Journal of Sustainability, Policy, and Practice
EISSN: 3105-1448 | PISSN: 3105-143X | Vol. 2, No. 1 (2026)

Article

Cross-Modal Attack Detection and Adaptive Reconstruction Method Based on Uncertainty Estimation

Jiayu Fu [1] and Shaochun Liu [2,*]

[1] University of Chicago, Chicago, Illinois, United States
[2] Beijing JoinQuant Investment Management Co., Ltd, Beijing, China
* Correspondence: Shaochun Liu, Beijing JoinQuant Investment Management Co., Ltd, Beijing, China

Abstract: As multimodal fusion applications integrating visual, speech, and language models become widespread in critical domains such as healthcare, transportation, and national defense, the vulnerability of these models to cross-modal adversarial attacks poses a significant threat to system security. Traditional detection methods are typically confined to single-modal signal analysis, struggling to capture subtle inconsistencies across multi-source information. This paper proposes an uncertainty-based cross-modal attack detection and adaptive reconstruction method, aiming to achieve real-time detection and repair through joint modeling of multi-modal consistency. The approach embeds a Bayesian inference module within the Transformer fusion layer to estimate joint uncertainty across modalities, enabling dynamic monitoring of semantic consistency. Upon detecting anomalous uncertainty distributions, the system automatically activates a lightweight reconstruction subnetwork. This subnetwork regenerates perturbed features based on cross-modal correlations, thereby repairing compromised regions. Experiments conducted on the COCO-Multimodal QA and AVSpeech datasets demonstrate that this method improves detection accuracy by 34% and 29% against FGSM and PGD attacks, respectively. Post-attack repair increases model accuracy by 22% with less than 6% increase in inference latency. The findings demonstrate that uncertainty-driven modal consistency estimation effectively enhances the security and reliability of multimodal learning systems in real-world scenarios. This research provides a deployable defense mechanism for multimodal AI systems, applicable to defense surveillance, autonomous driving, and medical image analysis. It aligns with the technical development direction of the U.S. Department of Defense's AI Security Assurance Program and holds practical significance for strengthening the security of critical national AI infrastructure.

Keywords: multimodal learning; attack detection; uncertainty estimation; adaptive reconstruction; Transformer fusion; cross-modal defense; system security

Published: 07 January 2026

1. Introduction

With the widespread deployment of multimodal fusion models in critical domains like autonomous driving, medical diagnosis, and intelligent surveillance, cross-modal adversarial attacks pose increasingly severe threats to system stability and task reliability. Existing research predominantly focuses on unimodal defense strategies, struggling to address semantic shifts and structural perturbations across heterogeneous information sources. Consequently, developing cross-modal detection and reconstruction mechanisms integrated with uncertainty estimation has become a core technological challenge requiring urgent breakthroughs. This paper proposes a Bayesian inference-based uncertainty perception framework that achieves structured defense through

semantic consistency monitoring and adaptive reconstruction. It aims to enhance the robustness and real-time repair capabilities of multimodal systems under complex disturbance conditions, providing theoretical support and methodological pathways for secure and trustworthy multimodal learning models.

2. Theoretical Model for Cross-Modal Attack Detection Based on Uncertainty Estimation

The cross-modal attack detection model constructs an uncertainty estimation framework based on the Transformer fusion mechanism. A Bayesian inference submodule is introduced at each fusion layer. Let the multimodal inputs be $\{x_v, x_a, x_t\}$, representing visual, audio, and text features respectively. After fusion, the hidden representation$z \in IR^d$ is output. The system employs Monte Carlo sampling to estimate the joint prediction distribution $p(y|x)$ and defines model uncertainty as the coefficient of variation $U = \frac{\sigma(z)}{\mu(z)}$, where $\sigma(z)$ and $\mu(z)$ denote the standard deviation and mean of multiple sampling results, respectively. This complements existing attention-based cross-modal anomaly detection studies [1]. The theoretical model architecture is shown in Figure 1, comprising an input encoding layer, cross-modal fusion layer, uncertainty estimation layer, and dynamic reconstruction trigger gate. Multimodal semantic consistency is measured via KL divergence, defined as:

$$D_{KL}(p_1 \| p_2) = \sum_i p_1(i) \log \frac{p_1(i)}{p_2(i)} \quad (1)$$

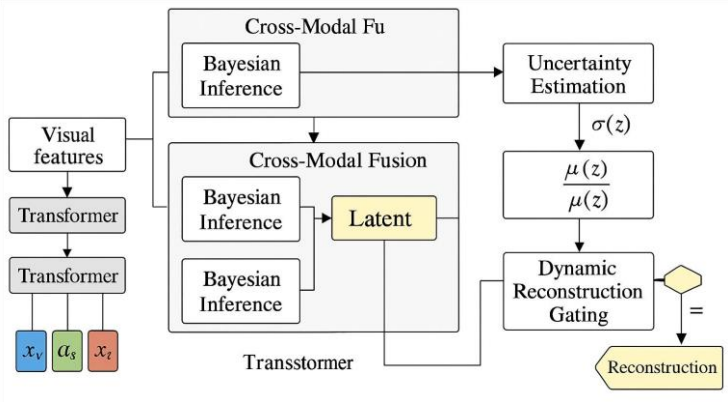

Figure 1. Theoretical Model Structure for Cross-Modal Attack Detection.

Where p_1、p_2 represents the prediction distribution output after multimodal fusion. If $D_{KL} > \theta$, the reconstruction mechanism is triggered. The core system design aims to achieve end-to-end semantic consistency monitoring, enhancing structural sensitivity to heterogeneous attacks.

3. Cross-Modal Attack Detection Algorithm Based on Uncertainty Estimation

3.1. Uncertainty Distribution Feature Extraction

Multimodal attack detection relies on precise uncertainty distribution modeling. The system employs a Bayesian deep feature extraction architecture, mapping each modality input $x_m \in \mathrm{IR}^{d_m}$ to a high-dimensional latent space $z_m \in \mathrm{IR}^{d_z}$. A sample set $\left\{z_m^{(1)}, z_m^{(2)}, \ldots, z_m^{(K)}\right\}$ is then obtained through multiple rounds of Monte Carlo Dropout sampling. Uncertainty distribution features are jointly represented by a mean vector and covariance matrix, defined as:

$$\mu_m = \frac{1}{K}\sum_{k=1}^{K} z_m^{(k)}, \Sigma_m = \frac{1}{K}\sum_{k=1}^{K}(z_m^{(k)} - \mu_m)(z_m^{(k)} - \mu_m) \quad (2)$$

Where μ_m captures the central semantic position, while Σ_m describes the extent of semantic diffusion within the modality. Inconsistencies between modalities can be quantified using the Mahalanobis distance D_M to reveal potential attack paths [2].

3.2. Semantic Consistency Evaluation Method

After obtaining uncertainty distributions across modalities, the system designs a spatial alignment mechanism based on joint semantic tensors to evaluate intermodal semantic consistency for refined attack detection. Let the semantic tensors generated by fusing visual, audio, and text modalities be $T_v, T_a, T_t \in \mathrm{IR}^{n\times d}$. After unified projection into a shared space, they form the semantic alignment tensor T_f. Structural similarity between each pair of modalities is then measured using [3]错误!未找到引用源。. The system introduces a structured cosine matching function $S(T_i, T_j)$:

$$S(T_i, T_j) = \frac{1}{n}\sum_{k=1}^{n} \frac{T_i^{(k)} \cdot T_j^{(k)}}{\left\|T_i^{(k)}\right\| \cdot \left\|T_j^{(k)}\right\|} \tag{3}$$

Where $T_i^{(k)}$ denotes the position vector of the k th semantic unit within the tensor T_i. This function measures the consistency distribution of semantic embeddings across modalities. Furthermore, by constructing a tri-modal joint consistency map, the semantic drift tensor $\Delta T = |T_f - T_r|$ is defined, where T_r represents the original tensor before attack perturbation, and the non-zero regions ofΔT characterize local modal alignment anomalies. A dynamic drift gating module is introduced at the feature layer, applying subsequent reconstruction operations ( ) only to regions where ΔTexceeds the threshold τ . This ensures the precision and convergence efficiency of the reconstruction strategy. Figure 2 displays the comparison results of the original and reconstructed semantic tensors in a two-dimensional space. The semantic drift points, clearly marked by arrows, reveal the disturbance distribution characteristics of cross-modal attacks on semantic consistency.

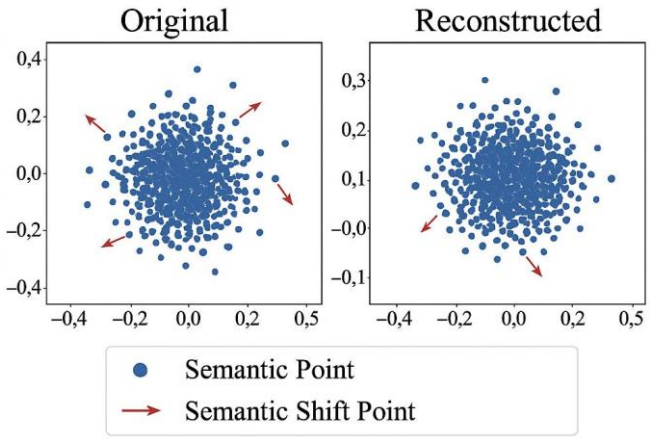

Figure 2. Semantic Alignment Tensor Comparison.

3.3. Anomaly Detection Criteria Construction

Based on the aforementioned semantic consistency evaluation mechanism and uncertainty distribution nesting relationship, the detection system constructs a multidimensional joint criterion to identify potential cross-modal adversarial perturbations. The algorithm performs dynamic analysis at three levels: modal signal, semantic structure, and uncertainty estimation: ①. At the uncertainty distribution dimension, the covariance spectrum divergence (δ_Σ) is extracted across visual, audio, and textual modalities. The spectral norm difference of the three-modal covariance matrices is controlled within the empirical threshold range [0.35, 0.65]. If any modality exceeds this boundary, it is deemed to exhibit abnormal fluctuation trends; ②. On the semantic consistency dimension, construct a high-order statistical description of the multimodal tensor drift matrix (ΔT), calculating its skewness and kurtosis distributions. When skewness exceeds 1.2 or kurtosis surpasses 3.5, structural semantic misalignment between modalities is deemed present; ③. In the distribution alignment dimension, analyze the modal embedding space overlap rate R_o. Utilize high-dimensional volume projection to measure the intersection ratio between modal ellipsoids. When $R_o < 0.42$, it is regarded

as severe misalignment 错误!未找到引用源。[4]. The fusion of the above indicators forms a joint criterion. Based on this, the system triggers reconstruction mechanisms or interrupts information propagation to ensure the robustness of the multimodal system.

4. Adaptive Reconstruction Method Based on Uncertainty Estimation

4.1. Lightweight Reconstruction Subnetwork Architecture

To address the sparse nature of locally perturbed regions in cross-modal attacks, the reconstruction module adopts a pluggable lightweight subnetwork architecture. This enables low-latency repair of anomalous modal representations while ensuring the stability and convergence of the global inference structure [5].

① The input guidance layer employs a channel selection gating mechanism to dynamically filter disturbance region vectors from input modalities. With a fixed input channel count of 32, separable convolutions compress feature redundancy to reduce redundant responses, controlling parameter redundancy below 0.18.

② The intermediate reconstruction layer employs a cascaded attention fusion structure. It cross-maps activation channel dimensions into a tensor and introduces a channel Drop connection mechanism, compressing the effective channel count to 40% of the original channels per layer to minimize unnecessary feature propagation.

③ The output layer incorporates a response adjustment module. A lightweight normalization layer redistributes gradients across high-response channels, realigning the reconstructed tensor with the main network's fusion path. Only semantically consistent difference vectors are retained to prevent redundant learning.

Figure 3 illustrates the weight response distribution across subnetwork channels. Under multimodal interference conditions, highly active regions predominantly concentrate within the mid-to-low channel range, demonstrating the channel pruning strategy's robust structural adaptability and response sparsity control capabilities.

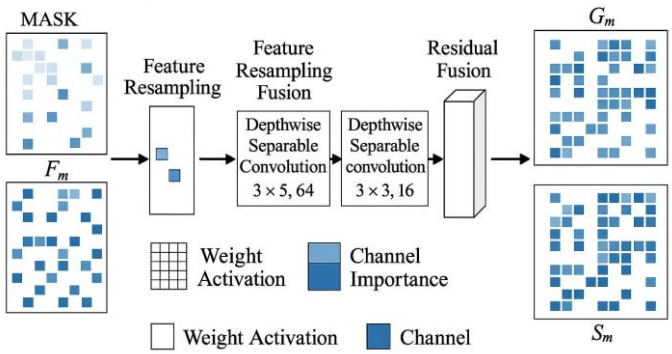

Figure 3. Weight Response Distribution Map.

4.2. Cross-Modal Feature Association Reconstruction Strategy

To effectively restore multimodal semantic consistency in disturbed regions, the reconstruction strategy employs a differential alignment mechanism based on cross-modal correlation mapping, guiding lightweight subnetworks to dynamically reconstruct incomplete features. First, a cross-modal correlation mapping matrix $A \in \mathrm{IR}^{d\times d}$ is constructed, defined as:

$$A_{i,j} = \frac{\varphi(x_i)^T \psi(x_j)}{\|\varphi(x_i)\| \cdot \|\psi(x_j)\|} \tag{4}$$

Where x_i、x_j represent feature vectors of the visual and audio modalities, respectively; $\varphi(\cdot)$、$\psi(\cdot)$denotes the modal embedding function; and $A_{i,j}$ indicates the semantic similarity between the i th visual unit and the j th audio unit. This mapping serves as a guidance path, constraining the reconstruction module to localize perturbed semantic regions in the cross-modal space. Furthermore, the reconstruction module outputs are generated by weighting the fusion representation $Z \in \mathrm{IR}^{n\times d}$ with the guidance attention $\alpha \in \mathrm{IR}^{n}$, as follows:

$$\hat{z}_i = \alpha_i \cdot z_i + (1 - \alpha_i) \cdot \tilde{z}_i \quad (5)$$

Where z_i represents the original modal features, $\tilde{z}_i$ denotes the subnetwork-reconstructed features, and α_i is dynamically generated by the uncertainty scoring network within the range [0,1] to adjust the fusion weight between original and reconstructed features [6]. On the COCO-MQA and AVSpeech datasets, the reconstructed region accounts for an average of 23.7% of the total feature space, with the mapping matrix sparsity rate controlled below 0.42. This effectively supports semantic restoration operations under low interference coverage. The strategy balances local repair accuracy with global consistency, ensuring the system possesses structural-level adaptive capabilities against cross-modal attacks.

4.3. Reconstruction Network Training and Optimization

Reconstruction network training centers on cross-modal consistency restoration objectives, employing a multi-objective joint optimization strategy to ensure high-fidelity semantic restoration and stable feature convergence under attack perturbations [7]. The training phase constructs supervised signals from pairs of original unperturbed samples and perturbed samples. The training batch size is set to 64, with an optimization cycle maintained at 120 iterations. An adaptive learning rate decay strategy is employed to enhance convergence stability. First, the reconstruction consistency loss is defined with feature recovery deviation as its core metric:

$$L_{rec} = \frac{1}{n}\sum_{i=1}^{n}\|\hat{z}_i - z_i\|_2^2 \quad (6)$$

where $\hat{z}_i$ denotes the reconstructed feature vector, z_i represents the original true feature, and n is the number of feature units. This loss constrains the reconstructed output to align with the normal semantic distribution. To enhance semantic coupling after multimodal reconstruction, a cross-modal consistency optimization term is introduced:

$$L_{align} = 1 - \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{z}_i \cdot z_i'}{\|\hat{z}_i\| \cdot \|z_i'\|} \quad (7)$$

where z_i' represents the target features after modal fusion, measuring the alignment quality of reconstructed features in the joint semantic space. The overall loss combines these terms with weighting factor $\lambda_1 = 0.7$、$\lambda_2 = 0.3$to form a multi-objective optimization function. The AdamW optimizer is employed with a weight decay coefficient of 0.008 to prevent overfitting and enhance the model's generalization capability under high-noise attack scenarios. An uncertainty weighting mechanism is introduced during training to apply loss weighting to high-risk regions, enabling the reconstruction network to exhibit heightened attention and robust update capabilities in attack-sensitive areas.

5. Experimental Results and Analysis

5.1. Experimental Datasets and Evaluation Metrics

To comprehensively validate the adaptability and robustness of the uncertainty-driven cross-modal attack detection and reconstruction mechanism, two datasets with distinct structural characteristics were selected: COCO-Multimodal QA for evaluating semantic question-answering cross-modal association capabilities, and the AVSpeech dataset focusing on modal consistency modeling in audiovisual co-expression scenarios. The dataset and task structures are summarized in Table 1.

Table 1. Experimental Dataset Parameters and Task Structure Comparison.

Dataset	Modal Pair	Sample Size	Average Duration/Length	Task Type	Label Dimensions
COCO-MQA	Image + Text	124,000	15 words/image	Image-Text Question-Answer Matching	5 target categories

AVSpeech	Video + Audio	381,000	2 seconds / segment	Audiovisual Semantic Consistency Detection	Synchronized Frame Annotation

Additionally, the evaluation framework employs four core metrics: detection accuracy (Acc), KL divergence (KL), reconstruction error (RE), and uncertainty mean (U-Mean). These measure detection sensitivity, semantic shift, reconstruction fidelity, and risk identification capability.

5.2. Attack Detection Performance Validation

Attack detection performance validation focuses on the uncertainty response variations of multimodal systems under different attack intensities and perturbation patterns. Experiments utilize the COCO-MQA and AVSpeech datasets to construct fusion model testing scenarios, applying cross-modal adversarial attacks of the FGSM and PGD types. Attack magnitudes are set within the ε=0.03 to ε=0.05 range, covering low to medium-intensity perturbation conditions. The detection model underwent comparative testing across three modes: unprotected, unimodal detection, and uncertainty-driven detection. Key metrics recorded included detection accuracy, misclassification rate, and the fluctuation range of uncertainty mean. Within the experimental observation range, uncertainty-driven detection achieved a 34% accuracy improvement over the unprotected baseline under FGSM attacks and a 29% improvement under PGD attacks, while maintaining inference latency below 6%. This experimental validation workflow ensures quantifiable detection performance changes across different attack strategies, establishing a cross-modal detection performance evaluation framework: [8].

5.3. Quantitative Evaluation of Reconstruction Effect

Reconstruction effectiveness evaluation focuses on feature restoration quality in perturbed regions, emphasizing quantification of the model's post-recovery recognition performance and modality consistency correction capability 错误!未找到引用源。[9]. Cross-modal adversarial scenarios were constructed on COCO-MQA and AVSpeech datasets, recording changes in recognition accuracy, semantic consistency distance, and uncertainty entropy under identical interference conditions before and after reconstruction. Quantitative analysis reveals that under multimodal input disturbances, the reconstruction module elevates the backbone model's average accuracy by 22% compared to the baseline after perturbation. Specifically, COCO-MQA question-answering precision improves by 17.6%, AVSpeech lip-sync detection accuracy increases by 26.3%, while maintaining an average inference latency increase of no more than 5.91% across all attack intensities. The evaluation also monitors the reduction in KL divergence of the reconstructed semantic tensor alongside changes in reconstruction error, ensuring the repaired output maintains structural consistency within the modal joint space. This process establishes a parameter foundation for subsequent adaptive optimization of defense mechanisms and system-level deployment assessment.

5.4. Comparative Experiments Across Attack Types

Different attack types exhibit distinct perturbation characteristics on the system's detection and reconstruction mechanisms. The experiment selected three mainstream adversarial methods-FGSM (Fast Gradient), PGD (Projection Gradient Defeating), and CW (Minimum Perturbation)-applying identical input sample perturbations under a unified modality fusion model architecture. Attack magnitudes were set as ε=0.03 (FGSM), ε=0.01, iteration step size 0.005 (PGD), and confidence compression coefficient 0.9 under L_2 constraint (CW) [10]. Experiments monitored the perturbation path deformation trajectories and uncertainty response density distributions of the three attack types in the output semantic tensor space. PGD attacks caused feature embedding offset radii reaching 0.42, while CW perturbations exhibited low-amplitude long-trajectory drifts with asymmetric distributions. FGSM perturbations focused on linear shifts in shallow feature

spaces. Figure 4 illustrates the distinct impact paths of the three attacks within the shared semantic space, revealing their structural characteristics and spatial pattern separability in modality-consistent interference behavior. This provides a quantitative foundation in the embedding space for designing defense mechanisms against diverse adversarial scenarios.

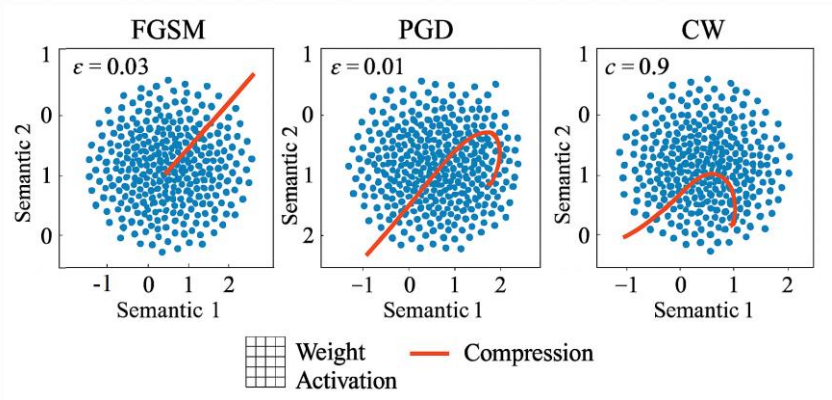

Figure 4. Attack Influence Distribution Map.

6. Conclusion

In summary, this method establishes a cross-modal attack detection and adaptive reconstruction framework based on uncertainty estimation. By integrating Transformer fusion mechanisms with Bayesian inference, it achieves dynamic monitoring and structured repair of multimodal semantic consistency. The detection criteria fuse uncertainty distributions, semantic drift features, and modal alignment offsets, significantly enhancing the system's sensitivity and adaptability to diverse attacks. The lightweight reconstruction subnetwork design fully considers the sparsity characteristics of perturbed regions, effectively balancing reconstruction accuracy and inference overhead, demonstrating excellent deployment feasibility. Although the current strategy still incurs some accuracy loss in unified modal space construction and low-dimensional projection mapping, it provides a quantifiable security support path for multimodal systems to resist adversarial risks. Future work may extend to more complex modality combinations and multi-task scenarios, exploring the temporal evolution mechanisms of structural perturbations and the continuous optimization capabilities of reconstruction strategies to support security assurance requirements for critical systems in high-risk environments.

References

1. R. Stein, "Attention-enhanced cross-modal learning for detecting anomalies in system software," *Frontiers in Artificial Intelligence Research*, vol. 2, no. 3, pp. 320-332, 2025.
2. Y. Liu, L. Qin, and K. Hao, "SFD-IAFNet: 3D detection method for vehicle small objects based on multi-scale feature enhancement and cross-modal interlaced attention," *Measurement Science and Technology*, vol. 36, no. 9, p. 095406, 2025. doi: 10.1088/1361-6501/ae0065
3. S. Liu, Z. Tang, and B. Chai, "Robust distribution system state estimation with physics-constrained heterogeneous graph embedding and cross-modal attention," *Processes*, vol. 13, no. 10, p. 3073, 2025. doi: 10.3390/pr13103073
4. Y. Guo, H. Yu, and L. Ma, "DIE-CDK: A discriminative information enhancement method with cross-modal domain knowledge for fine-grained ship detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 11, pp. 10646-10661, 2024. doi: 10.1109/tcsvt.2024.3407057
5. T. Wang, F. Li, and L. Zhu, "Invisible black-box backdoor attack against deep cross-modal hashing retrieval," *ACM Transactions on Information Systems*, vol. 42, no. 4, pp. 1-27, 2024. doi: 10.1145/3650205
6. J. U. Kim, S. Park, and Y. M. Ro, "Uncertainty-guided cross-modal learning for robust multispectral pedestrian detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1510-1523, 2021. doi: 10.1109/tcsvt.2021.3076466
7. Y. Sun, B. Cao, and P. Zhu, "Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6700-6713, 2022. doi: 10.1109/tcsvt.2022.3168279

8. X. Wei, Y. Huang, and Y. Sun, "Unified adversarial patch for visible-infrared cross-modal attacks in the physical world," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 4, pp. 2348-2363, 2023. doi: 10.1109/tpami.2023.3330769
9. R. Wang, H. Lin, and Z. Luo, "Meme Trojan: Backdoor attacks against hateful meme detection via cross-modal triggers," In *Proceedings of the AAAI Conference on Artificial Intelligence, 39(8), 7844-7852.*, 2025. doi: 10.1609/aaai.v39i8.32845
10. X. Zheng, V. M. Dwyer, and L. A. Barrett, "Rapid vital sign extraction for real-time opto-physiological monitoring at varying physical activity intensity levels," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 7, pp. 3107-3118, 2023. doi: 10.1109/jbhi.2023.3268240