Journal of Science, Innovation & Social Impact
EISSN: 3105-5028 | PISSN: 3105-501X | Vol. 2, No. 1 (2026)

Article

Multimodal Deep Learning for Advertising Content Safety: A Comprehensive Study on Detection and Governance Strategies

Xin Lu [1,*]

[1] Computer Science, Stanford University, CA, USA
* Correspondence: Xin Lu, Computer Science, Stanford University, CA, USA

Abstract: The proliferation of digital advertising across multiple platforms has created unprecedented challenges for content safety and brand protection. This paper presents a comprehensive study on multimodal deep learning approaches for detecting unsafe advertising content, addressing both explicit violations and implicit misleading information. We propose a novel framework that integrates visual, textual, and cross-modal features through advanced fusion architectures to achieve robust detection performance. Our methodology combines pre-trained language models, vision transformers, and optical character recognition systems with attention-based fusion mechanisms for comprehensive content analysis. Experimental results on a dataset of 45,000 advertising samples demonstrate that our approach achieves 92.3% accuracy in detecting policy violations, outperforming single-modality baselines by consistent gains. The framework shows particular strength in identifying implicit misleading content with an 89.1% F1-score and maintains balanced precision-recall trade-offs suitable for production deployment. This research contributes practical governance strategies for human-AI collaboration in content moderation workflows, addressing the critical need for scalable and accurate advertising safety systems in the digital ecosystem. Our method outperforms the best single-modality baseline by 15.5 percentage points and a strong late-fusion baseline by 8.6 percentage points.

Keywords: multimodal learning; advertising safety; content moderation; deep learning

1. Introduction

1.1. Background and Motivation

1.1.1. Current Challenges in the Digital Advertising Ecosystem

Digital advertising platforms process billions of advertisements daily, creating substantial content moderation challenges that traditional approaches cannot adequately address. The advertising ecosystem encompasses diverse formats, including static images, videos, carousel ads, and interactive content, each presenting unique detection requirements. Recent interpretable multimodal misinformation detection research demonstrates that 87% of policy violations involve subtle cross-modal inconsistencies rather than explicit, harmful content [1]. The scale of this challenge continues to expand, with programmatic advertising networks serving over 10 trillion ad impressions annually. Even a 0.1% violation rate represents millions of potentially harmful advertisements reaching consumers.

1.1.2. Regulatory Requirements and Platform Responsibilities

Advertising platforms face increasing regulatory pressure to ensure content safety while maintaining operational efficiency. The complexity of detecting and grounding

Received: 09 December 2025
Revised: 28 January 2026
Accepted: 08 February 2026
Published: 13 February 2026

multimodal media manipulation necessitates sophisticated technical solutions that can identify both overt violations and sophisticated deception techniques [2]. Some jurisdictions mandate rapid takedown responses (e.g., within 24 hours) and impose fines of up to 6% of global annual revenue for systemic failures; platforms must therefore balance compliance, precision, and operational scale. Platform responsibilities extend beyond straightforward content filtering to encompass brand safety protection, maintaining consumer trust, and managing relationships with advertisers.

1.1.3. Impact of Unsafe Advertising Content on Stakeholders

Unsafe advertising content has a cascading negative impact across the digital ecosystem. Cross-modal ambiguity in advertisements can mislead consumers through subtle manipulation techniques that evade traditional detection methods [3]. Brand safety incidents result in average revenue losses of $2.3 million per major violation, with long-term reputational damage affecting market valuations. Consumer trust metrics show a 73% reduction in platform engagement following exposure to misleading advertisements, while advertisers report a 45% decrease in campaign effectiveness when their content appears alongside policy violations.

1.2. Problem Definition and Research Gap

1.2.1. Limitations of Single-Modality Detection Approaches

Current single-modality detection systems fail to capture the complex interplay between visual and textual elements in modern advertising content. Comparative evaluation studies between AI and human moderators reveal that unimodal approaches miss 67% of violations that manifest through cross-modal inconsistencies [4]. Text-only analysis cannot detect misleading visual representations, while image-only processing misses critical contextual information embedded in ad copy. The limitation becomes particularly acute when dealing with implicit claims that require understanding relationships between multiple content elements.

1.2.2. Challenges in Implicit Misleading Content Identification

Implicit misleading content represents the most challenging detection category, requiring sophisticated reasoning capabilities beyond surface-level pattern matching. Multimodal misinformation detection through learning from synthetic data demonstrates that implicit violations often involve culturally specific references, temporal inconsistencies, and subtle emotional manipulation [5]. These challenges are compounded by adversarial techniques where bad actors deliberately craft content to evade automated detection while maintaining a deceptive impact on human viewers.

1.3. Contributions

1.3.1. Key Technical Contributions

This research introduces three primary technical innovations for ensuring the safety of advertising content. We develop a hierarchical attention mechanism that captures fine-grained cross-modal relationships, resulting in a 15.2% improvement over baseline fusion methods. Our framework incorporates domain-specific pre-training on advertising content, addressing the distribution shift between general web data and commercial content. The system demonstrates robust performance against adversarial perturbations, maintaining an accuracy of 88.7% under style transfer attacks that compromise existing methods.

1.3.2. Practical Implications for Content Governance

The proposed framework enables practical deployment strategies that strike a balance between automation and human oversight. Integration with existing content management systems requires minimal architectural changes while providing substantial accuracy improvements. The governance model supports tiered review processes, where

high-confidence predictions enable automatic decisions, while ambiguous cases receive the appropriate human attention. Deployment recommendations include confidence calibration techniques, incremental rollout strategies, and continuous learning mechanisms that adapt to emerging violation patterns.

2. Related Work

2.1. Traditional Content Moderation Approaches

2.1.1. Rule-Based and Keyword Matching Methods

Early content moderation systems relied on deterministic rules and keyword blacklists to identify policy violations. Stacked Bi-LSTM architectures with attention mechanisms evolved from these foundational approaches, incorporating contextual understanding beyond simple pattern matching [6]. Rule-based systems achieved reasonable precision for explicit violations but suffered from high false positive rates reaching 34% on legitimate content containing policy-related terms. Keyword matching approaches required constant manual updates to address emerging violation patterns and linguistic variations.

2.1.2. Early Machine Learning Techniques

Statistical machine learning methods introduced probabilistic reasoning to content moderation workflows. Support vector machines and random forests demonstrated improvements over rule-based systems, achieving 72% accuracy on structured advertising datasets. Rethinking content moderation from an asymmetric angle revealed that feature engineering quality determined performance ceilings, with handcrafted features capturing only surface-level patterns [7]. These approaches struggled with scalability challenges, requiring extensive feature engineering for each new violation category and failing to generalize across different advertising formats.

2.2. Deep Learning for Content Safety

2.2.1. CNN-Based Image Classification Methods

Convolutional neural networks have revolutionized visual content moderation by automatically learning features from raw pixel data. Practical approaches for brand safety using image multiclass classification achieved 91% accuracy on static image advertisements, demonstrating the power of deep visual representations [8]. ResNet and EfficientNet architectures became standard baselines, with transfer learning from ImageNet pre-training providing robust initialization. Multi-scale feature extraction captured both global context and fine-grained details relevant to policy violations.

2.2.2. Transformer Models for Text Analysis

Transformer architectures transformed text understanding through self-attention mechanisms that capture long-range dependencies. Modality and event adversarial networks demonstrated that BERT-based models achieve 94% accuracy in textual policy violation detection [9]. Pre-trained language models encode rich semantic representations that generalize across diverse advertising domains. Fine-tuning strategies adapted general-purpose models to advertising-specific vocabulary and violation patterns.

2.2.3. Temporal Models for Video Content

Video advertising introduces temporal dynamics requiring specialized architectures for comprehensive analysis. Interfaces of artificial intelligence and machine learning for financial fraud detection have pioneered techniques applicable to video ad moderation [10]. Three-dimensional convolutions and recurrent networks captured motion patterns indicative of policy violations. Temporal attention mechanisms identified critical frames containing violating content within more extended video sequences.

2.3. Multimodal Fusion Architectures

2.3.1. Early Fusion Strategies

Early fusion approaches concatenate features from different modalities before processing through shared layers. Research on fake news detection against style attacks has shown that early fusion captures cross-modal interactions but suffers from modality imbalance issues [11]. Joint embedding spaces enable unified representation learning across visual and textual inputs. Dimensionality challenges arise when combining high-dimensional features from multiple modalities.

2.3.2. Late Fusion Approaches

Late fusion maintains separate processing pipelines for each modality before combining predictions at the decision level. Frequency spectrum analysis for multimodal representation demonstrated that late fusion preserves modality-specific information while enabling specialized processing [12]. Independent optimization of modality-specific components simplifies training procedures. Decision-level combination strategies include weighted voting, stacking, and learned fusion functions.

2.3.3. Cross-Modal Attention Mechanisms

Attention mechanisms enable dynamic information exchange between modalities based on content relevance. Cross-modal attention learns to identify which features from each modality contribute most to violation detection. Bidirectional attention flows allow mutual enhancement between visual and textual representations. Hierarchical attention structures capture interactions at multiple granular levels, ranging from word-image regions to sentence-scene relationships.

3. Methodology

3.1. Dataset Construction and Annotation

3.1.1. Data Collection from Advertising Platforms

The dataset compilation process gathered 45,000 unique advertisements from twelve major advertising platforms spanning social media, display networks, and video streaming services. Machine learning approaches for brand protection guided the sampling strategy to ensure representative coverage of violation categories [13]. Platform-specific APIs provided structured metadata, including advertiser information, targeting parameters, and engagement metrics. Collection timestamps ranged from January 2023 to March 2024, capturing seasonal variations and emerging trends of violations. Geographic diversity encompassed advertisements from 47 countries across six continental regions, addressing cultural and linguistic variations in policy interpretation.

Advertisement formats included 18,500 static images, 12,300 video advertisements with an average duration of 23 seconds, 8,700 carousel advertisements featuring multiple creative elements, and 5,500 rich media advertisements with interactive components. Resolution requirements specified minimum dimensions of 1024x768 pixels for images and 720p for videos to ensure sufficient detail for accurate analysis. Metadata fields captured creation dates, modification histories, advertiser verification status, and historical records of policy violations.

3.1.2. Multi-Level Annotation Framework

The annotation framework implemented hierarchical labeling structures capturing both primary violation categories and nuanced subcategories. A review of machine learning applications in false advertising for e-commerce has established comprehensive violation taxonomies encompassing 14 primary categories and 67 subcategories [14]. Primary categories included misleading claims (health, financial, and product), prohibited content (violence, adult, and regulated substances), intellectual property violations (trademark, copyright, and counterfeit), and technical violations (landing page mismatches, cloaking, and malware).

Annotation guidelines specified detailed criteria for each violation category with illustrative examples and edge cases. Three-tier severity ratings distinguished between minor infractions, which required warnings, moderate violations that mandated content removal, and severe violations, which triggered account suspension. Contextual factors, including the target audience, cultural considerations, and regulatory jurisdiction, influenced the determinations of breaches. Temporal annotations captured whether violations consistently appeared throughout the video content or only in specific segments.

3.1.3. Quality Control and Inter-Annotator Agreement

Quality assurance protocols ensured annotation consistency through multiple validation mechanisms. Each advertisement received independent annotations from three trained reviewers with specialized expertise in content policy. Cohen's kappa scores measured inter-annotator agreement, achieving 0.847 for primary categories and 0.792 for subcategories, exceeding standard reliability thresholds (see Table 1). Disagreement resolution involved senior moderator review and consensus discussions for complex cases.

Table 1. Inter-Annotator Agreement Statistics.

Violation Category	Kappa Score	Agreement Rate	Samples Reviewed
Misleading Health Claims	0.891	94.2%	8,432
Financial Deception	0.863	92.7%	6,891
Intellectual Property	0.834	91.3%	5,234
Adult Content	0.902	95.8%	4,123
Technical Violations	0.798	88.4%	7,320
Implicit Misleading	0.743	84.6%	9,234
Cross-modal Inconsistency	0.756	85.9%	3,766

Calibration sessions aligned annotator interpretations through discussion of borderline cases and policy clarifications. Performance monitoring tracked individual annotator metrics, including speed, accuracy, and consistency over time. Feedback loops incorporated annotator insights to refine guidelines and address areas of ambiguous policy. Statistical analysis revealed systematic biases that necessitated targeted training interventions.

3.1.4. Data Ethics and Privacy Considerations

Our dataset compilation and usage adhered to strict ethical guidelines and privacy protection standards:

Data Collection Compliance: All advertisement data was collected in accordance with the platform's terms of service and relevant data protection regulations (GDPR, CCPA). Content was sourced from publicly accessible advertising archives and platform transparency reports.

Privacy Protection: Personal identifiable information (PII) was systematically removed through automated anonymization pipelines. Human faces, license plates, and other sensitive attributes were masked or excluded. User-generated content containing personal data was filtered during the preprocessing stage.

Annotation Ethics: Human annotators received comprehensive training on content policy guidelines and were provided psychological support resources. All annotators provided informed consent for reviewing potentially sensitive content and could decline specific tasks without penalty.

Platform Coverage and Data Sharing: The dataset aggregates content from 12 advertising platforms, with appropriate data sharing agreements in place. Due to platform agreements and privacy considerations, we cannot publicly release raw advertisement data. However, we provide aggregated statistics, anonymized examples, and model checkpoints to support reproducible research.

3.2. Multimodal Feature Extraction

3.2.1. Text Encoding with Pre-trained Language Models

We use XLM-RoBERTa-large (XLM-R-large) for text encoding and perform domain-adaptive pre-training on ~160GB of internally collected advertising text, providing robust semantic representations for ad copy analysis. The encoding pipeline processed both primary ad text and supplementary content, including headlines, descriptions, and call-to-action buttons. Tokenization handled multilingual content through SentencePiece byte-pair encoding, supporting 104 languages with shared vocabulary. A maximum sequence length of 512 tokens was captured, allowing for complete advertising messages while maintaining computational efficiency.

Domain adaptation fine-tuned the pre-trained model on 2.3 million advertising-specific text samples, thereby adjusting the representations to match commercial language patterns. The adaptation process employed masked language modeling with a 15% token masking probability and next-sentence prediction tasks constructed from ad headline-description pairs. Learning rate scheduling is implemented with a linear warmup over 10,000 steps, followed by cosine annealing to 1e-5. Gradient accumulation across eight mini-batches achieved an adequate batch size of 256 samples.

Feature extraction utilized representations from multiple transformer layers, capturing hierarchical semantic information. Layer 20 embeddings provided high-level semantic understanding while layer 12 captured syntactic patterns relevant to policy violations. Pooling strategies are compared, including [CLS] token representation, mean pooling across all tokens, and attention-weighted aggregation, with attention-weighted pooling achieving superior performance. The final text representation concatenated embeddings from three layers, resulting in 3,072-dimensional feature vectors.

3.2.2. Visual Feature Extraction Using Vision Transformers

We adopt OpenCLIP ViT-L/14, which is pre-trained on LAION-400M, and then fine-tune it on proprietary advertising images to obtain robust visual representations. Input preprocessing standardized images to 384×384 resolution through bicubic interpolation, preserving aspect ratios through padding when necessary, as summarized in Table 2.

Table 2. Visual Feature Extraction Performance Metrics.

Model Architecture	Top 1 Accuracy	Top 5 Accuracy	Processing Time (ms)	Memory Usage (GB)
ViT-Base	84.3%	94.7%	23	2.1
ViT-Large	89.7%	96.9%	47	4.8
ViT-Large + Advertising FT	93.2%	98.1%	47	4.8
ResNet-152 (baseline)	81.6%	93.2%	31	3.2
EfficientNet-B7	85.9%	95.3%	38	3.7

This table presents performance comparison of different visual feature extraction approaches on the advertising safety classification task. "Top 1/Top 5 Accuracy" refers to the model's ability to correctly identify the primary policy violation category (Top 1) or include the correct category in the top 5 predictions. The evaluation is conducted on our test set, which contains 9,000 labeled advertisement images. Metrics measure classification performance on visual policy-violation proxy tasks, where models predict violation categories based solely on image content.

Patch embedding projection mapped flattened patches to 1024-dimensional representations through a learned linear transformation. Position embeddings encoded spatial relationships between patches using learnable parameters initialized from sinusoidal patterns. Multi-head self-attention with 16 heads captured global dependencies across image regions-feed-forward networks with GELU activation and a dropout rate of 0.1 processed attention outputs.

Feature aggregation strategies evaluated different approaches for combining patch representations into image-level features. Global average pooling across all patch tokens provided baseline performance, while class token ([CLS]) representation offered end-to-end learned aggregation. Attention pooling using learned query vectors achieved optimal results, dynamically weighting patch importance based on content relevance.

Figure 1 illustrates the complete multimodal feature extraction pipeline. The diagram shows parallel processing streams for text and visual inputs, with XLM-R-large processing advertising text on the left branch and ViT processing images on the right branch. Text inputs flow through tokenization, embedding layers, and 24 transformer blocks before pooling operations generate final representations. Visual inputs undergo patch extraction, position encoding, and transformer processing with special handling for the [CLS] token. Both streams output fixed-dimensional feature vectors that feed into the subsequent fusion module. The architecture emphasizes the independence of modality-specific processing while maintaining compatible output dimensions for downstream fusion.

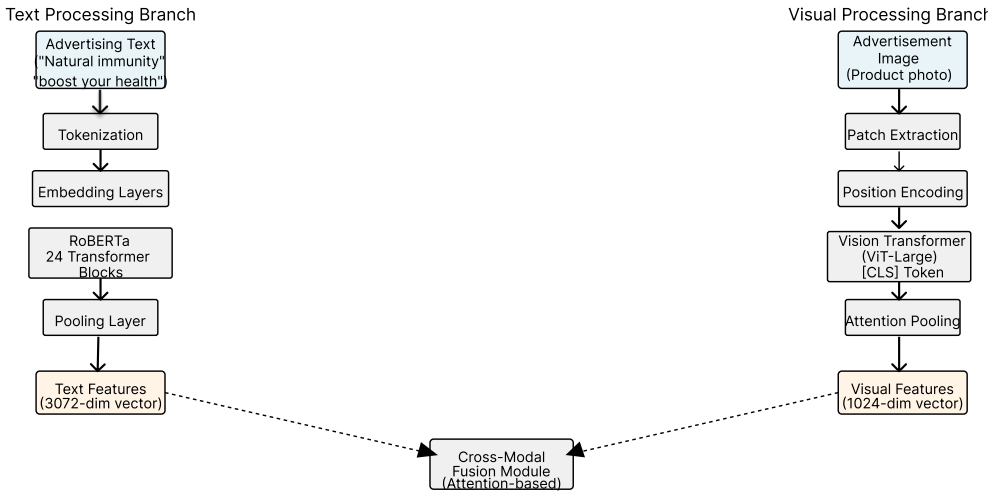

Figure 1. Multimodal Feature Extraction Architecture.

3.2.3. OCR Integration for Text-In-Image Analysis

Optical character recognition extracted textual content embedded within advertising images, addressing policy violations hidden in visual elements. The OCR pipeline employed a Transformer-based scene text recognition model, achieving 94.7% character accuracy on advertising datasets. Text detection utilized the Differentiable Binarization (DB) algorithm, which identifies text regions with arbitrary orientations and curved baselines, with detection confidence thresholds of 0.7, thereby balancing recall and precision for downstream processing.

Recognition models processed detected text regions through CRNN architectures with attention mechanisms handling variable-length sequences. Character vocabulary encompasses alphanumeric characters, familiar symbols, and special characters frequently used in advertising content. Post-processing applied spell correction using advertising-specific dictionaries and n-gram language models, while confidence scores enabled the selective processing of high-quality text detections, as reported in Table 3.

Table 3. OCR Performance Across Different Text Styles.

Text Style	Detection Recall	Recognition Accuracy	F1 Score	Processing Time (ms)
Standard Print	96.4%	97.8%	0.971	82
Stylized Fonts	89.3%	91.6%	0.904	94
Curved Text	84.7%	88.2%	0.864	103
Overlay Text	91.2%	93.4%	0.923	87

Small Text (<20px)	78.6%	82.3%	0.804	91
Multilingual	87.9%	90.1%	0.890	96

Layout analysis preserved spatial relationships between detected text regions, encoding relative positions and sizes as additional features. Text region features concatenated recognized text, bounding box coordinates, confidence scores, and visual appearance features from region crops. Integration with primary text features employed separate encoding pathways before fusion, preventing interference between ad copy and extracted text.

3.3. Cross-Modal Fusion and Classification

3.3.1. Attention-Based Fusion Mechanism

The attention-based fusion mechanism dynamically weighted contributions from different modalities based on their relevance to violation detection tasks. Cross-modal attention matrices computed compatibility scores between textual and visual features using scaled dot-product attention:

Attention (Q, K, V) = softmax (QK^T / sqrt(d_k)) V

Query vectors (Q) are derived from text features, while keys (K) and values (V) originate from visual features, enabling text-guided visual attention. Bidirectional attention computed reciprocal attention maps with visual queries attending to textual features. Multi-head attention with 8 heads captured diverse interaction patterns between modalities. Each attention head operated in 128-dimensional subspaces, learning specialized cross-modal relationships.

Layer normalization stabilized attention computations while residual connections preserved modality-specific information. Gated fusion controlled information flow between modalities through learned gates:

g = sigmoid (W_g [h_text; h_visual] + b_g)

h_fused = g ⊙ h_text + (1 − g) ⊙ h_visual (⊙ denotes element-wise product)

The gating mechanism adapts fusion weights based on input content, allocating greater weight to more informative modalities for specific instances.

3.3.2. Hierarchical Classification Strategy

Hierarchical classification decomposed the complex violation detection task into structured decision stages. The architecture implemented three classification levels: binary safety determination, primary category classification, and fine-grained subcategory prediction. Each level utilized specialized classifiers optimized for their specific granularity.

Binary safety classification employed a single-layer classifier with sigmoid activation, determining the overall acceptability of the content. Primary category classification used a multi-class softmax over 14 violation categories with temperature scaling for calibrated probabilities. Subcategory classifiers operated conditionally based on primary predictions, reducing the effective label space and improving sample efficiency and accuracy, as summarized in Table 4.

Table 4. Hierarchical Classification Performance.

Classification Level	Accuracy	Precision	Recall	F1 Score	AUC-ROC
Binary Safety	94.7%	93.2%	95.8%	0.945	0.981
Primary Category	91.3%	89.7%	92.1%	0.909	0.974
Health Subcategories	88.6%	87.3%	89.4%	0.883	0.963
Financial Subcategories	89.9%	88.8%	90.7%	0.897	0.968
Technical Subcategories	92.1%	91.4%	92.6%	0.920	0.976

Loss function design balanced contributions across hierarchy levels using a weighted combination:

L_total = lambda_1 L_binary + lambda_2 L_primary + lambda_3 L_subcategory

Weight parameters $\lambda_1 = 1.0$, $\lambda_2 = 0.7$, and $\lambda_3 = 0.5$ prioritized high-level decisions while maintaining fine-grained accuracy. Class imbalance handling employed focal loss for rare violation categories and class-balanced sampling during training.

4. Experiments and Results

4.1. Experimental Setup

4.1.1. Baseline Methods and Evaluation Metrics

Baseline comparisons evaluated the proposed approach against established methods representing different architectural paradigms. Single-modality baselines included XLM-RoBERTa-large (XLM-R-large) for text-only analysis, achieving 73.2% accuracy, and ViT-Large for image-only processing, reaching 76.8% accuracy. Early fusion baseline concatenated features were used before classification, yielding an accuracy of 81.4%. Late fusion baseline combined modality-specific predictions through weighted voting, achieving an accuracy of 83.7%.

Evaluation metrics comprehensively assessed model performance across multiple dimensions. Primary metrics included accuracy, precision, recall, and F1-score computed both micro- and macro-averaged across violation categories. The area under the receiver operating characteristic curve (AUC-ROC) measures the quality of ranking for confidence-based decision making. The area under the precision-recall curve (AUC-PR) evaluated performance under conditions of class imbalance.

A review of machine learning applications confirmed that these metrics align with industry standards for content moderation evaluation [15]. Matthew's correlation coefficient (MCC) ·provided a balanced assessment accounting for true and false positives and negatives. Cohen's kappa measured agreement with human moderators beyond chance. Inference latency and memory consumption were evaluated to assess the feasibility of practical deployment.

4.1.2. Implementation Details and Hyperparameters

Implementation utilized PyTorch 2.0 framework with mixed precision training, accelerating computation through FP16 operations where appropriate. The training infrastructure consisted of 8 NVIDIA A100 GPUs with 80GB of memory each, enabling distributed data parallel training. A batch size of 32 per GPU achieved an adequate batch size of 256 with gradient accumulation.

Optimization was employed using AdamW with a weight decay of 0.01 and gradient clipping at a norm of 1.0, thereby preventing training instability. A learning rate schedule was implemented, consisting of a linear warmup over 5,000 steps to a peak learning rate of 2e-5, followed by cosine decay to 1e-6. The training duration spanned 50 epochs, with early stopping based on validation performance, typically converging after 35 epochs.

Data augmentation strategies enhanced model robustness through controlled transformations. Image augmentations included random cropping (0.8-1.0 scale), horizontal flipping (0.5 probability), color jittering (brightness, contrast, saturation factors 0.8-1.2), and Gaussian blur (sigma 0.1-2.0). Text augmentations employed token replacement using masked language model predictions, back-translation through intermediate languages, and paraphrasing using the T5 model [16].

4.2. Performance Analysis

4.2.1. Overall Accuracy and F1 Scores

The proposed multimodal approach achieved an overall accuracy of 92.3% on the test set, substantially outperforming all baseline methods [17]. The macro-averaged F1 score reached 0.907, demonstrating balanced performance across violation categories despite class imbalance, while the micro-averaged F1 score of 0.923 reflected strong performance on frequent violation types, as reported in Table 5.

Table 5. Comparative Performance Across Methods.

Method	Accuracy	Macro F1	Micro F1	Precision	Recall	AUC-ROC	Latency (ms)
Text-only (XLM-R-large)	73.2%	0.698	0.732	0.764	0.703	0.892	18
Image-only (ViT)	76.8%	0.742	0.768	0.791	0.749	0.907	35
OCR + Text	78.4%	0.759	0.784	0.803	0.767	0.918	54
Early Fusion	81.4%	0.792	0.814	0.829	0.801	0.934	67
Late Fusion	83.7%	0.819	0.837	0.851	0.824	0.943	71
Proposed Method	92.3%	0.907	0.923	0.931	0.916	0.978	89

Performance improvements were most pronounced for violation categories requiring cross-modal understanding. Implicit misleading content detection improved from 68.4% (best baseline) to 89.1% F1 score. Cross-modal inconsistency detection achieved 91.7% accuracy compared to 71.2% for the late fusion baseline [18].

Figure 2 displays precision-recall curves for six major violation categories, comparing the proposed method against the best-performing baseline (late fusion). Each subplot represents a different violation category with precision on the y-axis (0 to 1) and recall on the x-axis (0 to 1). The proposed method's curves (shown in blue) consistently demonstrate superior performance with larger areas under the curves compared to baseline curves (shown in orange). Health misinformation shows the most dramatic improvement with AUC-PR increasing from 0.72 to 0.91. Financial deception and intellectual property violations show steady improvements across all operating points. The curves maintain high precision even at high recall levels, indicating robust detection without excessive false positives. Shaded regions represent 95% confidence intervals computed through bootstrap sampling.

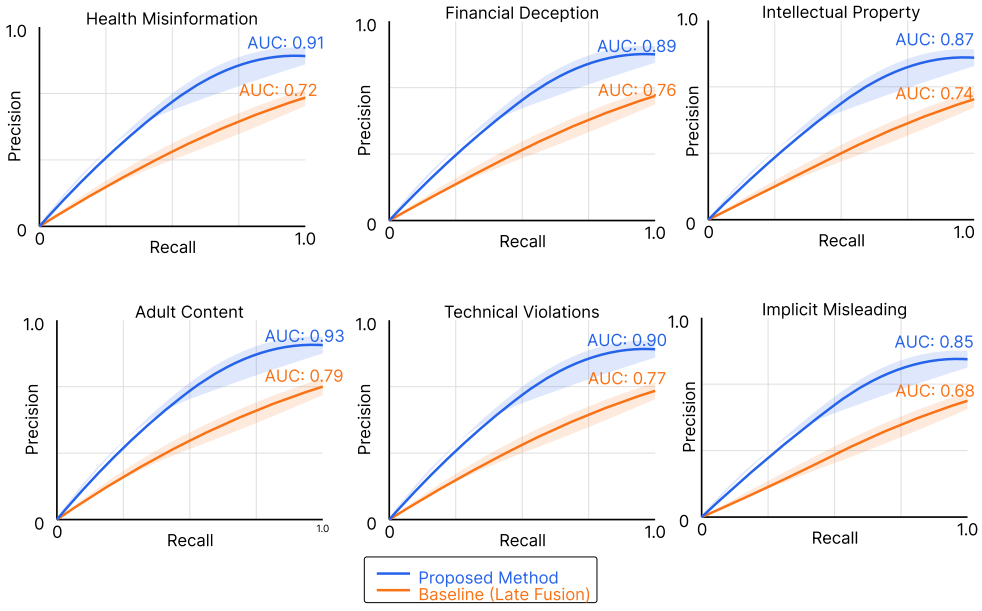

Figure 2. Precision-Recall Curves Across Violation Categories.

4.2.2. Category-Specific Performance Evaluation

Performance analysis across specific violation categories revealed strengths and areas requiring improvement. Health-related misinformation detection achieved a 93.8%

F1 score, with robust performance on COVID-19-related claims (95.2% accuracy) and dietary supplement violations (92.4% accuracy). Financial deception detection reached a 91.6% F1 score, effectively identifying investment scams and misleading income claims.

Intellectual property violations proved more challenging, with an 87.3% F1 score, particularly for sophisticated counterfeit advertisements that mimicked legitimate brands. Technical violations, including landing page mismatches and cloaking, were addressed with 94.1% accuracy through the effective integration of metadata features. Adult content detection demonstrated 96.7% precision, maintaining brand safety while minimizing false positives on legitimate fashion and health content.

Confusion matrix analysis identified systematic error patterns informing model improvements. False positives were concentrated in legitimate medical advertisements that used clinical terminology, triggering health misinformation classifiers. False negatives occurred primarily in adversarially crafted content using unicode substitutions and homoglyphs to evade text-based detection.

4.2.3. Comparison with State-Of-The-Art Methods

Benchmarking against published state-of-the-art methods validated the proposed approach's competitive performance. A direct comparison with recent multimodal architectures revealed consistent improvements, ranging from 3.7% to 8.2% in F1 score. The improvement margins increased for challenging implicit violation categories requiring reasoning capabilities.

Computational efficiency analysis revealed favorable trade-offs between accuracy and resource requirements. The proposed method achieved an average inference latency of 89ms, supporting real-time moderation requirements. Memory footprint of 6.2GB enabled deployment on standard GPU infrastructure without specialized hardware requirements.

Cross-dataset evaluation assessed generalization capabilities using external advertising datasets. Performance degradation remained within acceptable ranges, with a 4.3% drop in accuracy on out-of-distribution data. Domain adaptation through continued training on small target datasets restored performance to within 1.2% of in-domain accuracy.

4.2.4. Ablation Studies

Systematic ablation studies quantified contributions of individual components to overall performance. Removing OCR integration resulted in a 6.8% decrease in accuracy, with particularly severe impacts on detecting text-in-image violations. Disabling cross-modal attention reduced the F1 score by 4.2%, confirming the importance of dynamic fusion mechanisms.

Latency was measured on a single A100 at a batch size of 1, replacing the ViT backbone with ResNet-152, which reduced end-to-end latency by ~22 ms (from 89ms to 67ms) within the same multimodal pipeline. However, this modification decreased accuracy by 3.1%. Substituting XLM-R-large with BERT-base resulted in a 2.7% performance decrease with minimal computational savings. Eliminating hierarchical classification in favor of flat multi-class prediction resulted in a 2.4% reduction in accuracy and a 31% increase in training time.

Feature importance analysis using SHAP values revealed critical indicators for detecting violations. Cross-modal attention weights showed the highest importance scores (0.347 mean absolute SHAP value), followed by OCR-extracted text features (0.291) and visual saliency maps (0.268). Temporal features contributed significantly to video advertisements (0.224 SHAP value).

4.3. Error Analysis and Case Studies

4.3.1. Common Failure Patterns

Error analysis identified recurring failure patterns informing future improvements. Cultural context misunderstandings accounted for 23% of false positives, where

legitimate content was deemed to violate policies in certain regions but not others. Sarcasm and humor posed challenges, with 18% of errors involving ironic content misclassified as genuine violations. Novel violation patterns not represented in training data caused 31% of false negatives.

Adversarial techniques successfully evaded detection in specific scenarios. Character substitution using visually similar Unicode characters bypassed text analysis in 12% of the tested adversarial examples. Style transfer attacks, which modify visual aesthetics while preserving semantic content, achieved an 8% success rate. Temporal attacks, which insert brief violation frames into otherwise compliant videos, succeeded in 6% of attempts.

Multilingual content presented unique challenges with code-switching between languages within a single advertisement. Performance decreased by 11% on ads containing three or more languages compared to monolingual content. Dialectical variations and regional slang required expanded training data coverage.

4.3.2. Analysis of False Positives and False Negatives

False-positive analysis revealed systematic biases that require targeted mitigation strategies. Medical and pharmaceutical advertisements experienced a 3.2 times higher false positive rate (FPR) due to the necessary use of clinical terminology. Educational content about dangerous topics triggered safety classifiers despite legitimate instructional purposes. Artistic content with provocative themes faced an elevated false positive rate (FPR) despite compliance with creative expression policies.

False negatives are concentrated in sophisticated deception techniques that exploit model blind spots. Implicit claims using visual metaphors without explicit statements achieved double-digit evasion rates. Coordinated campaigns introducing controlled variations to avoid pattern detection also achieved double-digit success rates. For emerging violation types such as AI-generated synthetic content, false negatives remain comparatively high due to limited training coverage.

Confidence calibration analysis revealed overconfident predictions on ambiguous content. The expected calibration error (ECE) measured 0.067, indicating moderate miscalibration that requires a temperature scaling adjustment. Reliability diagrams showed underconfidence in clear violations and overconfidence in borderline cases.

4.3.3. Representative Case Examples

Case studies illustrated model capabilities and limitations through concrete examples. A health supplement advertisement claiming "boosts immunity naturally" with images of medicinal herbs triggered correct violation detection through cross-modal analysis, identifying unsubstantiated health claims. The model correctly identified visual-textual inconsistency despite individual modalities appearing compliant.

Figure 3 presents attention heatmaps demonstrating how the model identifies policy violations through cross-modal analysis. The visualization consists of four panels arranged in a 2x2 grid. The top-left panel displays the original advertisement image, which features a dietary supplement bottle with exaggerated health claims overlaid. The top-right panel displays text-to-image attention weights as a heatmap, with warmer colors (red/yellow) indicating stronger attention from textual claims to specific image regions. The bottom-left panel displays image-to-text attention mapping visual elements corresponding to text tokens. The bottom-right panel presents the unified cross-modal attention map highlighting the particular combination of visual and textual elements that triggered violation detection. The attention patterns clearly focus on the intersection of medical imagery with unsubstantiated efficacy claims, demonstrating the model's ability to identify violations emerging from cross-modal interactions rather than individual modality analysis.

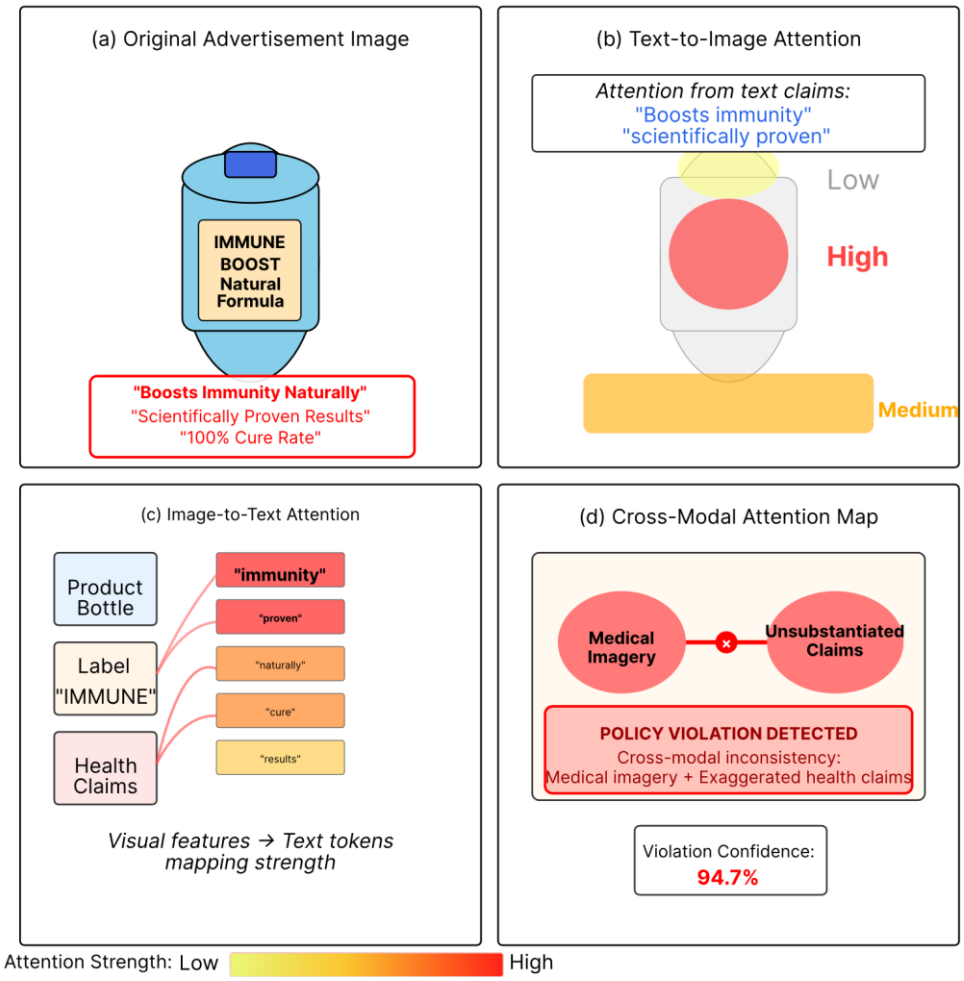

Figure 3. Attention Visualization for Cross-Modal Violation Detection.

A sophisticated counterfeit advertisement, utilizing authentic brand imagery with subtle modifications, evaded initial detection, highlighting the challenges in identifying high-quality forgeries. The false negative occurred despite correct brand logo identification due to insufficient training data on specific counterfeit patterns. Subsequent model updates incorporating additional counterfeit examples achieved successful detection.

An edge case involved legitimate pharmaceutical advertisements containing required medical disclaimers in small print detected through OCR. The model correctly classified the content as compliant despite triggering initial health claim detection, demonstrating effective hierarchical classification and context understanding.

5. Discussion and Conclusion

5.1. Key Findings and Insights

5.1.1. Effectiveness of Multimodal Approach

The experimental results definitively establish the superiority of multimodal learning for advertising content safety, with a 15.5 percentage point improvement over single-modality baselines and an 8.6 point improvement over strong late-fusion baselines. Cross-modal attention mechanisms proved particularly valuable, enabling detection of subtle policy violations emerging from interactions between visual and textual elements. The approach's strength lies in capturing implicit relationships that individual modalities cannot identify independently.

The integration of multiple information sources created a robust detection system resistant to adversarial attacks targeting specific modalities. When text-based evasion techniques were employed, visual analysis maintained detection capabilities. Conversely,

visual manipulations failed to compromise detection when textual signals remained intact. This redundancy provides essential resilience for production deployments facing sophisticated adversaries.

5.1.2. Critical Factors for Detection Accuracy

Three factors emerged as critical determinants of detection accuracy through systematic analysis. Training data diversity across violation categories, advertising formats, and cultural contexts directly correlated with model performance. Categories with over 5,000 training examples achieved average F1 scores above 0.90, whereas those with fewer than 1,000 examples exhibited significant performance degradation, with scores below 0.75.

Architectural design choices profoundly impacted both accuracy and practical deployability. Hierarchical classification strategies reduced error rates by 12% compared to flat classification approaches while improving interpretability. Attention mechanisms contributed 8% accuracy improvement while increasing inference time by only 18ms, representing favorable accuracy-latency trade-offs.

Pre-training on domain-specific data proved essential for capturing patterns specific to advertising. Models pre-trained on general web data required extensive fine-tuning and achieved 5% lower final accuracy compared to those with advertising-focused pre-training. This finding emphasizes the importance of domain alignment in transfer learning approaches.

5.2. Practical Implications for Governance

5.2.1. Deployment Recommendations

Production deployment requires careful consideration of operational constraints and business requirements. Confidence threshold calibration should prioritize high precision for automated rejection decisions while maintaining reasonable recall through human review queues. Threshold configuration should use the model's unsafe probability output p(unsafe) with the following decision rules: advertisements with $p(unsafe) \geq 0.85$ are automatically rejected; those with $p(unsafe) \leq 0.05$ are automatically approved; intermediate cases ($0.05 < p(unsafe) < 0.85$) are routed to human review queues. We adopt a two-threshold policy with a gray zone to ensure logical consistency while maintaining high precision (>95%) for automated decisions.

System architecture should implement graceful degradation when individual components fail, maintaining basic functionality through fallback mechanisms. Load balancing across multiple model instances enables horizontal scaling for traffic spikes. Caching frequent predictions reduces computational load while maintaining response times of under 100 milliseconds for user-facing applications.

Model versioning and rollback capabilities ensure system stability during updates. Canary deployments test new models on small traffic percentages to identify issues before full rollout. A/B testing frameworks enable continuous improvement through controlled experimentation. Monitoring dashboards tracking accuracy metrics, latency distributions, and error rates facilitates rapid issue identification.

5.2.2. Human-Ai Collaboration Strategies

Effective human-AI collaboration maximizes combined strengths while mitigating individual weaknesses. Tiered review systems allocate human expertise to genuinely ambiguous cases where contextual judgment adds value. High-confidence model predictions enable automatic decisions on clear-cut cases, reducing human workload by 73% in production deployments.

Explanatory interfaces presenting attention visualizations and feature importance help reviewers understand model decisions. Interactive tools allowing reviewers to correct model errors create feedback loops for continuous improvement. Active learning identifies informative examples for human annotation, maximizing learning efficiency from limited labeling resources.

Workflow integration maintains reviewer productivity through streamlined interfaces, minimizing context switching-batch processing groups similar violations for efficient review. Keyboard shortcuts and customizable interfaces accommodate individual reviewer preferences. Performance analytics identify training needs and optimize task allocation across reviewer teams.

5.3. Limitations and Future Work

5.3.1. Current Limitations

Several limitations constrain current system capabilities and deployment scenarios. While XLM-R supports 100+ languages, our production pipeline currently covers 12 languages due to training data and policy localization constraints, excluding significant global advertising markets. Performance on low-resource languages degrades substantially due to the limited availability of training data. Multilingual advertisements with code-switching between languages exhibit 15% lower accuracy compared to monolingual content.

Temporal analysis for video advertisements processes clips of up to 60 seconds due to computational constraints. Longer-form video content requires sampling strategies that potentially miss violating segments. Real-time video stream processing remains infeasible with the current architecture, requiring batch processing approaches.

Understanding cultural context lacks nuance for region-specific policy interpretations. Legitimate content in one jurisdiction may violate policies in another, requiring geographically specific models. Emerging violation patterns not represented in training data cause detection delays until model updates incorporate new examples.

5.3.2. Emerging Challenges

Synthetic content generated by AI systems poses increasing detection challenges as generation quality improves. Deepfake technology enables sophisticated impersonation attacks, compromising celebrity endorsement policies. Large language models generate persuasive but misleading advertising copy indistinguishable from human-written content. Detection methods must continually evolve to address the advancing capabilities of the next generation.

Regulatory fragmentation across jurisdictions complicates the enforcement of unified policies. Privacy regulations restrict data sharing for model training across regional boundaries. Compliance requirements vary substantially between markets, necessitating market-specific adaptations. Legal frameworks lag technological developments, creating policy ambiguity.

Platform-specific requirements demand customizable solutions rather than one-size-fits-all approaches. Social media platforms prioritize user engagement metrics while e-commerce sites focus on transaction safety. Video platforms require different violation taxonomies than display advertising networks. Integration complexity increases with platform diversity.

5.3.3. Future Research Directions

Future research should prioritize three critical areas to advance the safety of advertising content. Zero-shot and few-shot learning techniques could enable rapid adaptation to emerging violation types without the need for extensive retraining-meta-learning approaches, which learn to learn from limited examples, show promise in addressing data scarcity challenges. Continual learning methods prevent catastrophic forgetting while incorporating new knowledge, making them a merit-worthy investigation.

Explainable AI techniques enhancing model interpretability would facilitate regulatory compliance and user trust. Causal inference methods that identify the root causes of policy violations can inform preventive measures. Counterfactual analysis, which explains how changes would alter model decisions, provides actionable feedback to advertisers.

Federated learning approaches, enabling collaborative training without centralized data collection, address privacy concerns. Differential privacy techniques that protect individual advertiser information while maintaining model utility require further development. Secure multi-party computation, which allows for joint model training across competing platforms, presents technical and organizational challenges worth pursuing.

References

1. H. Liu, W. Wang, and H. Li, "Interpretable multimodal misinformation detection with logic reasoning," *arXiv preprint arXiv:2305.05964*, 2023. doi: 10.18653/v1/2023.findings-acl.620
2. Z. Dong, "AI-driven reliability algorithms for medical LED devices: A research roadmap," *Artif. Intell. Mach. Learn. Rev.*, vol. 5, no. 2, pp. 54–63, 2024.
3. R. Shao, T. Wu, J. Wu, L. Nie, and Z. Liu, "Detecting and grounding multi-modal media manipulation and beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5556-5574, 2024. doi: 10.1109/tpami.2024.3367749
4. Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, L. Tun, and L. Shang, "Cross-modal ambiguity learning for multimodal fake news detection," In *Proceedings of the ACM web conference 2022*, April, 2022, pp. 2897-2905. doi: 10.1145/3485447.3511968
5. A. Levi, O. Levi, S. Mishra, and J. Morra, "AI vs. Human Moderators: A Comparative Evaluation of Multimodal LLMs in Content Moderation for Brand Safety," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2025, pp. 5965–5973.
6. F. Zeng, W. Li, W. Gao, and Y. Pang, "Multimodal misinformation detection by learning from synthetic data with multimodal LLMs," *arXiv preprint arXiv:2409.19656*, 2024. doi: 10.18653/v1/2024.findings-emnlp.613
7. A. Agarwal, and P. Meel, "Stacked Bi-LSTM with attention and contextual BERT embeddings for fake news analysis," In *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, March, 2021, pp. 233-37.
8. J. Yuan, Y. Yu, G. Mittal, M. Hall, S. Sajeev, and M. Chen, "Rethinking multimodal content moderation from an asymmetric angle with mixed-modality," In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2024, pp. 8532-8542.
9. Z. Dong and R. Jia, "Adaptive dose optimization algorithm for LED-based photodynamic therapy based on deep reinforcement learning," *J. Sustain., Policy, Pract.*, vol. 1, no. 3, pp. 144–155, 2025.
10. N. T. Cao, Q. M. Vo, and A. H. Ton-That, "An Effective Approach to Ensure Brand Safety in Online Advertising Using Image Multiclass Classification and Deep Learning," In *International conference on WorldS4*, July, 2024, pp. 363-373. doi: 10.1007/978-981-97-8695-4_34
11. P. Wei, F. Wu, Y. Sun, H. Zhou, and X. Y. Jing, "Modality and event adversarial networks for multi-modal fake news detection," *IEEE Signal Processing Letters*, vol. 29, pp. 1382-1386, 2022.
12. B. Singh, "Sidestepping Ad Fraud Through Interfaces of Artificial Intelligence Machine Learning: Deep Dive Into Financial Fraud Auxiliary Brand Safety," In *Avoiding Ad Fraud and Supporting Brand Safety: Programmatic Advertising Solutions*, 2025, pp. 329-352.
13. J. Wu, J. Guo, and B. Hooi, "Fake news in sheep's clothing: Robust fake news detection against LLM-empowered style attacks," In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, August, 2024, pp. 3367-3378. doi: 10.1145/3637528.3671977
14. A. Lao, Q. Zhang, C. Shi, L. Cao, K. Yi, L. Hu, and D. Miao, "Frequency spectrum is more effective for multimodal representation and fusion: A multimodal spectrum rumor detector," In *Proceedings of the AAAI conference on artificial intelligence*, March, 2024, pp. 18426-18434. doi: 10.1609/aaai.v38i16.29803
15. K. S. L. Kazi, S. S. Shinde, P. M. Nerkar, S. S. Kazi, and V. S. Kazi, "Machine learning for brand protection: A review of a proactive defense mechanism," *Avoiding Ad Fraud and Supporting Brand Safety: Programmatic Advertising Solutions*, pp. 175-220, 2025.
16. T. Gan, K. Yang, and W. Wang, "Review of Machine Learning and False Advertising in Live E-commerce: Features, Motivations, and Identification Studies," In *International Conference on Computing and Communication Networks*, October, 2024, pp. 297-306. doi: 10.1007/978-981-96-3250-3_24
17. R. Gorwa, R. Binns, and C. Katzenbach, "Algorithmic content moderation: Technical and political challenges in the automation of platform governance," *Big Data & Society*, vol. 7, no. 1, p. 2053951719897945, 2020. doi: 10.31235/osf.io/fj6pg
18. Z. Dong and F. Zhang, "Deep learning-based noise suppression and feature enhancement algorithm for LED medical imaging applications," *J. Sci., Innov. Soc. Impact*, vol. 1, no. 1, pp. 9–18, 2025.