{hist}$) and **Emerging Policies** ($\Delta\mathcal{P}$).

B.1.1 Emerging Policy Definitions ($\Delta\mathcal{P}$)

To evaluate ARGUS's adaptability to non-stationary environments, we define five emerging policy domains ($\Delta\mathcal{P}$) that represent recent regulatory shifts and sophisticated deceptive tactics. These categories are characterized by high semantic ambiguity and require deep intent-level reasoning, as shown in Table 5.

B.1.2 Data Composition and Statistics

The dataset comprises approximately $680,000$ high-quality samples, as detailed in Table 6:

- **Historical Policies ($\mathcal{P}_{hist}$):** This subset contains $168,000$ positive samples covering over 50 foundational categories. These include *Basic Content Safety* (e.g., pornography, violence, bloodiness, hate speech, and prohibited items) and *Business Compliance* (e.g., absolute terminology, exaggerated marketing, and IP infringement).
- **Emerging Policies ($\Delta\mathcal{P}$):** We collect $32,700$ samples representing five critical emerging policy shifts. These domains are characterized by their high semantic ambiguity and evolving deceptive tactics.
- **Negative Samples:** To ensure robust training and minimize false positives, we incorporated approximately $480,000$ negative samples (compliant ads) that share similar visual or linguistic features with non-compliant ones, creating a challenging "needle-in-a-haystack" scenario.

B.2 Public Dataset

To evaluate the generalization and robustness of ARGUS on diverse social media content, we incorporate the ToxiCN MM dataset (Lu et al., 2024), the first comprehensive Chinese harmful meme dataset. ToxiCN MM consists of 12,000 multimodal samples with fine-grained annotations across multiple dimensions of harmfulness on the Chinese internet.

In our experimental setup, we adapt this dataset to simulate a policy evolution scenario by re-categorizing its original taxonomy into historical and emerging domains:

- **Historical Policies ($\mathcal{P}_{hist}$):** We select 3 categories (*Targeted Harmful*, *General Offensive*, and *Sexual Innuendo*) to represent the established historical policies. These categories align with the foundational content safety mandates in our industrial dataset.
- **Emerging Policy ($\Delta\mathcal{P}$):** We define *Dispirited Culture* (Sang Culture) as the emerging policy. This category is particularly challenging as it involves subtle linguistic metaphors and pessimistic sentiments that deviate from traditional toxic content, requiring advanced reasoning to identify its potential negative social impact.

By bridging industrial advertising data with this public social media dataset, we create a rigorous

PolicyID	Policy Category	Description ($\Delta\mathcal{P}$)
P33	**K12 Achievement-Driven Tutoring (K12-T)**	Targets achievement-driven academic tutoring for K-12 students. It prohibits promoting "exam shortcuts" or "guaranteed admission" that exploit parental anxiety and utilitarian educational goals.
P34	**Body & Aesthetic Anxiety (Aest-A)**	Regulates content that promotes singular beauty standards (e.g., extreme thinness) or implies that physical flaws are barriers to a successful life, thereby inducing psychological distress and body dysmorphia.
P35	**Information Arbitrage Inducement (Info-A)**	Prohibits inducing financial investment through claims of "insider info" or "unclosed trends." It targets the masking of fraudulent risks under the guise of "wealth shortcuts" or exclusive "circle privileges."
P36	**High-Pressure Purchase Inducement (Press-I)**	Regulates the use of artificial urgency (e.g., fake countdowns, false stock limits) and compulsive logic (e.g., "regret for life") designed to bypass rational decision-making in e-commerce.
P37	**Non-Medical Health Exaggeration (Heal-E)**	Targets non-medical supplements claiming therapeutic effects (e.g., "curing cancer" or "restoring physiological indicators"). It prohibits substituting professional medical treatment with vague health-related efficacy claims.

Table 5: Definitions of the 5 emerging policies ($\Delta\mathcal{P}$). These categories represent high-stakes domains characterized by significant semantic ambiguity and adversarial evolution. The policy ids follow by the historical policies (total 32).

Category	Count	Type
Historical Policy Overall ($\mathcal{P}_{hist}$)	168,000	Positive
Emerging Domains ($\Delta\mathcal{P}$)		
P33: K12-T	4,500	Positive
P34: Aest-A	9,800	Positive
P35: Info-A	7,500	Positive
P36: Press-I	5,200	Positive
P37: Heal-E	5,700	Positive
Negative Samples (Compliant Ads)	480,000	Negative
Total	**680,700**	–

Table 6: Statistics of the industrial advertising dataset. Emerging policies focus on domains with high regulatory shifts and semantic ambiguity, while negative samples provide the necessary balance for robust industrial training.

Evolving Reinforcement Stages	Hist. Rec.	Avg. $\Delta\mathcal{P}$ Prec.	Avg. $\Delta\mathcal{P}$ Rec.
Stage I: Policy Seeding	0.785	0.753	0.733
+ Stage II: (Adversarial Label Rectification)	0.824	0.758	0.792
+ Stage III: (Latent Knowledge Discovery)	**0.841**	**0.795**	**0.836**

Table 7: Ablation study on the evolving reinforcement stages.

benchmark that tests whether ARGUS can transfer its tri-party evolutionary logic from structured commercial policies to highly informal, culturally-specific internet memes. This dual-dataset evaluation ensures that the performance gains of ARGUS are not confined to a specific data distribution but are representative of a generalized capability in policy-driven content moderation.

Component Variant	Prec.	Rec.
Full ARGUS Dialectic	**0.795**	**0.836**
w/o Prosecutor	0.732	0.695
w/o Defender	0.684	0.812
w/o Rationale (Labels Only)	0.715	0.742

Table 8: Component-wise ablation of the Multi-Agent Dialectic. Results are averaged across all $\Delta\mathcal{P}$ categories.

Adversarial Strategy	Std. SFT	GPT-4o	ARGUS
Normal Samples	0.711	0.582	**0.835**
Adversarial Samples	0.440	0.473	**0.783**

Table 9: Detection recall under adversarial evasion.

C Dialectical Prompting Design

The efficacy of the adversarial evolution process hinges on the quality of the perspectives generated by the agents. We design a dialectical prompting scheme that enforces role-specific constraints, ensuring a high-entropy debate that covers the full spectrum of policy interpretation.

Prosecutor: Rigorous Scrutiny. The Prosecutor agent is prompted to act as a strict regulatory inspector. Its objective is to identify any potential violation of the newly emerging policy $\Delta\mathcal{P}$, no matter how subtle. We employ *negative-bias prompting*, instructing the model to focus on deceptive visual layouts, exaggerated textual claims, and potential legal risks.

Defender: Adversarial Justification. To pre-

vent the system from becoming over-sensitive (i.e., excessive False Positive Rate), the Defender acts as "sophisticated legal counsel". It is tasked with providing alternative, benign interpretations for every point of contention raised by the Prosecutor. The prompt encourages *Contextual Re-framing*, for instance, interpreting an exaggerated claim as "artistic hyperbole" or a high-pressure countdown as "legitimate seasonal promotion". By exploring the boundary of creative integrity, the Defender forces the debate to remain grounded, sharpening the model's ability to distinguish between "gray-area" creatives and true policy violations.

Umpire: Logical Triangulation. The Umpire is designed as a neutral adjudicator with *RAG-enhanced objectivity*. Unlike the biased Prosecutor and Defender, the Umpire's prompt enforces a "Fact-first, Logic-second" hierarchy. It is required to first validate the specific policy clauses fetched from $\mathcal{P}_{new}$ and then evaluate the dialectical consistency of the opposing CoTs. We incorporate *Reasoning Pruning* constraints, instructing the Umpire to explicitly reject any hallucinations or irrelevant arguments presented during the debate. The final output is a synthesized reasoning chain $\mathcal{C}^*$ that represents the optimal balance between regulatory rigor and creative tolerance.

D Ablation Study

D.1 Study on Evolving Reinforcement Stages

To investigate the individual contribution of each stage within the ARGUS framework, we conduct an incremental ablation study. The results, summarized in Table 7, demonstrate the cumulative performance gains from our multi-stage evolutionary strategy.

Foundational Alignment via Stage I. Stage I (Policy Seeding) establishes the initial cross-domain alignment. It provides a baseline historical recall of 0.785 and an average $\Delta\mathcal{P}$ recall of 0.733. This stage ensures that the auditor agent internalizes the basic semantic boundaries of new policies before engaging in complex adversarial reasoning, preventing the model from starting the reinforcement process with unstable or biased gradients.

Adversarial Rectification in Stage II. The introduction of the Stage II (Adversarial Label Rectification) triggers a significant performance leap. These gains suggest that the dialectical conflict between the Prosecutor and Defender effectively rectifies mislabeled "gray-area" samples, forcing the model to develop a more robust reasoning logic rather than relying on surface-level pattern matching.

Boundary Refinement via Stage III. Stage III (Latent Knowledge Discovery) further polishes the model's sensitivity by mining and synthesizing hard adversarial cases. This stage achieves the peak performance across all metrics. Notably, the continued improvement in Historical Recall indicates that the hard sample mining process doesn't just aid new policy adaptation, it also reinforces the model's overall logical intuition, making it more resilient across the entire policy spectrum.

D.2 Ablation on Multi-Agent Dialectic Components

We conduct a component-wise ablation to verify the synergy between the agents. Results in Table 8 highlight the distinct contribution of each role to the overall policy evolution.

Precision-Recall Balance. The Prosecutor and Defender serve as the primary drivers for Recall and Precision, respectively. Removing the Prosecutor leads to a 14.1% drop in recall (0.695), as the model fails to uncover latent non-compliance. Conversely, removing the Defender causes precision to plunge to 0.684. This confirms that the Defender prevents "over-censorship" by forcing the model to consider benign justifications in gray-area cases.

The Power of Rationale. A crucial finding is that the **linguistic debate** itself is indispensable. When the agents provide only binary labels without detailed rationales (*w/o Rationale*), performance drops across all metrics. This proves that ARGUS's strength lies in "policy reasoning" rather than simple pattern matching, the model requires the logical depth of the debate to internalize complex regulatory boundaries.

D.3 Robustness against Adversarial Evasion

In real-world scenarios, evasion techniques like homophone replacement and visual blurring are usually used to bypass the filters. We evaluate ARGUS on an Adversarial Evaluation Set (2,000 samples) featuring these sophisticated obfuscation strategies. As shown in Table 9, ARGUS exhibits superior resilience compared to baselines. While the standard SFT model's recall plunges by 38.1% (from 0.711 to 0.440) when facing adversarial samples, ARGUS maintains a robust average recall of 0.783, experiencing only a marginal 6.2% decrease. This suggests that the tri-party game forces the model to

focus on intent-level cues rather than surface-level patterns.

E Limitations and Future Work

A primary constraint of this work is that ARGUS is currently evaluated on image-text advertisements. While effective for spatial reasoning, this focus does not account for the unique challenges of the video domain, which has become a dominant medium in digital marketing. Video governance requires capturing complex temporal dynamics and multi-modal synchronization, where non-compliance often emerges from the sequence of frames and audio cues rather than a single image. Extending ARGUS to support temporal-aware reasoning and video-based policy evolution remains a key direction for our future research.