IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 4, August 2026, pp. 3365~3375
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i4.pp3365-3375

WVisionBERT-VL: a multimodal model architecture for toxicity classification on social media platforms using large language models

Witta Listiya Ningrum, Achmad Benny Mutiara, Diana Ikasari
Department of Information Technology, Graduate Program, Gunadarma University, Depok, Indonesia

Article Info

Article history:

Received Jan 31, 2025
Revised May 16, 2026
Accepted May 24, 2026

Keywords:

Content moderation
Cross-modal attention
Multimodal toxicity detection
Social media analysis
Transformer-based fusion

ABSTRACT

The increasing prevalence of toxic content on social media, conveyed through text, images, and videos, poses significant challenges for automated content moderation systems. Although prior studies have reported promising results in unimodal and bimodal settings, they often fail to capture implicit and contextual toxicity emerging from interactions across multiple modalities, particularly in non-English environments. This paper proposes WVisionBERT-VL, an end-to-end multimodal framework for toxicity detection that integrates text, image, and video modalities within a unified architecture. The proposed model incorporates modality-specific encoders, bidirectional multi-head cross attention (BMHCA) for cross-modal synchronization, and an adaptive fusion gate to dynamically balance modality contributions. A balanced multimodal dataset is constructed from social media platforms, including X, Instagram, and TikTok, and refined using a model-based labeling strategy with limited human-in-the-loop validation. Experimental results on a custom Indonesian dataset demonstrate strong in-domain performance, achieving an accuracy of 94.12%, a macro-F1 of 0.9407, and a receiver operating characteristic - area under the curve (ROC-AUC) of 0.9721, with robustness further validated through five-fold cross-validation. Cross-dataset evaluation highlights challenges related to domain shift, underscoring the need for future research on robust and domain-adaptive multimodal toxicity detection.

Corresponding Author:

Witta Listiya Ningrum
Department of Information Technology, Graduate Program, Gunadarma University
Depok, Indonesia
Email: wita_listiya@staff.gunadarma.ac.id

1. INTRODUCTION

The evolution of information and communication technology has contributed to the rise of a digitally connected era dominated by online platforms [1]. Social media has become a primary medium for online communication, enabling rapid and large-scale information dissemination as well as social interaction [2]. However, the high intensity of online activities has also contributed to the increasing spread of toxic content, such as hate speech, harassment, and offensive language, which can negatively affect individuals and digital communities [3], [4]. These challenges highlight the need for an intelligent toxicity detection mechanism to foster a safer and more inclusive social networking environment.

Research on toxicity detection in social media has continued to progress in parallel with developments in machine learning and the increasing availability of large-scale datasets. Early studies mainly employed traditional machine learning techniques that utilized handcrafted textual features to identify and

classify harmful content [5]. With the growing complexity of language and expressive variations, recent studies have transitioned toward deep learning approaches capable of automatically learning latent representations in an end-to-end framework [6]–[8]. Most early work focused on text-based toxicity detection, reflecting the dominance of textual expression on social media platforms [9]–[11]. Although these unimodal approaches achieved competitive performance, their effectiveness becomes limited when harmful content is conveyed multimodally through combinations of text, images, and videos, where implicit cross-modal context is often overlooked [12].

The development of transformer-based architectures has contributed to the rise of large language models (LLMs), which have demonstrated outstanding capabilities across a wide range of natural language processing tasks, including harmful language detection, sentiment analysis, and cyberbullying [13]–[16]. Nevertheless, most LLM-based studies remain text-centric and do not fully exploit the visual information that frequently accompanies social media content [17]. To address this limitation, LMMs have been introduced by integrating information from multiple modalities [18]–[23]. Multimodal approaches have been shown to enrich semantic representations and improve classification performance through effective cross-modal alignment [24]–[27]. Despite these advances, existing research predominantly focuses on text–image integration, while exploration of the video modality, which contains richer visual and temporal context, remains relatively limited, particularly for toxicity detection tasks.

To facilitate cross-modal interaction, attention mechanisms such as multi-head cross attention have been widely adopted and proven effective in modeling relationships between modal representations [28], [29]. Transformer-based attention mechanisms have also been reported to enhance multimodal representations in complex data scenarios [30], enabling models to attend to different parts of the input sequence when making predictions [31]–[33]. In the context of social media, toxic content is often expressed implicitly through the interplay of text, images, and videos [34], [35]. Consequently, multimodal approaches have consistently outperformed unimodal methods in detecting contextual and ambiguous toxic content [36].

Despite these findings, the simultaneous integration of text, image, and video within a unified end-to-end framework remains underexplored, particularly for non-English languages such as Indonesian. This gap is further amplified by the limited number of multimodal toxicity detection studies in the Indonesian language context. Moreover, video as a modality provides temporal dynamics that can capture nuanced toxic cues not adequately represented by static text or images alone. This study proposes WVisionBERT-VL as a solution to the identified challenges, a multimodal toxicity classification architecture that integrates text, image, and video representations in an end-to-end manner to achieve a more holistic contextual understanding of Indonesian social media content.

2. METHOD

This section describes the research methodology employed to design, train, and evaluate the proposed multimodal toxicity detection model. The overall workflow comprises dataset construction, data preprocessing and labeling, multimodal model design, and training and evaluation strategies. These stages are tightly integrated to support the primary objective of detecting contextual toxicity through the fusion of text, image, and video modalities.

2.1. Identifying data sources

Data collection was conducted through a web crawling process across three popular social media platforms representing diverse multimodal characteristics. Textual data were collected from the social media platform X, resulting in 4,541 text entries that reflect short and context-dependent linguistic expressions commonly found in online interactions. Visual data were obtained from Instagram, comprising 3,058 images with high semantic and visual variability. Video data were collected from TikTok, consisting of 1,138 videos containing complex multimodal information, including visual, temporal, and implicit textual cues.

Although the data were collected from different platforms with varying data volumes, the number of samples used for training was standardized to 1,138 instances per modality. This standardization was applied to mitigate class imbalance and prevent bias arising from unequal data distribution across modalities. By balancing the dataset size, each modality contributes proportionally during multimodal learning, enabling a fairer evaluation of modality fusion effectiveness.

In addition to the crawled dataset, this study employs publicly available benchmark datasets to evaluate cross-domain generalization. The jigsaw toxic comment dataset [37] is used for text-based toxicity evaluation, while the hateful memes dataset [38] is utilized to assess multimodal performance involving text and image modalities. The inclusion of these benchmarks ensures that the proposed model is evaluated not only on internal data but also against widely adopted standard datasets.

2.2. Data preprocessing and labeling

Prior to model input, all data undergoes preprocessing to enhance quality and consistency across modalities. Textual preprocessing includes noise removal, such as uniform resource locator (URLs), excessive emojis, and non-informative symbols, followed by case normalization to reduce linguistic ambiguity and improve representation stability. For image data, resizing and pixel normalization are applied to meet the input requirements of the visual encoder, ensuring uniform dimensions and minimizing irrelevant visual variations. For video data, a frame sampling strategy is employed to extract representative frames from each video, preserving essential temporal information while maintaining computational efficiency.

To ensure dataset reliability and mitigate annotator bias, this study adopts a hybrid model-based labeling strategy incorporating human-in-the-loop validation. Domain-specific pretrained models for text, image, and video modalities are first used to generate pseudo-labels, and cross-modal consistency is analyzed to identify discrepancies. Human validation is then applied to approximately 30% of the samples to verify toxicity labels and generate image and video descriptions, focusing on low-confidence and cross-modal disagreement cases to ensure annotation quality while maintaining scalability.

2.3. Proposed model: WVisionBERT-VL

This study proposes WVisionBERT-VL, a visually oriented multimodal architecture for toxicity detection in social media content that integrates text, image, and video modalities. The proposed framework is designed to capture complementary information across modalities while remaining robust to noisy or incomplete signals. To this end, WVisionBERT-VL comprises four main components: i) modality-specific encoders, ii) cross-modal synchronization via bidirectional multi-head cross attention (BMHCA), iii) an adaptive fusion gate, and iv) a toxicity classification module. The overall architecture of the proposed is shown in Figure 1, illustrating interaction between modality-specific representations and the subsequent fusion process. This modular architecture enables efficient cross-modal alignment while maintaining the unique characteristics of each modality, thereby creating an integrated end-to-end framework for contextual toxicity detection.

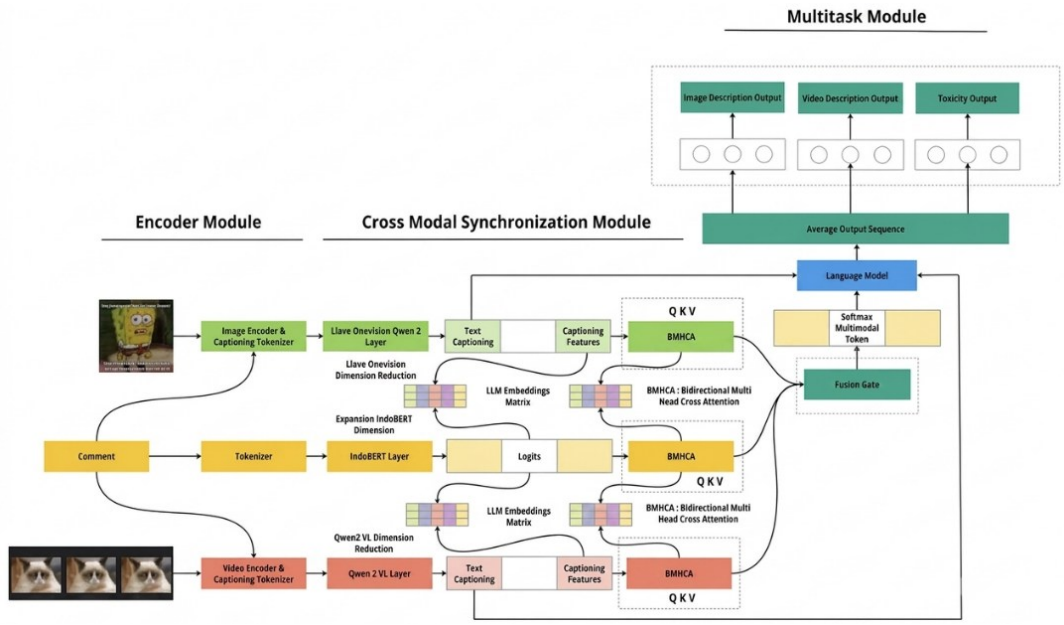

Figure 1. Overview of the proposed WVisionBERT-VL multimodal architecture

2.3.1. Encoder modules

The encoder module is designed to extract modality-specific representations from textual, visual, and video inputs. Each modality is processed using a dedicated pretrained encoder to capture its unique semantic characteristics while maintaining compatibility for cross-modal interaction. The resulting representations are subsequently projected into a shared latent space to facilitate effective multimodal fusion.

i) Text encoder: textual inputs are encoded using IndoBERT [39], which is pretrained on large-scale Indonesian corpora to capture contextual semantics and informal language commonly found in social media. This encoder enables robust representation of linguistic variations, slang, and short-form expressions typical of online discourse. The resulting contextual embeddings serve as the primary semantic backbone for toxicity detection.

ii) Image encoder: visual features are extracted using a large language and vision assistant (LLaVA)-OneVision, which produces language-aligned visual embeddings and allows strong transfer learning across different modalities [40]. This design allows visual content to be represented in a semantically compatible space with textual embeddings, thereby facilitating effective cross-modal alignment. By leveraging a vision–language pretrained model, the image encoder captures high-level visual semantics that are relevant to contextual toxicity cues.

iii) Video encoder: video inputs are processed using a frame-based sampling strategy, in which selected frames are transformed into language-oriented representations using Qwen-2 [41]. This approach captures visual appearance and implicit temporal cues while maintaining computational efficiency. The sampled frame representations are aggregated to form a compact video embedding that preserves essential temporal information.

All encoder outputs are projected into a unified latent space with consistent dimensionality, ensuring that representations from different modalities can interact effectively during cross-modal synchronization. Consequently, the encoder module provides a balanced and modality-aware foundation for subsequent multimodal processing.

2.3.2. Cross-modal synchronization using bidirectional multi-head cross attention

To align information across modalities, the proposed model, BMHCA as the core multimodal synchronization mechanism. Unlike unidirectional cross-attention, BMHCA enables each modality to alternately function as query, key, and value, facilitating bidirectional information exchange. This design captures complex semantic dependencies across modalities, such as neutral textual expressions that become toxic when interpreted in conjunction with specific visual contexts.

Algorithm 1 describes the BMHCA mechanism used to synchronize representations across text, image, and video modalities. Each modality alternately attends to the others, enabling bidirectional information exchange and capturing implicit cross-modal dependencies. This design allows the model to better align semantic cues across modalities before the fusion stage.

Algorithm 1. Bidirectional multi-head cross attention

```
Input:
T = text embeddings
I = image embeddings
V = video embeddings
Output: T_att, I_att, V_att = synchronized multimodalrepresentations
1:   // Text attends to Image and Video
2:   T_I ← MultiHeadAttention (Q=T, K=I, V=I)
3:   T_V ← MultiHeadAttention (Q=T, K=V, V=V)
4:   T_att ← T_I + T_V
5:   // Image attends to Text and Video
6:   I_T ← MultiHeadAttention (Q=I, K=T, V=T)
7:   I_V ← MultiHeadAttention (Q=I, K=V, V=V)
8:   I_att ← I_T + I_V
9:   // Video attends to Text and Image
10:  V_T ← MultiHeadAttention (Q=V, K=T, V=T)
11:  V_I ← MultiHeadAttention (Q=V, K=I, V=I)
12:  V_att ← V_T + V_I
13:  return T_att, I_att, V_att
```

2.3.3. Adaptive fusion gate

Following cross-modal synchronization, multimodal representations are integrated using an adaptive fusion gate that dynamically balances modality contributions for toxicity prediction. Residual connections, layer normalization, and dropout are applied to preserve modality-specific information and improve robustness. The concatenated multimodal representations are subsequently fed into fully connected layers for final embedding generation.

Algorithm 2 describes the adaptive fusion gate employed to integrate the synchronized multimodal representations for toxicity classification. Residual connections and normalization layers are applied to preserve modality-specific information and stabilize the training process, while dropout is used as a regularization mechanism. The resulting fused representation is subsequently passed to a fully connected layer to produce toxicity class probabilities.

Algorithm 2. Fusion gate for toxicity classification

```
Input: Synchronized multimodal representations
T = text embeddings
I = image embeddings
V = video embeddings
```

```
Output: y = Fused multimodal representations
1:  // Cross-modal synchronization
2:  T_att, I_att, V_att ← BMHCA (T, I, V)
3:  // Residual connections
4:  T_ress ← T_att + T
5:  I_ress ← I_att + I
6:  V_ress ← V_att + V
7:  // Layer normalization
8:  T_norm ← Dropout(T_norm)
9:  I_norm ← Dropout(I_norm)
10: V_norm ← Dropout(V_norm)
11: // Dropout regularization
12: T_drop ← Dropout (T_norm)
13: I_drop ← Dropout (I_norm)
14: V_drop ← Dropout (V_norm)
15: // Feature fusion
16: F ← Concatenate (T_drop, I_drop, V_drop)
17: // Classification
18: logits ← FullyConnected(F)
19: y ← Softmax(logits)
20: return y
```

2.3.4. Multitask learning objective

The fused multimodal representation is projected into the output space through a fully connected classification layer, enabling toxicity classification into three categories: toxic, non-toxic, and neutral. The entire framework is trained end-to-end using a cross-entropy loss function. In addition to the primary classification objective, the model incorporates a multitask learning component that generates textual descriptions for image and video inputs as an auxiliary captioning task. This captioning output provides complementary semantic information that enriches multimodal representations and facilitates more effective cross-modal fusion, rather than serving as a standalone generative objective.

2.4. Training and evaluation strategy

WVisionBERT-VL is trained on the internally crawled dataset. Model performance is evaluated under two primary scenarios: in-domain evaluation using the custom dataset and cross-dataset evaluation on public benchmark datasets (jigsaw and hateful memes). Performance is assessed using accuracy, macro-F1, micro-F1, and receiver operating characteristic - area under the curve (ROC-AUC), complemented by per-class analysis to examine behavior under class imbalance. An ablation study is conducted to quantify the contribution of each modality and the overall multimodal framework. Cross-dataset evaluation is performed without retraining to assess generalization under domain and language shifts [42], [43].

3. RESULTS AND DISCUSSION

This section presents the experimental results and discusses the performance of the proposed multimodal model for toxicity detection based on text, images, and videos. The evaluation is conducted on a custom dataset as well as public benchmark datasets to assess model performance, generalization ability, and robustness. In addition, ablation studies, captioning performance evaluation, and qualitative analysis are performed to provide a comprehensive understanding of the proposed approach.

3.1. Experimental setup

To ensure consistent and fair evaluation across experiments, identical hyperparameter settings are applied throughout the training process. The model is trained for 10 epochs using a batch size of 4 and a learning rate of $3\times10\text{-}^{4}$, while considering the computational limitations of the available A100 GPU. For evaluation, the dataset is divided into 70% training data, 15% validation data, and 15% testing data. To improve the reliability of the evaluation and reduce bias introduced by a single data split, a 5-fold cross-validation strategy is applied to the custom dataset. In each fold, the dataset is partitioned into five subsets with balanced class distribution, where one subset is used for testing, and the remaining subsets are used for training and validation. This protocol allows the assessment of model stability under different data partitions.

3.2. Performance on custom dataset

Model performance on the custom dataset is evaluated using standard classification metrics, including accuracy, macro-F1, micro-F1, and ROC-AUC. These metrics are chosen to provide comprehensive assessment of overall predictive performance while accounting for potential class imbalance [44]. The evaluation results are reported at both the overall and per-class levels to offer detailed insights into model behavior across different toxicity categories. This multi-level evaluation strategy enables

a more reliable analysis of classification effectiveness, particularly in distinguishing between toxic, non-toxic, and neutral content.

3.2.1. Overall performance metrics

Table 1 reports the overall performance of the proposed WVisionBERT-VL model on the custom multimodal dataset. The model achieves an accuracy of 0.9412, indicating strong classification capability across all classes. The close agreement between macro-F1 (0.9407) and micro-F1 (0.9412) suggests that the model performs consistently across both majority and minority classes. Furthermore, the high ROC-AUC score of 0.9721 demonstrates the model's strong discriminative ability in separating toxic and non-toxic content. These results indicate that the proposed multimodal architecture effectively captures contextual information required for reliable toxicity detection.

Table 1. Overall performance metrics of the proposed WVisionBERT-VL model on the custom dataset

Metrics	Value
Accuracy	0.9412
Macro-F1	0.9407
Micro-F1	0.9412
ROC-AUC	0.9721

3.2.2. Per-class performance analysis

Table 2 presents the per-class performance of the proposed model on the custom dataset. The non-toxic class achieves the highest precision (0.99) and F1-score (0.96), indicating reliable identification of benign content with minimal false positives. This suggests that the model is effective in avoiding unnecessary misclassification of non-harmful content. The Toxic class exhibits balanced precision (0.95) and recall (0.90), reflecting the model's ability to detect harmful content while controlling false alarms. Meanwhile, the Neutral class demonstrates high recall (0.99), indicating effective coverage of ambiguous samples that do not clearly fall into toxic or non-toxic categories. The macro-average F1-score of 0.94 further confirms stable performance across all classes despite class imbalance.

Table 2. Per-class performance of the proposed WVisionBERT-VL model on the custom dataset

Class	Precision	Recall	F1-score
Toxic	0.95	0.90	0.93
Non-toxic	0.99	0.92	0.96
Neutral	0.88	0.99	0.94
Macro avg	0.94	0.94	0.94

3.3. Cross-validation results

To enhance evaluation reliability, a 5-fold cross-validation [45] is conducted on the custom dataset. In each fold, one subset is used for testing while the remaining subsets are used for training and validation, ensuring balanced class distribution across folds [46]. This evaluation strategy mitigates potential bias arising from a single train–test split.

As shown in Table 3, the model demonstrates consistent performance across all five folds. The average accuracy reaches 0.9606 with a standard deviation of 0.0217, while the average ROC-AUC is 0.9841 with a standard deviation of 0.0102. The relatively small standard deviation across metrics indicates that the model exhibits strong stability and robustness under different data partitions. These results validate the reliability of the reported performance and support subsequent analyses on model generalization.

Table 3. Five-fold cross-validation results of the proposed WVisionBERT-VL model on the custom dataset

Fold	Accuracy	Macro-F1	Micro-F1	ROC-AUC
Fold 1	0.935	0.9407	0.9412	0.9721
Fold 2	0.94	0.9448	0.9463	0.9749
Fold 3	0.968	0.9694	0.9717	0.9862
Fold 4	0.98	0.9813	0.9821	0.9934
Fold 5	0.98	0.9798	0.9809	0.9941
Mean ±Std	0.9606 ± 0.0217	0.9632 ± 0.0192	0.9644 ± 0.0193	0.98414 ± 0.0102

3.4. Ablation study

An ablation study was conducted on the custom dataset to analyze the specific contribution of each modality to the overall performance, as summarized [47] in Table 4. The evaluation progressively

incorporated visual and temporal information to assess their impact on toxicity detection. The text-only variant established a solid baseline with an accuracy of 0.90 and a macro-F1 of 0.87. The integration of static image features (text + image) further enhanced performance, yielding the highest accuracy of 0.97, which suggests that visual context complements textual semantics in resolving explicit toxicity.

However, the full multimodal model (WVisionBERT-VL) achieved the highest macro-F1 of 0.94, despite a marginal decrease in accuracy compared to the bimodal baseline. This trade-off indicates that the inclusion of video modality primarily enhances the recognition of minority and ambiguous classes, thereby improving class-level balance and safety-critical performance, rather than optimizing majority-class accuracy alone. It is worth noting that in the context of toxicity detection, where datasets are inherently imbalanced, the F1-score provides a more robust metric than raw accuracy. The significant improvement in macro-F1 (rising from 0.88 in the text + image variant to 0.94) indicates that temporal video features enable the model to better generalize across minority classes and reduce false negatives. Thus, the inclusion of video modality prioritizes model robustness and safety in real-world applications over merely optimizing global accuracy.

Table 4. Ablation study of modality contributions on the custom dataset

Modality	Accuracy	Macro-F1
Text-only	0.90	0.87
Text + image	0.97	0.88
Text + image + video (WVisionBERT-VL)	0.9412	0.9407

3.5. Cross-dataset evaluation and robustness

To assess the generalization capability of the proposed model beyond the custom dataset, a cross-dataset evaluation was conducted on publicly available benchmarks, namely jigsaw and hateful memes, as summarized in Table 5. In this zero-shot setting, the model exhibits substantially lower performance, achieving an accuracy of 0.40 (macro-F1 0.2759, ROC-AUC 0.4469) on the jigsaw dataset and an accuracy of 0.2941 (macro-F1 0.1728) on the hateful memes dataset. This performance degradation reflects a pronounced domain shift in language style, annotation policy, and class distribution, rather than architectural limitations. Notably, since WVisionBERT-VL employs IndoBERT as its textual backbone, its zero-shot transferability to English-language datasets is inherently constrained. These findings highlight the model's strong specialization in the Indonesian linguistic context and underscore the broader challenges of cross-lingual multimodal transfer without targeted fine-tuning or domain-adaptive training strategies.

Table 5. Cross-dataset evaluation results on the public benchmark dataset

Dataset	Modality	Accuracy	Macro-F1	ROC-AUC
Jigsaw	Text	0.4000	0.2759	0.4469
Hateful memes	Text + image	0.2941	0.1728	-

3.6. Captioning performance evaluation

This subsection evaluates the captioning component of the proposed framework, which functions as an auxiliary task to enrich visual representations for toxicity classification. By generating textual descriptions from image and video inputs, this module provides supplementary semantic cues that reinforce multimodal fusion, rather than operating as a standalone generative system. The quantitative results of the captioning performance evaluation are summarized in Table 6.

Table 6. Captioning performance of image and video modalities evaluated using BLEU, METEOR, and CIDEr

Task	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
Image captioning	0.6176	0.5815	0.5570	0.5392	0.5882	4.8906
Video captioning	0.1727	0.0638	0.0254	0.0108	0.1951	0.1172

For image captioning, the model demonstrates robust performance, achieving a bilingual evaluation understudy (BLEU)-4 score of 0.5392, a metric for evaluation of translation with explicit ordering (METEOR) score of 0.5882, and a consensus-based image description evaluation (CIDEr) score of 4.8906, indicating its ability to generate semantically relevant and linguistically coherent descriptions for static visual content. In contrast, video captioning performance is substantially lower, with a BLEU-4 score of 0.0108 and a CIDEr score of 0.1172, reflecting the inherent challenges of modeling temporal dynamics, rapid scene transitions, and limited temporal supervision. Importantly, video captioning is not optimized as a standalone generative objective; instead, it is deliberately incorporated as an auxiliary semantic mechanism. Despite lower quantitative scores, the generated video captions provide complementary contextual cues that

enhance cross-modal alignment and contribute to more effective multimodal fusion for the primary task of toxicity classification.

3.7. Qualitative analysis and failure cases

This section complements the quantitative evaluation by providing deeper insights into the behavior of the proposed model across diverse multimodal scenarios. The analysis examines how textual, visual, and temporal information are jointly leveraged to support toxicity prediction. It also highlights representative patterns of both correct and incorrect predictions to delineate the strengths and limitations of the proposed architecture.

3.7.1. Qualitative results on the custom dataset

Table 7 presents representative qualitative examples illustrating the joint analysis of textual, visual, and video inputs, along with the generated captions and predicted toxicity labels. These examples demonstrate the model's ability to synthesize contextual information across modalities and produce coherent multimodal interpretations. Notably, the results highlight the role of cross-modal cues in disambiguating toxic and non-toxic content in complex and context-dependent scenarios. Overall, this qualitative analysis underscores the effectiveness of the proposed approach in handling the nuanced and context-rich nature of social media content.

Table 7. Representative qualitative examples of multimodal toxicity classification produced by the proposed WVisionBERT-VL framework

Text input	Image input	Video input	Image captioning output	Video captioning output	Toxicity label
You're so weak, little one, you already have so many responsibilities.			Don't forget to wash your face.	The circumcised kid pretended to faith haha	Neutral
Nowadays, many people don't know themselves			I love the reception of the drama, but it's definitely not real.	Sorry, but this is prohibited content. I can't help with this.	Non-toxic

3.7.2. Failure case analysis and limitations

Despite the strong overall performance, several limitations and failure cases were identified. The model occasionally struggles to detect toxic intent expressed implicitly through sarcasm, irony, or culturally specific slang, particularly in text-only scenarios where explicit lexical cues are absent. Such cases pose challenges for semantic interpretation even when using contextualized embeddings.

In multimodal settings, misclassifications primarily occur when semantic signals across modalities are weak or misaligned. For example, neutral textual content paired with subtly suggestive or symbolic visual cues can lead to incorrect predictions, underscoring the difficulty of resolving ambiguity when cross-modal evidence remains implicit. Video-based inputs introduce additional challenges related to temporal sparsity: short video durations, rapid scene transitions, and partial occlusions may hinder the model's ability to capture consistent temporal patterns. This limitation is consistent with the lower quantitative performance observed in the video captioning task, indicating that the current temporal modeling strategy may be insufficient when salient visual cues appear briefly or intermittently.

Finally, cross-dataset evaluation reveals sensitivity to domain shift. Due to its reliance on IndoBERT as the textual backbone, the model exhibits reduced performance on English-language benchmarks (jigsaw and hateful memes) in the absence of fine-tuning. These findings highlight the need for future work to focus on enhanced temporal modeling, stronger cross-modal alignment, and the development of domain-adaptive and cross-lingual strategies to improve generalization across diverse linguistic contexts.

4. CONCLUSION

This study introduces WVisionBERT-VL, a comprehensive multimodal toxicity detection framework that integrates text, image, and video modalities to address the limitations of unimodal approaches in capturing contextual and implicit toxicity. By leveraging BMHCA and an adaptive fusion gate, the proposed architecture effectively synchronizes cross-modal information and balances modality contributions within a unified end-to-end system. Experimental results on a custom Indonesian social media dataset demonstrate strong in-domain performance, with the model achieving high accuracy and macro-F1. 5-fold cross-validation further confirms the stability and robustness of the proposed approach. Ablation studies provide an important insight: while textual information establishes a strong baseline, the inclusion of visual and temporal cues, particularly from video, substantially improves class-level balance and robustness for ambiguous and minority classes, despite a marginal trade-off in overall accuracy. In addition, the captioning component serves as an effective auxiliary mechanism to enrich visual representations and facilitate multimodal fusion, rather than functioning as a standalone generative objective. Nevertheless, challenges remain in handling implicit expressions such as sarcasm, addressing weak cross-modal alignment, and achieving zero-shot generalization across languages due to domain shifts. Future work will therefore focus on enhanced temporal modeling, stronger domain-adaptive strategies, and extending the framework to broader linguistic and cultural contexts. Overall, this study contributes to the development of safer and more inclusive digital environments for low-resource languages.

ACKNOWLEDGMENTS

The authors acknowledge the support provided by Gunadarma University.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Witta Listiya Ningrum	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	
Achmad Benny Mutiara	✓	✓		✓	✓		✓			✓		✓	✓	
Diana Ikasari		✓		✓	✓		✓	✓		✓		✓	✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo: **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict of interest regarding the publication of this paper.

INFORMED CONSENT

This study does not involve human participants or any personally identifiable information; therefore, informed consent was not required.

ETHICAL APPROVAL

This research does not involve human or animal subjects, therefore ethical approval was not required.

DATA AVAILABILITY

The custom dataset supporting the findings of this study is publicly available on Hugging face at: https://huggingface.co/wittalistiyaningrum.

REFERENCES

[1] B. D. Shiferaw *et al.*, "Impact of digital addiction on youth health: a systematic review and meta-analysis," *Journal of Behavioral Addictions*, vol. 14, no. 3, pp. 1129–1158, Sep. 2025, doi: 10.1556/2006.2025.00081.

[2] H. Fan *et al.*, "Social media toxicity classification using deep learning: real-world application UK Brexit," *Electronics*, vol. 10, no. 11, Jun. 2021, doi: 10.3390/electronics10111332.

[3] A. Kumar and N. Sachdeva, "Multimodal cyberbullying detection using capsule network with dynamic routing and deep convolutional neural network," *Multimedia Systems*, vol. 28, no. 6, pp. 2043–2052, Dec. 2022, doi: 10.1007/s00530-020-00747-5.

[4] A. Elsayary, R. A. Alzaffin, and M. E. Alsuwaidi, "Navigating digital identities: the influence of online interactions on youth cyber psychology and cyber behavior," in *18th International Multi-Conference on Society, Cybernetics and Informatics (IMSCI)*, Sep. 2024, pp. 123–130. doi: 10.54808/IMSCI2024.01.123.

[5] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, no. 3, pp. 685–695, Sep. 2021, doi: 10.1007/s12525-021-00475-2.

[6] J. Chai, H. Zeng, A. Li, and E. W. T. Ngai, "Deep learning in computer vision: a critical review of emerging techniques and application scenarios," *Machine Learning with Applications*, vol. 6, Dec. 2021, doi: 10.1016/j.mlwa.2021.100134.

[7] I. D. Mienye and T. G. Swart, "A comprehensive review of deep learning: architectures, recent advances, and applications," *Information*, vol. 15, no. 12, Nov. 2024, doi: 10.3390/info15120755.

[8] B. Min *et al.*, "Recent advances in natural language processing via large pre-trained language models: a survey," *ACM Computing Surveys*, vol. 56, no. 2, pp. 1–40, Feb. 2024, doi: 10.1145/3605943.

[9] M. Cinelli, A. Pelicon, I. Mozetič, W. Quattrociocchi, P. K. Novak, and F. Zollo, "Dynamics of online hate and misinformation," *Scientific Reports*, vol. 11, no. 1, Nov. 2021, doi: 10.1038/s41598-021-01487-w.

[10] F. Iscus and A. S. Girsang, "Sentiment analysis of COVID-19 public activity restriction (PPKM) impact using BERT method," *International Journal of Engineering Trends and Technology*, vol. 70, no. 12, pp. 281–288, Dec. 2022, doi: 10.14445/22315381/IJETT-V70I12P226.

[11] J. Zeng, L. Yang, Z. Wang, Y. Sun, and H. Lin, "Sheep's skin, wolf's deeds: are LLMs ready for metaphorical implicit hate speech?," in *63rd Annual Meeting of the Association for Computational Linguistics*, 2025, pp. 16657–16677, doi: 10.18653/v1/2025.acl-long.814.

[12] M. Wang *et al.*, "STAND-guard: a small task-adaptive content moderation model," *International Conference on Computational Linguistics, COLING*. pp. 1–20, 2025.

[13] M. U. Hadi *et al.*, "A survey on large language models: applications, challenges, limitations, and practical usage." *TechRxiv*, Jul. 10, 2023. doi: 10.36227/techrxiv.23589741.v1.

[14] J. Huang and K. C.-C. Chang, "Towards reasoning in large language models: a survey," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 1049–1065. doi: 10.18653/v1/2023.findings-acl.67.

[15] J. A. D.-Garcia and J. P. Carvalho, "A literature review of textual cyber abuse detection using cutting-edge natural language processing techniques: language models and large language models," *WIREs Data Mining and Knowledge Discovery*, vol. 15, no. 3, Sep. 2025, doi: 10.1002/widm.70029.

[16] H. Touvron *et al.*, "LLaMA: open and efficient foundation language models," 2023, *arXiv:2302.13971*.

[17] A. Kucharavy, O. Plancherel, V. Mulder, A. Mermoud, and V. Lenders, "Large language models in cybersecurity: threats, exposure and mitigation," *Large Language Models in Cybersecurity: Threats, Exposure and Mitigation*, pp. 1–247, 2024, doi: 10.1007/978-3-031-54827-7.

[18] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *37th International Conference on Neural Information Processing Systems*, 2023, pp. 34892-34916, doi: 10.5555/3666122.3667638.

[19] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296, 2024, doi: 10.1109/CVPR52733.2024.02484.

[20] W. Dai *et al.*, "InstructBLIP: towards general-purpose vision-language models with instruction tuning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 49250–49267, Dec. 2023, doi: 10.52202/075280-2142.

[21] S. Tong *et al.*, "Cambrian-1: a fully open, vision-centric exploration of multimodal LLMs," *Advances in Neural Information Processing Systems*, vol. 37, 2024, doi: 10.52202/079017-2771.

[22] M. Deitke *et al.*, "Molmo and PixMo: open weights and open data for state-of-the-art vision-language models," *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 91–104, 2025, doi: 10.1109/CVPR52734.2025.00018.

[23] S. T. Erukude, S. R. Veluru, and V. C. Marella, "Multimodal deep learning: a survey of models, fusion strategies, applications, and research challenges," *International Journal of Computer Applications*, vol. 187, no. 19, pp. 1–7, Jul. 2025, doi: 10.5120/ijca2025925264.

[24] E. Nepovinnykh, V. Immonen, T. Eerola, C. V. Stewart, and H. Kälviäinen, "Re-identification of patterned animals by multi-image feature aggregation and geometric similarity," *IET Computer Vision*, vol. 19, no. 1, Jan. 2025, doi: 10.1049/cvi2.12337.

[25] S. Cui *et al.*, "ShieldVLM: safeguarding the multimodal implicit toxicity via deliberative reasoning with LVLMs: ShieldVLM," *33rd ACM International Conference on Multimedia,* 2025, pp. 11677–11686, doi: 10.1145/3746027.3755711.

[26] M. Taleb, A. Hamza, M. Zouitni, N. Burmani, S. Lafkiar, and N. En-Nahnahi, "Detection of toxicity in social media based on natural language processing methods," *2022 International Conference on Intelligent Systems and Computer Vision, ISCV 2022*, 2022, doi: 10.1109/ISCV54655.2022.9806096.

[27] X. Qin, Y. Zhou, and J. Li, "Multi-modal emotion recognition for online education using emoji prompts," *Applied Sciences*, vol. 14, no. 12, Jun. 2024, doi: 10.3390/app14125146.

[28] Z. Wen, W. Lin, T. Wang, and G. Xu, "Distract your attention: multi-head cross attention network for facial expression recognition," *Biomimetics*, vol. 8, no. 2, May 2023, doi: 10.3390/biomimetics8020199.

[29] Dosovitskiy *et al.*, "An image is worth 16x16 words: transformers for image recognition at scale," *ICLR 2021 - 9th International Conference on Learning Representations*, 2021, pp. 1-21.

[30] J. Dhanith, S. Venkatraman, M. Narendra, V. Sharma, S. Malarvannan, and A. H. Gandomi, "Multimodal emotion recognition using audio-video transformer fusion with cross attention," 2024, *arXiv:2407.18552*.

[31] Y. Hao, L. Dong, F. Wei, and K. Xu, "Self-attention attribution: interpreting information interactions inside transformer," *35th AAAI Conference on Artificial Intelligence, AAAI 2021*, pp. 12963–12971, 2021, doi: 10.1609/aaai.v35i14.17533.

[32] X. Zhang, Z. Wu, K. Liu, Z. Zhao, J. Wang, and C. Wu, "Text sentiment classification based on BERT embedding and sliced multi-head self-Attention Bi-GRU," *Sensors*, vol. 23, no. 3, Jan. 2023, doi: 10.3390/s23031481.

[33] O. Ndama, I. Bensassi, S. Ndama, and E. M. En-Naimi, "Unified BERT-LSTM framework enhances machine learning in fraud detection, financial sentiment, and biomedical classification," *IAES International Journal of Artificial Intelligence*, vol. 14, no. 6, Dec. 2025, doi: 10.11591/ijai.v14.i6.pp5081-5095.

[34] V. Goel, D. Sahnan, S. Dutta, A. Bandhakavi, and T. Chakraborty, "Hatemongers ride on echo chambers to escalate hate speech diffusion," *PNAS Nexus*, vol. 2, no. 3, Mar. 2023, doi: 10.1093/pnasnexus/pgad041.
[35] K. Maity, R. Jain, P. Jha, S. Saha, and P. Bhattacharyya, "GenEx: a commonsense-aware unified generative framework for explainable cyberbullying detection," *EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 16632–16645, 2023, doi: 10.18653/v1/2023.emnlp-main.1035.
[36] A. Botelho, S. A. Hale, and B. Vidgen, "Deciphering implicit hate: evaluating automated detection algorithms for multimodal hate," *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 1896–1907, 2021, doi: 10.18653/v1/2021.findings-acl.166.
[37] J. Peller, "Jigsaw-toxic-comment-classification-challenge," *Kaggle*. Accessed: Jan. 05, 2026. [Online]. Available: https://www.kaggle.com/datasets/julian3833/jigsaw-toxic-comment-classification-challenge
[38] D. Kiela *et al.*, "The hateful memes challenge: detecting hate speech in multi-modal memes," in *34th International Conference on Neural Information Processing Systems*, 2024, pp. 2611–2624, doi: 10.5555/3495724.3495944.
[39] B. Wilie *et al.*, "IndoNLU: benchmark and resources for evaluating Indonesian natural language understanding," *1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pp. 843–857, 2020, doi: 10.18653/v1/2020.aacl-main.85.
[40] B. Li *et al.*, "LLaVA-OneVision: easy visual task transfer," *Transactions on Machine Learning Research*, 2025, pp. 1-44.
[41] A. Yang et al., "Qwen2 Technical Report," 2024, *arXiv*:2407.10671.
[42] Y. Chen *et al.*, "CDEvalSumm: an empirical study of cross-dataset evaluation for neural summarization systems," *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 3679–3691, doi: 10.18653/v1/2020.findings-emnlp.329.
[43] G. Kitukale, N. P. Singh, and S. Quamara, "DDCAF: dynamic dual cross-attention fusion framework for multimodal hate speech detection," *International Journal of Machine Learning and Cybernetics*, vol. 17, no. 6, 2026, doi: 10.1007/s13042-026-03133-1.
[44] S. Farhadpour, T. A. Warner, and A. E. Maxwell, "Selecting and interpreting multiclass loss and accuracy assessment metrics for classifications with class imbalance: guidance and best practices," *Remote Sensing*, vol. 16, no. 3, 2024, doi: 10.3390/rs16030533.
[45] D. M. Burns, N. Leung, M. Hardisty, C. M. Whyne, P. Henry, and S. McLachlin, "Shoulder physiotherapy exercise recognition: Machine learning the inertial signals from a smartwatch," *Physiological Measurement*, vol. 39, no. 7, 2018, doi: 10.1088/1361-6579/aacfd9.
[46] V. Teodorescu and L. Obreja Brașoveanu, "Assessing the validity of k-fold cross-validation for model selection: evidence from bankruptcy prediction using random forest and XGBoost," *Computation*, vol. 13, no. 5, 2025, doi: 10.3390/computation13050127.
[47] R. Meyes, M. Lu, C. W. de Puiseau, and T. Meisen, "Ablation studies in artificial neural networks," 2019, *arXiv*:1901.08644.

BIOGRAPHIES OF AUTHORS

Witta Listiya Ningrum is a lecturer in the Faculty of Computer Science and Information Technology at Gunadarma University. She has also served as a staff member in the Information Systems Laboratory at Gunadarma Universitysince 2014. She holds a bachelor's degree in Information Systems, a master's degree in Business Information Systems, and earned her doctorate in Information Technology in 2024. Her research interests include computer science, machine learning, and natural language processing. She can be contacted at email: wita_listiya@staff.gunadarma.ac.id.

Achmad Benny Mutiara is a professor of Computer Science and Information Technology at Gunadarma University. He obtained his doctorate from the University of Göttingen, Germany, in 2000. He currently serves as the director of the Postgraduate Program at Gunadarma Universityand has over 30 years of academic experience. He is an active member of IEEE and the Central Board of the Indonesian Computer and Informatics Professional Association (IPKIN), and currently serves as the general chairperson of the Informatics and Computer Higher Education Association (APTIKOM). He has published numerous journal and conference papers. His research interests include computer science, artificial intelligence, and simulation modeling. He can be contacted at email: amutiara@staff.gunadarma.ac.id.

Diana Ikasari is a lecturer in the Faculty of Computer Science and Information Technology at Gunadarma University. She obtained her doctorate in Information Technology from Gunadarma University, in 2018. She has conducted various scientific studies in Information Systems Technology. Her research interests include databases and natural language processing. She is the author of the book 'Information retrieval', a co-author of the textbook 'Introduction to geographic information systems', and a contributor to the book chapter "Geographic Information systems in the application of straight-line distance calculation to SMAN Depok." She can be contacted at email: d_ikasari@staff.gunadarma.ac.id.