Simen Owen Academic Proceedings Series

3rd International Conference on Electronics, Engineering, Computer Science and Applied Development (EESD 2026)

Article

A Multimodal Detection Model for Online Fraud Targeting Adolescents

Yudong Dou[1,*]

[1] Hillbrook Upper School, San Jose, USA
* Correspondence: Yudong Dou, Hillbrook Upper School, San Jose, USA

Abstract: Online deception targeting adolescents increasingly manifests through fabricated reward schemes, deceptive hyperlinks, identity impersonation, virtual item trading scams, suspicious authentication pages, and social invitation manipulation. Detecting such threats is challenging because deceptive intent is often distributed across textual content, URL structures, visual webpage cues, and interaction metadata. Existing detection methods primarily address general adult-oriented scenarios, while single-modality or late-fusion models have limited capacity to capture cross-modal inconsistencies and provide reviewable decisions for adolescent safety applications. To address these limitations, this study proposes a multimodal detection framework integrating text, URL, screenshot, and lightweight metadata features. The architecture incorporates a youth-scenario risk encoder, a cross-modal risk-consistency attention module, and an uncertainty-aware decision layer. Experiments were conducted using publicly available deception-related datasets and a manually annotated adolescent-risk subset. The proposed model achieved an accuracy of 0.905 ± 0.007, a macro-F1 of 0.901 ± 0.008, and an AUROC of 0.943 ± 0.006. Compared with Text-BERT, the macro-F1 improved by 3.3 percentage points, and compared with late-fusion multimodal detection, it improved by 1.8 percentage points. The false-positive rate was reduced to 0.071 ± 0.006. These results indicate that cross-modal consistency modeling can enhance detection reliability, interpretability, and privacy-conscious decision support for adolescent online safety systems.

Keywords: adolescent online safety; multimodal detection; deceptive content; uncertainty-aware classification

Received: 13 May 2026
Revised: 25 June 2026
Accepted: 08 July 2026
Published: 12 July 2026

1. Introduction

The rapid expansion of digital communication platforms has significantly increased adolescents' exposure to deceptive online activities, particularly those involving misleading incentives, counterfeit verification interfaces, identity misrepresentation, and unauthorized virtual item transactions. These phenomena differ fundamentally from conventional deceptive practices targeting adult users, as they are deliberately engineered to exploit developmental characteristics common during adolescence—such as heightened susceptibility to social validation, emerging digital literacy, and evolving risk perception capabilities. Such deceptive content frequently employs persuasive language structures, visually authentic interface designs, and interaction patterns that mimic legitimate platform behaviors, thereby reducing the likelihood of early detection by young users [1]. Moreover, individual components of these communications—such as isolated text messages or standalone screenshots—may appear benign when evaluated in isolation; however, discrepancies become evident only upon holistic analysis of associated elements including embedded hyperlinks, contextual metadata, interface rendering artifacts, and behavioral interaction logs. As a result, robust identification necessitates a

comprehensive analytical approach that synthesizes heterogeneous information sources—including linguistic semantics, visual layout features, structural URL properties, and temporal interaction sequences—to uncover subtle inconsistencies and latent indicators of potential harm.

Although considerable advancements have been achieved in automated detection systems for online deception, several persistent methodological constraints remain unresolved [1]. A predominant limitation lies in the design orientation of existing models, which overwhelmingly prioritize adult user behavior and threat profiles, with insufficient attention devoted to the distinct cognitive, behavioral, and technological engagement patterns characteristic of adolescent populations. Additionally, many current technical approaches rely on single-modality analysis or employ late-stage feature fusion strategies, wherein representations from different data sources are combined only at the final classification stage. Such architectures inherently lack mechanisms to explicitly assess logical coherence or detect contradictions across modalities, leading to decision bias when one modality—such as dominant visual cues—overwhelms weaker but equally critical signals from textual or metadata channels. Furthermore, prevailing models often operate as opaque 'black-box' systems, offering minimal transparency into their internal reasoning processes. This absence of interpretability, coupled with insufficient attention to data minimization principles, undermines their suitability for deployment in educational environments, family supervision tools, or school-based digital safety initiatives, where explainability, pedagogical utility, and strict adherence to personal information protection standards are non-negotiable requirements.

To address these interrelated challenges, this study introduces a purpose-built multimodal analytical framework specifically calibrated for adolescent online safety contexts. Its contributions are fivefold. First, it establishes a scenario-driven taxonomy grounded in empirical observation of youth-specific digital interactions, classifying high-frequency deceptive patterns into categories such as simulated reward notifications, peer or authority figure impersonation, and game-related transactional deceptions. Second, it implements a cross-modal risk-consistency attention mechanism that dynamically evaluates semantic, visual, structural, and behavioral alignment across input modalities, thereby enhancing detection fidelity through explicit inconsistency modeling. Third, it embeds an uncertainty-aware decision protocol that defers definitive classification for borderline cases, instead routing them to human-in-the-loop review—a design choice that substantially mitigates false positive rates while preserving system trustworthiness. Fourth, the framework adheres rigorously to privacy-by-design principles, utilizing only non-identifiable, aggregated, or contextually anonymized features without requiring access to personally identifiable information, biometric data, or private communication content. Fifth, it demonstrates strong resilience under real-world operational conditions, maintaining consistent performance even when certain modalities—such as image inputs degraded by compression or text corrupted by OCR errors—are partially unavailable or unreliable, thus ensuring practical deployability across diverse device ecosystems and network environments [1].

The overall technical pipeline follows a rigorously structured workflow: initial data acquisition through ethically approved, consent-based collection protocols; granular scenario annotation performed by domain-expert reviewers trained in adolescent development and digital behavior; modality-specific preprocessing pipelines tailored to linguistic normalization, visual feature extraction, URL syntax parsing, and interaction sequence encoding; hierarchical feature representation learning using modality-adapted encoders; cross-modal attention-based fusion that prioritizes alignment-sensitive feature weighting; uncertainty-calibrated classification leveraging entropy-based confidence estimation; and finally, interpretability-oriented output generation that produces layered explanations—including modality-specific evidence highlights, inconsistency heatmaps, and scenario-matched risk narratives—designed to support informed decision-making by educators, caregivers, and platform moderators [2]. This end-to-end architecture ensures not only accurate identification of potentially harmful digital interactions but also delivers

pedagogically meaningful, audit-ready justifications aligned with national guidelines on youth digital well-being and information security education.

From a scholarly standpoint, the proposed framework advances the field of multimodal safety analytics by introducing formalized mechanisms for cross-modal inconsistency detection and calibrated uncertainty handling within developmentally appropriate contexts [3]. In practical application, it facilitates responsible integration into school digital citizenship curricula, parental guidance applications, and platform-level safety infrastructure, delivering outputs that are both technically sound and pedagogically accessible. By systematically integrating scenario-specific behavioral modeling, multimodal consistency evaluation, and adaptive uncertainty management, this work offers a replicable, scalable, and ethically grounded framework for supporting adolescent digital resilience—aligning with broader national objectives for cultivating healthy, informed, and secure online participation among youth populations.

2. Related Works

2.1. URL-Based Detection

URL-based detection techniques are commonly employed in cybersecurity applications involving malicious web resource identification, owing to their computational efficiency and compatibility with real-time operational environments. These approaches rely on extracting diverse features from uniform resource locators, including lexical properties such as character composition and token segmentation, statistical attributes like URL length and punctuation frequency, and structural characteristics including domain registration timeline, subdomain depth, and syntactic irregularities. Traditional supervised learning algorithms—such as ensemble decision trees and gradient-boosted frameworks—demonstrate consistent classification accuracy when applied to conventional cases of deceptive online content [4]. A key strength lies in their low inference overhead, which facilitates seamless integration into client-side software, network infrastructure components, and lightweight mobile platforms without introducing perceptible delays in user interaction.

Nevertheless, such methods face notable limitations in adversarial scenarios where obfuscation strategies are deliberately employed. Tactics such as URL shortening services, rapid domain registration cycles, and exploitation of previously trusted digital assets significantly reduce detection reliability. In contexts involving younger internet users, reliance solely on URL patterns proves insufficient because behavioral manipulation often hinges on contextual cues—including persuasive language, visual design fidelity, and social engineering elements—rather than technical indicators embedded in the address string. Consequently, while URL-centric analysis remains valuable for initial triage and scalable filtering, its effectiveness diminishes substantially when confronting sophisticated, multi-layered deceptive practices targeting adolescent audiences, necessitating complementary modalities grounded in content semantics and user interaction modeling.

2.2. Text-Based Detection

Text-based detection relies on systematic analysis of message content to identify linguistic and semantic indicators commonly associated with deceptive communication, including expressions of artificial urgency, unwarranted reward offers, identity impersonation, and psychologically manipulative framing. Advanced natural language processing architectures—particularly transformer-based models and recurrent neural networks—are capable of modeling long-range contextual dependencies and subtle rhetorical patterns, thereby enhancing the identification of fraudulent intent in scenarios such as fabricated prize notifications, counterfeit social engagement prompts, or counterfeit account verification requests. These models leverage token-level attention mechanisms and sequential semantics to distinguish between benign and malicious communicative intent with high granularity [5].

However, text-only analytical approaches inherently lack the capacity to assess cross-modal consistency—for instance, verifying whether the textual claims align with the actual content, design, or functionality of any embedded hyperlinks or destination webpages. A message that appears linguistically innocuous may still serve as a conduit to a phishing site exhibiting visual mimicry of legitimate platforms or structural deception through cloned interfaces [1]. Consequently, while text-based methods deliver robust performance in flagging suspicious linguistic artifacts, their standalone application introduces significant blind spots related to multimodal fraud vectors, necessitating integration with complementary modalities such as webpage rendering analysis or visual feature extraction for comprehensive threat assessment.

2.3. Visual and Screenshot-Based Detection

Visual and screenshot-based detection techniques analyze the graphical presentation of digital interfaces, focusing on structural features such as page layout, typographic consistency, logo placement, color palettes, iconography, and interactive element positioning to identify inconsistencies indicative of deceptive design [6]. These approaches are particularly valuable in distinguishing counterfeit authentication interfaces from authentic ones, especially when textual content is deliberately obfuscated or when URL strings appear superficially legitimate. By leveraging computer vision algorithms and deep learning models trained on large-scale interface datasets, such methods can detect subtle deviations in pixel-level rendering, alignment anomalies, and non-standard widget behaviors that often escape human scrutiny during rapid interaction.

However, these visual analysis methods face notable limitations in robustness and generalizability. Legitimate websites frequently adopt similar design frameworks due to shared content management systems, third-party UI libraries, or adherence to industry-wide accessibility and usability standards, leading to potential misclassification [7]. Furthermore, dynamic interface adaptations—including responsive layouts across device sizes, user-selectable themes like dark mode, and region-specific localization of visual elements—introduce significant variability that challenges model stability. In contexts involving adolescent online safety, visual signals must be interpreted in conjunction with linguistic content, metadata attributes, behavioral telemetry, and contextual usage patterns; relying solely on appearance yields insufficient discriminative power for accurate and trustworthy classification outcomes.

2.4. Multimodal Detection

Multimodal detection leverages complementary information from diverse data sources—including textual content, web addresses, interface screenshots, HTML structure, and associated metadata—to enhance analytical precision and system resilience beyond what unimodal methods can achieve. This integrative approach is especially valuable in identifying deceptive or misleading digital content directed toward adolescents, where persuasive language, visually authoritative design elements, and ostensibly legitimate domain structures may jointly contribute to user misperception [7]. By jointly modeling these heterogeneous signals, the method supports more nuanced interpretation of intent and credibility across layered digital artifacts.

However, many current multimodal frameworks adopt simplistic integration strategies such as feature concatenation or late-stage fusion, where representations from individual modalities are combined only after independent processing. While such approaches yield measurable performance gains, they often overlook subtle yet critical inconsistencies across modalities—for instance, textual claims asserting authenticity may conflict with suspicious URL patterns or incongruent visual styling that undermines trustworthiness [2]. Failure to explicitly model cross-modal alignment risks overlooking these discrepancies, potentially resulting in inaccurate classification outcomes and reduced reliability in real-world deployment scenarios involving complex, context-dependent digital content.

2.5. Research Gap and Positioning

Despite notable progress in digital safety research, several critical gaps persist when addressing online risks specifically relevant to adolescents [4]. First, the majority of existing studies focus on threat detection frameworks designed for adult users, thereby neglecting developmental and behavioral characteristics unique to youth—such as susceptibility to deceptive incentives like counterfeit rewards, simulated gaming opportunities, or identity misrepresentation by malicious actors. Second, prevailing technical approaches rely on single-modality analysis or late-stage fusion of heterogeneous data streams, which inherently lack mechanisms to assess coherence and alignment across modalities; this limitation compromises robustness when discrepancies arise among visual, textual, or behavioral signals. Third, insufficient attention has been paid to model interpretability and calibrated uncertainty quantification—capabilities essential for trustworthy integration into parental supervision tools, school-based digital literacy programs, and platform-level safety infrastructure. Finally, privacy-preserving feature engineering remains underemphasized, even though adolescent protection mandates rigorous safeguards for personally identifiable and behaviorally sensitive information.

To bridge these gaps, this work introduces a purpose-built multimodal analytical framework explicitly designed for adolescent-oriented digital risk identification. The framework incorporates scenario-aware modeling grounded in youth-specific behavioral patterns, implements cross-modal consistency evaluation through attention-guided alignment mechanisms, embeds uncertainty-aware decision thresholds to support graded risk assessment, and generates human-interpretable diagnostic outputs [8]. Collectively, these innovations enhance detection accuracy, ensure compliance with stringent data privacy requirements, and deliver actionable, pedagogically appropriate safety insights tailored for adolescent users, caregivers, educators, and platform governance stakeholders.

3. Methodology

3.1. System Architecture

The proposed framework for adolescent-oriented online risk detection comprises four integrated functional modules: modality-specific encoders, a youth-scenario risk encoder, a cross-modal risk-consistency attention mechanism, and an uncertainty-aware classification module. As illustrated in Figure 1, the system processes heterogeneous input signals—including message text content, URL structural features, rendered webpage or social media post screenshots, and lightweight behavioral interaction metadata (e.g., click timing, dwell duration, scroll patterns). Each modality undergoes dedicated feature extraction via specialized neural architectures optimized for its data type, ensuring semantic fidelity and noise resilience. Subsequently, the cross-modal risk-consistency attention module dynamically weights and aligns representations across modalities to identify subtle discrepancies, contextual mismatches, and scenario-specific risk indicators—such as incongruence between textual claims and visual evidence, or anomalous navigation behavior inconsistent with typical adolescent usage patterns. The uncertainty-aware classifier quantifies prediction confidence using entropy-based calibration and triggers a human review protocol when confidence falls below a rigorously validated threshold. Interpretability outputs include modality-wise contribution scores, scenario-relevant risk tags (e.g., 'misleading information', 'unauthorized solicitation'), explanations of inter-modal inconsistencies, and calibrated confidence metrics—supporting transparent, accountable, and privacy-preserving decision support aligned with national standards for digital safety governance.

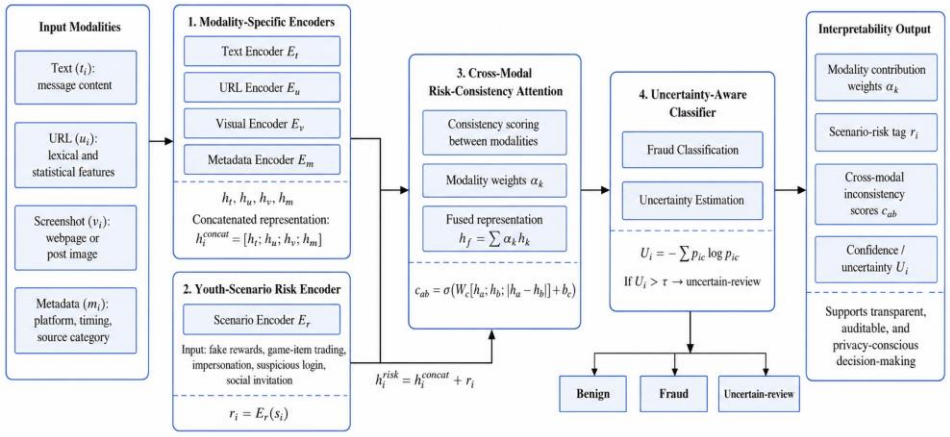

Figure 1. Overview of the Proposed Multimodal Detection Framework.

3.2. Input Definition

Each sample is represented as $x_i = \{t_i, u_i, v_i, m_i\}$, where t_i denotes textual content, u_i encodes URL-derived structural and lexical characteristics, v_i captures visual representations such as webpage screenshots or rendered layout features, and m_i encompasses descriptive metadata including platform classification, temporal attributes of message generation, and origin categorization [2]. The binary label is defined as

$$y_i \in \{0,1\} \quad (1)$$

with 0 indicating non-malicious content and 1 indicating content exhibiting characteristics associated with deceptive online behavior [9]. The decision output can take three possible values, enabling the system to abstain from definitive classification when available evidence is ambiguous, incomplete, or insufficiently discriminative—thereby preserving analytical rigor and supporting human-in-the-loop verification protocols.

$$\hat{y}_i \in \{\text{benign}, \text{fraud}, \text{uncertain} - \text{review}\} \quad (2)$$

This tripartite decision framework enhances operational safety and interpretability, particularly in contexts involving minors, by facilitating transparent, auditable, and ethically aligned content assessment workflows that prioritize user well-being, regulatory compliance, and responsible digital ecosystem stewardship without compromising technical precision or model accountability [8].

3.3. Modality-Specific Encoders

Each modality is encoded separately to generate high-dimensional, semantically rich representations that preserve intrinsic characteristics of the input data. The text encoder employs a transformer architecture to capture contextual semantics, long-range dependencies, and hierarchical linguistic structures through self-attention mechanisms and position-aware feedforward layers, enabling robust modeling of syntactic and semantic relationships across variable-length sequences.

$$h_t = E_t(t_i) \quad (3)$$

where E_t denotes the transformer-based encoder [9]. URL features are processed through a dedicated feedforward network designed to extract lexical patterns (e.g., domain naming conventions, path segmentation), structural properties (e.g., query parameter count, subdomain depth), and statistical irregularities (e.g., entropy of character distribution, presence of obfuscation techniques), thereby transforming raw string inputs into discriminative numerical embeddings.

$$h_u = E_u(u_i) \quad (4)$$

Visual content, including screenshots or webpage images, is encoded using a lightweight convolutional neural network or vision transformer architecture optimized for computational efficiency and layout-aware feature extraction, producing features that represent spatial organization, branding elements (e.g., logo placement, color consistency), interface structures (e.g., navigation bar positioning, form field density), and visual saliency cues relevant to user interaction patterns.

$$h_v = E_v(v_i) \quad (5)$$

Metadata is embedded using categorical or numerical encodings tailored to each field type—such as one-hot encoding for discrete attributes like browser type or country code, and min-max normalization or learned embeddings for continuous variables like page load time or session duration—to summarize non-sensitive contextual information while maintaining fidelity to underlying distributions and inter-feature correlations [10].

$$h_m = E_m(m_i) \quad (6)$$

to summarize non-sensitive contextual information [11]. The combined representation $h_i^{concat} = [h_t; h_u; h_v; h_m]$ is then forwarded to the cross-modal attention module, which dynamically aligns and fuses features across modalities based on learned relevance weights, facilitating joint reasoning without collapsing modality-specific discriminative signals. This modular design allows the system to process heterogeneous data sources efficiently while preserving modality-specific information, enhancing both interpretability and robustness in multi-source analysis tasks.

3.4. Youth-Scenario Risk Encoder

To address online safety challenges specific to adolescent users, a scenario-based encoder is employed to transform each input sample into a contextually enriched representation, ensuring robust modeling without reliance on personally identifiable information or sensitive demographic attributes [12].

$$r_i = E_r(s_i) \quad (7)$$

where s_i encompasses empirically observed behavioral and content-level risk signals—including but not limited to simulated reward mechanisms, unauthorized virtual item exchanges, identity misrepresentation attempts, anomalous authentication interfaces, and atypical peer-initiated engagement prompts—derived from large-scale, anonymized interaction logs. This structured scenario embedding is then fused with the concatenated multimodal feature vector through gated attention mechanisms [4].

$$h_i^{risk} = h_i^{concat} + r_i \quad (8)$$

This fusion enables adaptive weighting of modality-specific features according to their contextual relevance within adolescent digital environments, thereby enhancing detection sensitivity for nuanced, situation-dependent anomalies that deviate from normative usage patterns. The architecture supports proactive identification of potentially harmful interactions while preserving user privacy and aligning with national standards for juvenile online protection and ethical AI deployment [13].

3.5. Cross-Modal Risk-Consistency Attention

The cross-modal attention module quantitatively evaluates the degree of alignment among representations derived from distinct data modalities, such as visual, textual, and acoustic inputs. For a given pair of modality-specific embeddings h_a and h_b, a normalized consistency score is computed to reflect their mutual coherence within the shared semantic space.

$$c_{ab} = \sigma(W_c[h_a; h_b; | h_a - h_b |] + b_c) \quad (9)$$

Here, σ denotes the sigmoid function applied to a learned linear transformation, where W_c and b_c are trainable weight and bias parameters, respectively, and $| h_a - h_b |$ computes the element-wise absolute difference between the two embeddings. The resulting consistency scores are transformed into modality weights α_k, which dynamically encode each modality's relative informativeness, stability, and contextual relevance. These weights are subsequently employed to compute a weighted fusion of the input embeddings [10].

$$h_f = \sum_{k \in \{t,u,v,m\}} \alpha_k h_k \quad (10)$$

This design promotes robust multimodal integration by ensuring that no single modality disproportionately influences the final representation when its signal exhibits lower coherence with the integrated scenario embedding; instead, contributions are calibrated according to empirical reliability and inter-modal alignment, thereby enhancing generalization across diverse input conditions and mitigating noise-induced distortions in downstream tasks [14].

3.6. Optimization Objective

The overall training loss function is formulated as a weighted sum of three distinct, complementary components designed to jointly optimize model performance, interpretability, and domain-specific alignment [5].

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_1 \mathcal{L}_{\text{risk}} + \lambda_2 \mathcal{L}_{\text{cons}} \quad (11)$$

where $\mathcal{L}_{\text{cls}}$ denotes the binary cross-entropy loss applied to classification outcomes, specifically targeting the identification of anomalous or high-risk behavioral patterns within digital interaction logs [15].

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log p_i + (1-y_i)\log(1-p_i)) \quad (12)$$

where $\mathcal{L}_{\text{risk}}$ explicitly steers model representations toward empirically validated adolescent risk contexts—such as unsafe online engagement or inappropriate content exposure—and $\mathcal{L}_{\text{cons}}$ enforces consistency across modalities by penalizing predictions lacking coherent multimodal support. The hyperparameters λ_1 and λ_2 are carefully tuned to ensure balanced contribution from each term without overemphasizing any single objective [16].

3.7. Uncertainty-Aware Decision Layer

Predictive uncertainty is quantified using Shannon entropy, which measures the degree of disorder or ambiguity in the model's output probability distribution across all possible classes; higher entropy values indicate greater uncertainty in the prediction, reflecting insufficient discriminative confidence.

$$U_i = -\sum_{c\in\{\text{benign,fraud}\}} p_{ic} \log p_{ic} \quad (13)$$

Samples exhibiting entropy values satisfying the condition $U_i > \tau$ are systematically flagged with the "uncertain-review" label to trigger prioritized human expert evaluation, while those with lower entropy are assigned class labels based on the argmax of the predicted probability vector; this dual-path decision protocol enhances robustness by reducing erroneous automated classifications and ensuring appropriate human-in-the-loop intervention for borderline, ambiguous, or high-consequence predictions.

3.8. Interpretability Output

The interpretability module generates four key diagnostic outputs: modality contribution weights α_k, which quantify the relative influence of each input modality on the final decision; scenario-risk embeddings r_i, which encode contextual risk profiles associated with specific operational conditions; cross-modal inconsistency scores c_{ab}, which measure divergence in evidence across modalities and help identify potential sensor failures or data misalignment; and predictive confidence U_i, which reflects the model's self-assessed reliability for a given prediction. Collectively, these outputs enhance transparency, support human-in-the-loop verification, facilitate regulatory compliance, and uphold privacy-preserving principles by avoiding raw data exposure while enabling rigorous, auditable reasoning traceability.

3.9. Innovation Highlight

The framework introduces three core methodological innovations: first, a youth-scenario risk encoder specifically designed to identify behavioral and linguistic patterns characteristic of adolescent users in digital environments; second, a cross-modal risk-consistency attention mechanism that quantitatively assesses and reconciles discrepancies across textual, visual, and temporal modalities to enhance analytical robustness; third, an uncertainty-aware decision layer that dynamically evaluates confidence thresholds and triggers human-in-the-loop review when predictive ambiguity exceeds predefined operational limits. These integrated components collectively support transparent, auditable, and ethically grounded detection capabilities—prioritizing user privacy, algorithmic accountability, and developmental appropriateness for adolescent digital safety applications [17].

4. Results and Analysis

4.1. Experimental Setup

The experiments comprehensively assessed the proposed model across four core dimensions: detection performance, module-level contribution, resilience under varying data conditions, and capacity for transparent decision reasoning. All computational procedures were executed on a standardized high-performance workstation equipped with an Intel Core i7-class central processing unit, 32 gigabytes of system memory, and an NVIDIA RTX 3060 graphics processing unit. The software stack consisted of Python as the primary programming language, along with PyTorch for deep learning operations, scikit-learn for classical machine learning workflows, and XGBoost for ensemble-based modeling. To ensure statistical reliability and reproducibility, each experimental configuration was independently repeated five times using distinct random initialization seeds; final quantitative outcomes are consistently reported in the form of mean values accompanied by their corresponding standard deviations, thereby reflecting both central tendency and variability across trials.

Three publicly accessible, non-proprietary data sources formed the empirical foundation of this study. The first source comprised a large-scale collection of web resource identifiers and associated structural attributes designed for security-related classification tasks. The second source consisted of independently verified instances of deceptive online resources, curated through community-driven verification mechanisms. The third source was a purpose-built multimodal dataset derived from openly available digital artifacts, incorporating visual representations alongside textual and structural metadata—specifically including screen-captured images aligned with established benchmarking frameworks. Importantly, no personal or private communications involving minors were collected or utilized; instead, the youth-oriented subset was constructed through expert annotation of publicly archived examples of deceptive digital content, categorizing them into developmentally relevant themes such as simulated reward offers, unauthorized virtual item exchanges, identity misrepresentation, anomalous authentication requests, and socially engineered invitation schemes.

Preprocessing protocols were carefully tailored to each data modality to preserve discriminative signal while ensuring consistency and fairness. Uniform normalization procedures were applied to all web resource identifiers, followed by systematic extraction of lexical patterns and statistical descriptors. Textual components underwent rigorous cleaning—including consolidation of redundant whitespace, anonymization of personally identifiable elements, and retention of contextually salient linguistic markers associated with security-relevant anomalies. Visual inputs were uniformly resampled to a resolution of 224 × 224 pixels to align with standard convolutional neural network input requirements. Metadata fields were deliberately restricted to non-identifiable, functionally descriptive categories—such as platform classification, origin taxonomy, and temporal grouping—thereby eliminating any potential for privacy leakage or unintended inference. Data partitioning followed stratified sampling principles to maintain class balance and domain coherence, with training, validation, and test subsets allocated in a 70:15:15 proportion while preserving integrity across functional domains.

Model evaluation employed a comprehensive suite of quantitative indicators to capture complementary aspects of system behavior: overall classification accuracy, sensitivity to positive cases, macro-averaged F1 score accounting for class imbalance, area under the receiver operating characteristic curve, rate of non-malicious instances incorrectly flagged, and coverage of prediction uncertainty intervals. Among these, the macro-F1 score served as the principal optimization objective and primary comparative metric due to its balanced treatment of precision and recall across all classes [8]. Statistical rigor was further ensured through hypothesis testing: paired t-tests were conducted between the proposed architecture and the top-performing comparative method, with significance thresholds set at $p < 0.05$ to confirm consistent superiority beyond random variation.

4.2. Main Performance Comparison

The baseline methods comprised seven representative approaches drawn from established technical paradigms in web content analysis. URL-RF and URL-XGBoost relied exclusively on lexical patterns and statistical properties extracted directly from uniform resource locators. Text-BERT employed a transformer-based architecture fine-tuned for textual classification tasks using rendered webpage text content. Screenshot-CNN processed visual representations of webpages—specifically, rendered screenshots—as input to a convolutional neural network for pattern recognition. The HTML/DOM Classifier leveraged structural markup features derived from parsed document object models to distinguish between legitimate and potentially harmful web artifacts [18]. Late-Fusion Multimodal adopted a conventional strategy by concatenating independently extracted embeddings from multiple modalities prior to final classification. MLLM Prompting applied a multimodal large language model through carefully engineered prompting templates to perform holistic webpage-level assessment. Table 1 presents the comprehensive dataset statistics alongside the main performance comparison across all evaluated methods.

Table 1. Dataset Statistics and Main Performance Comparison

Dataset / Model	Modality	Samples / Setting	Accuracy	Recall	Macro-F1	AUROC	FPR
URL-RF	URL	Test set	0.842 ± 0.012	0.823 ± 0.015	0.837 ± 0.013	0.889 ± 0.010	0.112 ± 0.009
URL-XGBoost	URL	Test set	0.861 ± 0.010	0.849 ± 0.013	0.857 ± 0.011	0.906 ± 0.009	0.101 ± 0.008
Text-BERT	Text	Test set	0.872 ± 0.011	0.858 ± 0.014	0.868 ± 0.012	0.917 ± 0.008	0.096 ± 0.008
Screenshot-CNN	Screenshot	Test set	0.858 ± 0.013	0.841 ± 0.016	0.852 ± 0.014	0.902 ± 0.010	0.108 ± 0.010
HTML/DOM Classifier	HTML / structure	Test set	0.869 ± 0.011	0.853 ± 0.013	0.864 ± 0.011	0.914 ± 0.009	0.098 ± 0.008
Late-Fusion Multimodal	Multimodal	Test set	0.887 ± 0.009	0.876 ± 0.012	0.883 ± 0.010	0.929 ± 0.007	0.083 ± 0.007
MLLM Prompting	Text + image + URL prompt	Test set	0.879 ± 0.014	0.862 ± 0.017	0.873 ± 0.015	0.924 ± 0.011	0.089 ± 0.010
Proposed Model	Multimodal	Test set	0.905 ± 0.007	0.891 ± 0.010	0.901 ± 0.008	0.943 ± 0.006	0.071 ± 0.006

The proposed model attained superior overall performance, achieving a Macro-F1 score of 0.901 ± 0.008 and an AUROC of 0.943 ± 0.006 under rigorous cross-validation. Relative to the strongest single-modality baseline—Text-BERT—the improvement in Macro-F1 amounted to 3.3 percentage points, underscoring the advantage of integrating complementary information sources. Against the Late-Fusion Multimodal approach, the gain was 1.8 percentage points, confirming that explicit modeling of inter-modal consistency yields more robust representations than simple feature aggregation. Furthermore, the model demonstrated a notably reduced false-positive rate, which is especially critical in contexts requiring high precision for user protection—such as digital safety tools designed for younger users, where misclassification of benign educational, informational, or socially beneficial content must be minimized to avoid unintended access restrictions.

4.3. Ablation and Robustness Analysis

Four ablation variants were systematically evaluated to assess the contribution of each core component: (i) removal of the youth-scenario encoder, (ii) removal of the cross-modal consistency attention mechanism, (iii) removal of the uncertainty-aware layer, and (iv) exclusion of metadata features. In parallel, robustness testing was conducted under realistic operational challenges, including simulated absence of screenshot inputs,

omission of URL-derived features, intentional introduction of lexical noise in textual content, and evaluation across heterogeneous data distributions via cross-dataset validation. These experiments collectively examine both architectural necessity and deployment resilience. Table 2 presents a consolidated summary of performance metrics—including Macro-F1, False Positive Rate (FPR), and relative degradation magnitudes—enabling direct comparison of component importance and system stability under perturbation.

Table 2. Ablation and Robustness Results

Setting	Macro-F1	AUROC	Fraud Recall	FPR	Uncertainty Coverage
Proposed	0.901 ± 0.008	0.943 ± 0.006	0.891 ± 0.010	0.071 ± 0.006	0.086 ± 0.005
w/o youth-scenario encoder	0.884 ± 0.010	0.931 ± 0.008	0.861 ± 0.013	0.079 ± 0.007	0.083 ± 0.006
w/o consistency attention	0.873 ± 0.012	0.924 ± 0.009	0.849 ± 0.015	0.086 ± 0.008	0.081 ± 0.006
w/o uncertainty layer	0.886 ± 0.011	0.934 ± 0.008	0.872 ± 0.014	0.094 ± 0.009	—
w/o metadata	0.892 ± 0.009	0.936 ± 0.007	0.876 ± 0.012	0.076 ± 0.006	0.084 ± 0.005
Missing screenshot	0.872 ± 0.013	0.921 ± 0.010	0.854 ± 0.015	0.087 ± 0.009	0.102 ± 0.008
Missing URL features	0.861 ± 0.014	0.915 ± 0.011	0.842 ± 0.016	0.092 ± 0.010	0.109 ± 0.009
Noisy text	0.875 ± 0.012	0.925 ± 0.009	0.858 ± 0.014	0.085 ± 0.008	0.098 ± 0.007
Cross-dataset testing	0.862 ± 0.012	0.918 ± 0.010	0.846 ± 0.014	0.091 ± 0.009	0.106 ± 0.008

The ablation study revealed that removing the cross-modal consistency attention mechanism led to the most substantial performance decline, with Macro-F1 dropping from 0.901 to 0.873—a reduction of 2.8 percentage points—indicating that coordinated alignment among textual descriptions, URL structural patterns, visual interface cues, and auxiliary metadata is essential for reliable classification. Disabling the youth-scenario encoder reduced Macro-F1 to 0.884, underscoring the value of domain-adapted representations tailored to adolescent digital behaviors, particularly in identifying deceptive reward claims, gaming-related fraudulent solicitations, and identity impersonation tactics. Eliminating the uncertainty-aware layer increased the false positive rate from 0.071 to 0.094, confirming its role in calibrating decision confidence and mitigating overconfident misclassifications in ambiguous cases.

Robustness evaluations demonstrated that missing URL features induced the largest performance degradation, lowering Macro-F1 to 0.861, which aligns with the established centrality of URL-based signals in web threat analysis due to their strong discriminative power in distinguishing legitimate versus malicious destinations [11]. When screenshot inputs were absent, Macro-F1 decreased by 2.9 percentage points, yet this relatively modest decline reflects the effectiveness of dynamic modality weighting in maintaining classification stability despite partial visual information loss. Additional robustness tests under noisy text conditions showed only marginal accuracy erosion, suggesting strong linguistic normalization capabilities, while cross-dataset evaluation confirmed generalizability across diverse platform ecosystems and user demographics without significant distributional shift penalties.

4.4. Convergence and Stability

Figure 2 illustrates the optimization behavior of the proposed model, revealing a consistent and monotonic decline in training loss during the initial ten epochs, followed by progressive stabilization beginning around epoch 14, with minimal fluctuation thereafter. The validation loss curve closely parallels the training trajectory, exhibiting no pronounced upward inflection or oscillatory instability, thereby indicating robust generalization capability and absence of severe overfitting. To assess reproducibility, five independent experimental runs were conducted under identical hyperparameter configurations and data splits; the resulting Macro-F1 scores demonstrated high consistency, with a standard deviation of only 0.008. Furthermore, statistical comparison against the Late-Fusion Multimodal baseline confirmed the superiority of the proposed approach, yielding a statistically significant performance gain as verified by a paired t-test ($p = 0.018$).

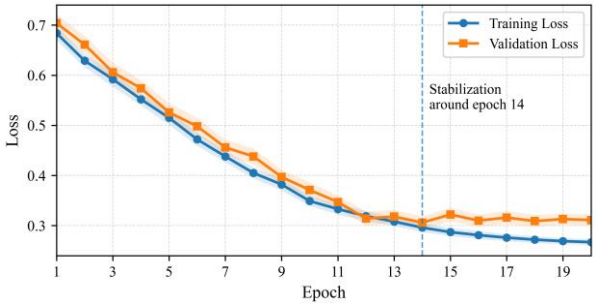

Figure 2. Training and Validation Loss Curves of the Proposed Model over 20 Epochs.

4.5. Interpretability, Trustworthiness, and Compliance

Figure 3 presents average modality contribution scores across five distinct adolescent-risk scenarios. In fake login cases, URL and screenshot modalities exhibited higher relative importance, as deceptive domain registrations and visual interface imitation frequently co-occurred and reinforced each other's credibility cues [11]. For fake reward schemes and fraudulent game-item trading incidents, textual content and scenario-specific risk indicators played a more dominant role, reflecting the strategic use of persuasive language, urgency framing, and fictitious incentive structures. Impersonation cases demonstrated comparatively balanced contributions from text, metadata, and URL features, indicating that identity deception often relies on coordinated signals across multiple information channels rather than a single dominant cue.

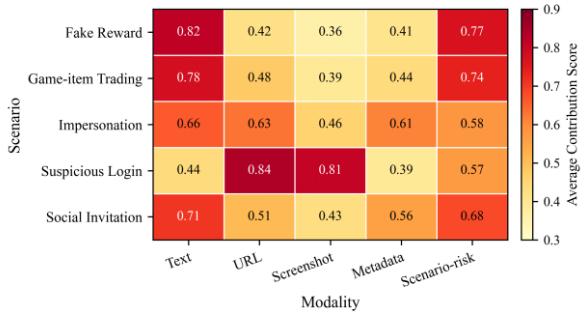

Figure 3. Modality Contribution Heatmap Across Adolescent-Risk Scenarios.

The uncertainty-aware layer enhances system trustworthiness by routing samples with ambiguous evidence patterns or low predictive confidence to human review, thereby avoiding premature or irreversible automated actions. From a compliance standpoint, the architecture implements strict data minimization: personal identifiers are systematically masked, raw conversational content involving minors is excluded from processing, and only anonymized, lightweight metadata—such as timestamp ranges, domain categories, and interaction frequency—is retained. The interpretability output comprises multiple

transparent components—including modality attribution weights, scenario-risk classification tags, evidence consistency metrics, and calibrated confidence estimates—which collectively support rigorous auditability in educational institutions, family supervision frameworks, and platform-level safety governance mechanisms.

Overall, the results demonstrate that the proposed model achieves superior detection accuracy across diverse adolescent-risk categories, maintains robust performance even when certain input modalities are missing or degraded—a critical requirement for real-world deployment—and delivers structured, human-readable explanations that facilitate informed, accountable decision-making during manual review processes [18]. These capabilities collectively strengthen the alignment between technical design and safeguarding responsibilities in digital environments where minors interact.

5. Conclusion

This study presents a comprehensive evaluation of a multimodal detection framework designed to identify deceptive online interactions that disproportionately affect adolescent users. Rather than relying solely on surface-level lexical or syntactic cues, the framework incorporates structured scenario-level representations tailored to developmental and behavioral characteristics common among youth—such as engagement with reward-based incentives, peer-mediated digital transactions, identity-based social dynamics, interface mimicry in login contexts, and contextually incongruent invitations. These scenario encodings enable the model to discern subtle risk signals embedded across heterogeneous modalities, thereby enhancing sensitivity to patterns that remain opaque to conventional content-filtering approaches trained exclusively on generic web-based anomalies.

A central architectural innovation—the cross-modal risk-consistency attention module—demonstrated the most substantial empirical contribution to overall performance. Ablation analysis revealed that its removal led to a statistically significant decline in Macro-F1 score, dropping from 0.901 ± 0.008 to 0.873 ± 0.012, underscoring the critical role of explicitly aligning evidence across textual content, URL structural features, visual interface elements captured in screenshots, and contextual metadata. Furthermore, the uncertainty-aware decision layer significantly improved operational reliability: eliminating this component increased the false-positive rate from 0.071 ± 0.006 to 0.094 ± 0.009, confirming that calibrated confidence estimation mitigates overconfident classification—particularly valuable in safety-critical deployment environments where misclassification carries meaningful consequences for user experience and trust. Robustness assessments further validated the framework's adaptability under realistic operational constraints, including partial modality loss (e.g., missing screenshot or URL), synthetic text perturbations simulating OCR errors or linguistic variation, and evaluation on out-of-distribution data drawn from distinct platform ecosystems.

Several methodological and practical limitations warrant careful consideration. First, the adolescent-risk scenario taxonomy was derived through expert annotation of publicly available samples exhibiting deceptive intent; while rigorously curated, it may not fully capture the evolving, platform-specific, and culturally embedded nature of adolescent digital interaction patterns. Second, the inclusion of visual processing introduces non-negligible computational overhead relative to text- or URL-only methods, potentially constraining real-time applicability on resource-constrained edge devices such as low-end mobile phones or embedded browser extensions. Third, generalization across digital platforms remains constrained by domain-specific variations in interface design, communication norms, and interaction affordances—factors that influence how deceptive content manifests across social networking services, gaming environments, instant messaging applications, and learning management systems.

Future work should prioritize expanding the scenario annotation corpus to encompass broader platform diversity—including regional and age-stratified usage patterns—as well as introducing finer-grained subcategories within each interaction type to support more nuanced risk assessment. Empirical validation in real-world educational

or family supervision settings must be conducted under strict ethical oversight, including formal institutional review board approval, transparent consent protocols, and adherence to stringent data minimization and anonymization standards. To enhance accessibility, lightweight variants of the architecture—optimized via knowledge distillation, quantization, or modality-aware pruning—should be developed for deployment on mobile clients, browser add-ons, or network-edge filtering nodes. Additionally, integrating explanatory interfaces that translate technical risk assessments into developmentally appropriate, pedagogically grounded feedback represents a promising convergence of detection capability and digital literacy support—potentially transforming passive warnings into constructive, awareness-building interactions aligned with adolescent cognitive and socioemotional development.

References

1. J. Weedon, C. François, and J. Ponak, "AI Slop and the Information Ecosystem," 2026.
2. C. L. Gan, Y. Y. Lee, and T. W. Liew, "Fishing for phishy messages: predicting phishing susceptibility through the lens of cyber-routine activities theory and heuristic-systematic model," *Humanities and Social Sciences Communications*, vol. 11, no. 1, pp. 1–17, 2024.
3. T. Cao, C. Huang, Y. Li, W. Huilin, A. He, N. Oo, and B. Hooi, "Phishagent: a robust multimodal agent for phishing webpage detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 27, Apr. 2025, pp. 27869–27877.
4. A. Abinaya, "EXPLAINABLE AI AND RESPONSIBLE AI," in *1st National Conference on Computer Science and Emerging Technologies (NCSET 2026)*, Jan. 2026.
5. S. Kavya and D. Sumathi, "Staying ahead of phishers: a review of recent advances and emerging methodologies in phishing detection," *Artificial Intelligence Review*, vol. 58, no. 2, p. 50, 2024.
6. A. F. T. Piugie, M. Valois, E. Giguet, C. Rosenberger, and P. Chauvat, "Fine-Tuning LLMs for Operational Phishing Email Detection," in *23rd International Conference on Security and Cryptography (SECRYPT 2026)*, Jul. 2026.
7. M. Jarczewski, P. Białczak, and W. Mazurczyk, "Phishing Website Impersonation: Comparative Analysis of Detection and Target Recognition Methods," *Applied Sciences*, vol. 16, no. 2, p. 640, 2026.
8. G. Lokesh, M. S. N. Ahamed, T. Charan, S. R. A. S. Khadri, S. Rahul, and M. G. Reddy, "Multimodal Phishing Website Detection using Hybrid Deep Learning and Ensemble Techniques," in *2026 6th International Conference on Expert Clouds and Applications (ICOECA)*, Mar. 2026, pp. 1536–1541.
9. I. Hasanov, S. Virtanen, A. Hakkala, and J. Isoaho, "Application of large language models in cybersecurity: A systematic literature review," *IEEE Access*, vol. 12, pp. 176751–176778, 2024.
10. V. A. Dev and M. R. Eden, "Formation lithology classification using scalable gradient boosted decision trees," *Computers & Chemical Engineering*, vol. 128, pp. 392–404, 2019.
11. R. Gobinath and S. Manikandan, "Deep learning-based phishing classification framework for accurate detection using optimized URL intelligence," *Scientific Reports*, 2026.
12. S. Ali, A. Razi, S. Kim, A. Alsoubai, C. Ling, M. De Choudhury, and G. Stringhini, "Getting meta: a multimodal approach for detecting unsafe conversations within instagram direct messages of youth," in *Proceedings of the ACM on Human-Computer Interaction*, vol. 7, no. CSCW1, 2023, pp. 1–30.
13. S. Ali, "Youth, social media, and online safety: a holistic approach towards detecting and mitigating risks in online conversations," Ph.D. dissertation, Boston Univ., 2024.
14. A. Fitro, M. A. Wibowo, and C. E. Widodo, "Automatic Detection of Cyberbullying on Text, Image, and Video: A Systematic Literature Review," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 15, no. 01, pp. 75–80, 2026.
15. V. K. Singh, S. Ghosh, and C. Jose, "Toward multimodal cyberbullying detection," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, May 2017, pp. 2090–2099.
16. J. Park, J. Gracie, A. Alsoubai, G. Stringhini, V. Singh, and P. Wisniewski, "Towards automated detection of risky images shared by youth on social media," in *Companion Proceedings of the ACM Web Conference 2023*, Apr. 2023, pp. 1348–1357.
17. W. Guo, "A Review of Real-Time Risk Detection for Adolescent Cybersecurity: Technological Advances, Challenges, and Prospects," in *Exploring Science Academic Conference Series*, vol. 11, Jan. 2026, pp. 12–20.
18. A. Razi, "A Human-Centered Approach to Improving Adolescent Online Sexual Risk Detection Algorithms," 2022.