A Bi-Mamba Fusion Framework With KAN Pattern Mining For Multi-modal Sensitive Content Detection

Ming Lian* and Teng Li

School of Artificial Intelligence, Anhui University, Hefei 230039 China

Corresponding author. E-mail: lianming@mail.ustc.edu.cn

Received: May. 07, 2026; Accepted: Jun. 04, 2026

The meteoric expansion of social media platforms has established multi-modal data as the dominant paradigm for global digital communication. However, this evolution has concurrently facilitated the rapid dissemination of sensitive and harmful material, including hate speech, extremist propaganda, and violent imagery, which presents a severe challenge to online security and public psychological well-being. Modern automated detection systems frequently falter because they fail to bridge the heterogeneous semantic gap between visual and linguistic modalities effectively, often relying on late-stage fusion or computationally expensive Transformer architectures that struggle with long-range dependencies. Furthermore, the conventional use of multi-layer perceptrons (MLPs) as predictors often fails to capture the irregular and complex decision boundaries inherent in high-dimensional multi-modal manifolds. To address these critical limitations, we propose MamKAN, a sophisticated multi-modal fusion framework that integrates State Space Models with spline-based neural architectures. Our methodology innovates at three distinct stages: First, we implement an implicit modal interaction strategy using textual inversion, which maps visual features into pseudo-word tokens within the text embedding space to achieve robust early-stage semantic alignment. Second, we introduce a bidirectional Mamba fusion module based on state space models; this mechanism achieves linear computational complexity while simultaneously capturing comprehensive vision-to-text and text-to-vision contextual dependencies through forward and backward scanning. Finally, we replace traditional MLPs with a Kolmogorov-Arnold Network (KAN) predictor. By employing learnable B-spline activation functions on the network edges, the KAN module adaptively fits complex non-linear distributions, significantly enhancing discriminative precision. Comprehensive experiments conducted on the HMC and HarMeme benchmark datasets demonstrate that MamKAN consistently and significantly outperforms state-of-the-art models across three metrics.

Keywords: State space models; Kolmogorov-Arnold network; Sensitive content detection

http://dx.doi.org/10.6180/jase.202610_33.030

1. Introduction

The digital landscape has undergone a paradigm shift characterized by the astronomical proliferation of social media platforms and diverse digital ecosystems. This expansion has catalyzed an unprecedented surge in multi-modal data, where visual imagery and linguistic descriptions are inextricably intertwined to articulate sophisticated and nuanced semiotic meanings [1–3]. While this connectivity serves as a cornerstone for global information exchange, it simultaneously acts as a conduit for the rapid dissemination of sensitive and deleterious content. The spread of hate speech, graphic violence, and extremist rhetoric poses severe threats to digital ecosystem integrity and the psychological well-being of global users, rendering automated content moderation an urgent societal and technical

necessity [4, 5].

However, the robust identification of sensitive multi-modal content remains a formidable challenge due to the complex, context-dependent synergy between disparate modalities. The toxicity of a particular post often resides not in a single modality but in the subtle interplay between them; for instance, a seemingly innocuous image can be transformed into a highly offensive message when juxtaposed with a provocative or derogatory caption. Consequently, there is an imperative demand for sophisticated detection architectures capable of bridging this heterogeneous semantic gap and deciphering the intricate cross-modal dependencies that define harmful content [6, 7]. Despite significant advancements, current methods face two key limitations that create a clear research gap. First, many frameworks struggle with effective and efficient multi-modal fusion. Traditional dual-stream architectures often process visual and textual features independently and combine them too late in the process, missing crucial early-stage interactions. While Transformer-based models offer better fusion through cross-attention, they are hampered by quadratic computational complexity, which is inefficient for modeling long-range dependencies [8–10]. Second, the majority of existing models employ multi-layer perceptrons (MLPs) as the final predictor. Since MLPs use fixed activation functions, they often fail to fit the highly irregular and complex decision boundaries inherent in fused multi-modal data, leading to suboptimal classification accuracy [11–13].

To overcome these critical limitations, this paper introduces MamKAN, a sophisticated Multi-modal Mamba fusion framework synergized with a Kolmogorov-Arnold Network (KAN) predictor for high-precision sensitive content detection. The architecture of MamKAN initiates with a multi-modal feature extraction module that strategically employs textual inversion to inject visual semantics directly into the linguistic embedding space, thereby ensuring robust early-stage modal awareness and semantic alignment. These modality-aware features are subsequently processed by a bidirectional Mamba fusion module, which capitalizes on the efficiency of State Space Models to capture the comprehensive and intricate vision-to-text and text-to-vision information flows. This mechanism facilitates the modeling of deep, non-linear contextual dependencies with linear computational complexity, effectively surmounting the scaling bottlenecks of traditional attention-based architectures. Finally, instead of relying on the conventional MLPs with static activation functions, we introduce an adaptive pattern mining module based on KAN. By utilizing learnable B-spline activation functions positioned on the edges of the network, the KAN predictor can adaptively approximate complex non-linear distributions and irregular decision boundaries. This structural innovation significantly enhances the model's discriminative power, allowing it to more accurately distinguish between benign and sensitive content within complex, high-dimensional manifolds.

The contributions of this work are three-fold

- We introduce an implicit modal interaction strategy via textual inversion during the feature extraction phase. By mapping visual features into pseudo-word tokens that are injected into the text prompt, we ensure that the textual representation is inherently aware of the visual context, facilitating early-stage semantic alignment.
- We propose a bidirectional Mamba fusion mechanism based on State Space Models. Unlike standard Transformers, this module achieves linear computational complexity while effectively capturing long-range, non-linear dependencies between modalities through forward and backward state-space scanning.
- By replacing traditional MLPs with KAN, our model leverages learnable spline-based transformations on edges to approximate complex decision boundaries more accurately, providing superior adaptability to irregular multi-modal feature distributions.

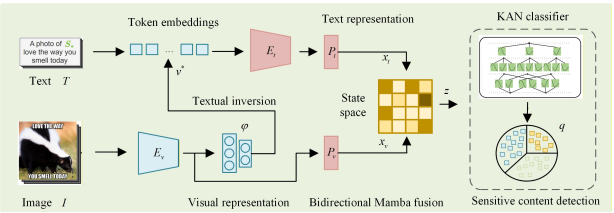

Fig. 1. The illustration of MamKAN. It consists of a multi-modal feature extraction with textual inversion, a bidirectional Mamba fusion, and an adaptive pattern mining

2. Method

Multi-modal sensitive content detection aims to determine whether a given pair of image and text contains inappropriate, harmful, or restricted information. Formally, let $\mathcal{D} = \{(I_n, T_n, y_n)\}_{n=1}^{N}$ denote a dataset of N samples, where $I_n \in \mathcal{I}$ represents the input image and $T_n \in \mathcal{T}$ represents the associated textual description. Each sample is assigned a binary ground-truth label $y_n \in \{0, 1\}$, where

$y_n = 1$ indicates that the multi-modal content is sensitive, and $y_n = 0$ denotes benign content. The fundamental objective is to learn a mapping function $\mathcal{F} : \mathcal{I} \times \mathcal{T} \to [0, 1]$, parameterized by Θ, which predicts the probability q_n of a sample being sensitive. To achieve the goal above, this paper proposes a multi-modal Mamba fusion with Kolmogorov-Arnold Network predictor for sensitive content detection (MamKAN), which consists of a multi-modal feature extraction with textual inversion, a bidirectional Mamba fusion, and an adaptive pattern mining. The architecture of MamKAN is shown in Fig. 1.

2.1. Multi-modal Feature Extraction with Textual Inversion

Traditional dual-stream architectures often process visual and textual features independently, resulting in weak semantic alignment during the early stages. To overcome this limitation, this module leverages textual inversion to introduce an implicit modal interaction during the feature extraction phase. By injecting visual features into the text embedding space, we ensure that the extracted text representation contains not only semantic information but is also aware of the specific image content, thereby providing robust foundational features for subsequent fusion.

Specifically, MamKAN leverages the pre-trained CLIP model [11] and the textual inversion network ϕ [14] for feature extraction. Given an input image I and its associated text T, MamKAN first extracts the visual features via the CLIP visual encoder E_v and applies an image projection layer P_v to obtain the visual representation:

$$x_v = P_v(E_v(I)) \in \mathbb{R}^d \tag{1}$$

where d denotes the dimension of features. Simultaneously, the textual inversion network ϕ maps the visual features to a pseudo-word token $v_* = \phi(E_v(I))$. Subsequently, MamKAN constructs an augmented prompt T_{prompt} = "a photo of S_*, T", where S_* is the pseudo-word associated with v_*. This prompt is processed by the CLIP text encoder E_t and a text projection layer P_t to generate the text representation:

$$x_t = P_t(E_t(T_{prompt})) \in \mathbb{R}^d \tag{2}$$

These text representations uniquely encompass rich visual semantics due to the inversion process.

2.2. Bidirectional Mamba Fusion

Detecting sensitive content often relies on complex, nonlinear contextual dependencies between vision and text, which simple concatenation fails to capture. To address this, MamKAN employs a bidirectional Mamba multi-modal fusion module. Unlike Transformers, Mamba is built on State Space Models (SSM), enabling efficient modeling of long-range dependencies with linear computational complexity. By utilizing bidirectional scanning, the model simultaneously captures the information flow from "vision-to-text" and "text-to-vision", generating a comprehensive fusion representation.

Specifically, the extracted features are concatenated into a sequence $\mathbf{s} = [x_v, x_t] \in \mathbb{R}^{2\times d}$. The bidirectional Mamba fusion module is built on a continuous-time SSM mapping, i.e.,

$$y(t) = \mathbf{C} \cdot h(t) + \mathbf{D} \cdot s(t), \tag{3}$$

$$\dot{h}(t) = \mathbf{A} \cdot h(t) + \mathbf{B} \cdot s(t), \tag{4}$$

where $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ are the system parameters. $s(t)$, $h(t)$, and $y(t)$ are the input, hidden state, and the output, respectively. $\dot{h}(t)$ is the time derivative of $h(t)$. For discrete computation, the module introduces a time-scale parameter Δ to discretize the system:

$$h_t = \overline{\mathbf{A}} h_{t-1} + \overline{\mathbf{B}} \mathbf{s}_t, \quad y_t = \mathbf{C} h_t \tag{5}$$

where t is the sequence index, $\overline{\mathbf{A}} = \exp(\Delta\mathbf{A})$, $\overline{\mathbf{B}} = (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - I)\Delta\mathbf{B}$. The module performs both forward and backward scans to capture bidirectional context and apply average pooling to obtain the final fusion representation $\mathbf{z}$:

$$\mathbf{z}_{fwd} = \mathrm{SSM}(\mathbf{s}), \quad \mathbf{z}_{bwd} = \mathrm{SSM}(\mathrm{Flip}(\mathbf{s})) \tag{6}$$

$$\mathbf{z} = \mathrm{AvgPool}([\mathbf{z}_{fwd}; \mathbf{z}_{bwd}]) \in \mathbb{R}^d \tag{7}$$

where AvgPool$(\cdot)$ and Flip$(\cdot)$ denote the average pooling operation and the reverse operation, respectively. $[\cdot]$ is the concatenation.

2.3. Adaptive pattern mining

Multi-modal fusion representations often exhibit complex distributions. Traditional MLPs with fixed activation functions struggle to fit irregular decision boundaries. We introduce the Kolmogorov-Arnold Network [15] as the predictor, which places learnable activation functions on edges rather than nodes. This design allows KANs to approximate complex non-linear functions more accurately than MLPs, enhancing classification precision.

Specifically, the fusion representation $\mathbf{z}$ is fed into a standard KAN classifier, where each layer is defined with a base transformation and a B-Spline transformation. The base transformation aims to leverage a linear transformation with a nonlinear activation to learn primary patterns from data, which can be expressed as follows:

$$\mathbf{z}_{base} = w_b \cdot \sigma(\mathbf{z}) \tag{8}$$

where w_b and σ denote the learning parameters and the nonlinear activation function, respectively.

The B-Spline transformation leverages smooth interpolations among data to adaptively fit complex patterns. It relies on a predefined grid structure $\mathbf{G} \in \mathbb{R}^{r+2\times o+1}$, characterized by the grid resolution r (which determines the number of intervals) and the polynomial order o. The transformation is mathematically formulated as a weighted summation of basis functions:

$$\mathbf{z}_{spline} = \sum_i c_i \cdot \mathbf{B}_i^o(\mathbf{z}) \tag{9}$$

where the coefficients c_i represent the trainable parameters. $\mathbf{B}_i^o(\cdot)$ signifies the i-th B-spline basis function of degree o. The grid points $\mathbf{G} = [g_{-o}, \ldots, g_{r+o}]$ are arranged uniformly within the bounded interval $[-1, 1]$, thereby governing the granularity of the spline interpolation. The basis functions $\mathbf{B}_i^o(\cdot)$ are derived recursively via:

$$\mathbf{B}_i^0(x) = \begin{cases} 1 & \text{if } x \in [g_i, g_{i+1}) \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

$$\mathbf{B}_i^o(x) = \frac{x - g_i}{g_{i+o} - g_i}\mathbf{B}_i^{o-1}(x) + \frac{g_{i+o+1} - x}{g_{i+o+1} - g_{i+1}}\mathbf{B}_{i+1}^{o-1}(x) \tag{11}$$

where $\mathbf{B}_i^0(\cdot)$ serves as the initial zero-degree basis, and $\mathbf{B}_i^o(\cdot)$ represents the basis of degree o. The subscript i indexes the specific spline component within the grid. Finally, the output of each KAN layer is defined as:

$$\mathbf{z}_{KAN} = \mathbf{z}_{base} + \mathbf{z}_{spline} \tag{12}$$

By stacking L KAN layers as the predictor, we obtain the final prediction result q of each sample.

2.4. The loss function

To integrate feature extraction, multi-modal fusion, and classification into a unified, end-to-end trainable framework, an objective function is required to effectively measure the divergence between the predicted probability distribution and the ground truth. Given that sensitive content detection is inherently a binary classification task, we employ the binary cross-entropy loss. Given a mini-batch of size N, let y_i denote the ground truth label for the i-th sample and let q_i be the predicted probability output by the KAN predictor. The total loss function $\mathcal{L}$ is defined as:

$$\mathcal{L} = -\frac{1}{N}\sum_{n=1}^{N}[y_n \log(q_n) + (1 - y_n)\log(1 - q_n)] \tag{13}$$

3. Results and discussion

3.1. Set up

Benchmark datasets: Aligning with the baselines [8, 9], our experiments are conducted on the HMC and HarMeme datasets. HMC is a large-scale dataset of synthetic memes generated by Facebook, aimed at detecting hate speech regarding race, religion, sex, and disability. The data split includes 8,500 items for training, 500 for validation, and 2,000 for testing. In contrast, HarMeme features real-world content related to COVID-19. We adopt the preprocessing strategy to convert its original three-tier annotation (very harmful, partially harmful, harmless) into a binary format by merging the first two labels into a hateful category. Consequently, the dataset provides 3,013 training memes and 354 testing memes.

Implementation details: The learning rate, the batch size, and the epoch number are set as 0.001, 48, 500 on across datasets, respectively. The feature dimension d is set as 768. The image size is set to 224×224. All the experiments are running on a ubuntu (20.04) server with an NVIDIA Tesla V100 GPU. In the experiments, the parameters of both the pre-trained CLIP model and the textual inversion network ϕ are kept frozen throughout the entire training phase. Kolmogorov-Arnold Network structured as $[768, 64, K]$ is configured with a grid size of r=2, a spline order of o=1, K is the class number. To comprehensively validate the effectiveness of our proposed method, we report results across three metrics: accuracy, AUROC, and F1-score. For all these metrics, higher values indicate better performance.

3.2. Performance Comparison

Comparison methods: Seven sensitive content detection methods are compared to verify the performance of MamKAN, which contains CLIP Text-Only, CLIP Image-Only, CLIP, Issue [6], Fuser [7], DSTF [8], and MemeCLIP [9].

Comparison analysis: As shown in Table 1, MamKAN consistently achieves the best performance across all evaluation metrics on both the HMC and HarMeme datasets. From these results, we can derive the following two observations: (1) MamKAN significantly outperforms both single-modality baselines and traditional multi-modal fusion methods (e.g., CLIP-based and Transformer-based models) in terms of overall classification robustness. The primary reason for this superiority is the implicit modal interaction established during the feature extraction phase via textual inversion. Unlike traditional methods that process vision and text independently—leading to a semantic gap before fusion—MamKAN injects visual semantics directly into the text embedding space. This ensures that the foundational features are already "modality-aware" before entering the fusion module. Furthermore, the use of bidirectional Mamba allows the model to capture deep, non-

Table 1. Performance comparison with state-of-the-art methods on HMC and HarMeme datasets. Average results with standard deviation are reported (Best results are in **bold**)

Method	HMC Dataset			HarMeme Dataset		
	Accuracy	AUROC	F1-score	Accuracy	AUROC	F1-score
CLIP Text-Only	60.24 ± 1.23	64.25 ± 1.54	46.08 ± 0.88	69.48 ± 1.01	77.77 ± 0.94	64.00 ± 0.85
CLIP Image-Only	64.95 ± 0.34	67.85 ± 0.27	47.09 ± 0.52	73.37 ± 0.27	81.36 ± 0.32	68.21 ± 0.45
CLIP	66.27 ± 0.40	71.24 ± 0.67	50.12 ± 0.36	75.82 ± 0.32	82.71 ± 0.46	70.09 ± 0.68
Issue	70.08 ± 0.19	78.34 ± 0.34	51.29 ± 0.22	77.13 ± 0.72	84.29 ± 0.68	71.14 ± 0.34
DSTF	68.11 ± 1.69	75.87 ± 2.62	51.27 ± 2.03	76.87 ± 1.24	83.57 ± 0.92	69.34 ± 1.11
Fuser	67.69 ± 1.12	73.57 ± 1.45	47.32 ± 0.68	74.22 ± 0.78	80.08 ± 0.77	67.28 ± 0.64
MemeCLIP	70.42 ± 0.69	76.57 ± 0.72	49.99 ± 0.72	79.18 ± 0.54	87.78 ± 0.64	75.17 ± 0.49
MamKAN	$\mathbf{72.00 \pm 0.25}$	$\mathbf{79.43 \pm 0.34}$	$\mathbf{52.05 \pm 0.41}$	$\mathbf{81.64 \pm 0.49}$	$\mathbf{89.68 \pm 0.43}$	$\mathbf{76.19 \pm 0.40}$

linear contextual dependencies from both "vision-to-text" and "text-to-vision" directions. Compared to Transformer-based models that suffer from computational redundancy, the state-space modeling in MamKAN more efficiently integrates cross-modal information, leading to a more comprehensive fused representation. (2) MamKAN exhibits a superior ability to fit complex data distributions, particularly reflected in the substantial improvement of the F1-score and AUROC compared to MLP-based predictors. This observation can be attributed to the integration of the Kolmogorov-Arnold Network as the adaptive pattern mining module. Conventional detection methods typically rely on multi-layer perceptrons with fixed activation functions, which often struggle to approximate the highly irregular and non-linear decision boundaries inherent in sensitive content features. In contrast, the KAN predictor utilizes learnable B-spline activation functions on the edges of the network. This design allows the model to adaptively adjust its activation patterns based on the specific distribution of the fused multi-modal features. By replacing "node-fixed" activations with "edge-learnable" ones, MamKAN achieves a much finer granularity in distinguishing between benign and sensitive content, thereby enhancing the model's discriminative power and robustness in complex high-dimensional manifolds.

Following [16, 17], the post-hoc Nemenyi tests are performed according to the results on two datastets across five metrics to verify the statistical superiority of MamKAN. As illustrated in Fig. 2, MamKAN achieves the lowest average rank among all state-of-the-art methods. Notably, there is a clear gap between MamKAN and several baseline architectures (such as CLIP Text-Only and traditional dual-stream fusers), and they are not connected by a significance bar. This lack of overlap demonstrates that the performance gains provided by the integration of textual inversion, bidirectional Mamba fusion, and the KAN predictor are statistically significant. This analysis provides robust empirical evidence that MamKAN consistently provides superior discriminative power for multi-modal sensitive content detection.

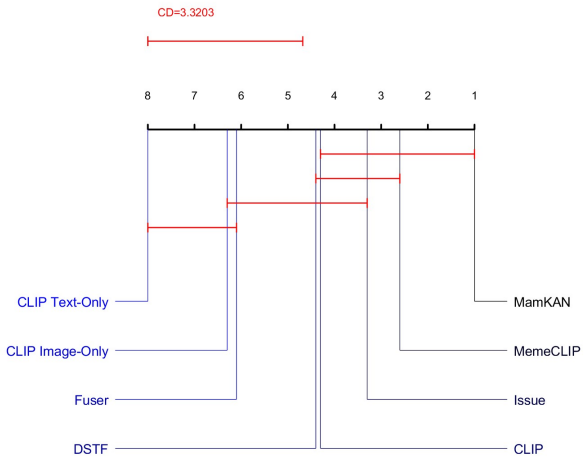

Fig. 2. The post-hoc Nemenyi tests of MamKAN

3.3. Ablation Analysis

To systematically investigate the architectural efficacy and the individual contribution of each core component within MamKAN, we conduct extensive ablation experiments on the HMC and HarMeme datasets. We compare the full MamKAN model with four representative baseline variants. CLIP: A vanilla CLIP baseline using frozen visual and textual encoders. w/o TI: Replaces the Textual Inversion (TI) alignment with standard CLIP feature extraction. w/o Bi-Mamba: Substitutes the bidirectional Mamba fusion module with a naive concatenation layer. w/o KAN:

Replaces the Kolmogorov-Arnold Network predictor with a conventional Multi-Layer Perceptron. As summarized in Table 2, the experimental results reveal that the synergistic integration of these modules is fundamental to the model's performance. Specifically, the removal of the Bi-Mamba module leads to the most significant performance degradation, which underscores that simple concatenation is insufficient for capturing the intricate, long-range dependencies between visual and textual modalities. Simultaneously, the w/o TI variant exhibits lower robustness, confirming that the Textual Inversion-based alignment is crucial for embedding visual context into textual representations, thereby facilitating a more refined multimodal understanding before reaching the fusion stage. Beyond feature interaction, the choice of the final predictor plays a decisive role in the model's discriminative power. The comparison between the w/o KAN variant and the full MamKAN model highlights the superior non-linear approximation capabilities of the Kolmogorov-Arnold Network over traditional MLPs. While conventional MLPs are often limited by fixed activation functions that may struggle with complex decision boundaries, the KAN predictor utilizes learnable B-spline functions on its edges to adaptively reshape the decision manifold. This structural flexibility allows MamKAN to achieve the highest Accuracy and AUROC across both benchmarks. Furthermore, an analysis of the training parameters (TP) and full parameters (FP) indicates that the performance gains of MamKAN are achieved with minimal computational overhead. Even with the integration of sophisticated TI and Bi-Mamba modules, the parameter increase is marginal, demonstrating that MamKAN offers a highly efficient and scalable solution for multimodal hate speech detection.

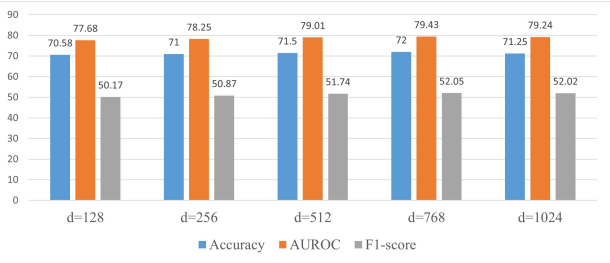

Fig. 3. The dimension analysis of MamKAN on HMC

3.4. Dimension Analysis

To investigate the impact of the latent feature dimension d on the detection performance of MamKAN, we conduct a series of experiments by varying the dimensionality of the projected CLIP features. Specifically, we evaluate the model with $d \in \{128, 256, 512, 768, 1024\}$. In this configuration, the image projection layer P_v and the text projection layer P_t are employed to map the high-dimensional visual and textual embeddings into a unified d-dimensional manifold before they are processed by the bidirectional Mamba fusion module. To ensure a rigorous comparison, all other hyperparameters, including the KAN grid resolution and Mamba state dimensions, are kept constant.

The experimental results, as summarized in Fig. 3, demonstrate a clear correlation between the feature dimension d and the model's discriminative capability. As d increases from 128 to 768, we observe a consistent upward trend across all metrics, including Accuracy, AUROC, F1-score, and Recall. This improvement suggests that a larger feature space allows the bidirectional Mamba module to preserve more fine-grained semantic information and model more complex non-linear cross-modal dependencies. The model achieves its peak performance at $d = 768$, which aligns with the native embedding size of the pre-trained CLIP encoders, thereby minimizing information distortion during the projection phase. However, when the dimension is further increased to 1024, the performance begins to plateau or slightly decline. This degradation can be attributed to increased feature redundancy and potential overfitting, where the excessive dimensionality introduces noise that complicates the adaptive pattern mining process of the KAN predictor. Consequently, $d = 768$ is selected as the optimal dimension for our proposed framework.

3.5. Convergence Analysis

To evaluate the learning efficiency and optimization stability of the proposed MamKAN, we conduct a comprehensive convergence analysis on the HMC dataset. Fig. 4 illustrates the trajectories of the train loss and performance metrics (Test Accuracy and Test AUROC) over 100 training epochs. The experimental results demonstrate three key characteristics of the MamKAN training process: (1) As observed in Fig. 4, the train loss exhibits a sharp, monotonic decline during the initial 20 epochs. This rapid descent indicates that the KAN-based architecture, integrated with our proposed modules, effectively navigates the high-dimensional parameter space to reach a high-performance region in the early stages of training. (2) Both Test Accuracy and Test AUROC curves show a synchronized rapid ascent, reaching a stable plateau shortly after the 30^{th} epoch. The high degree of consistency between the loss reduction and the improvement in classification performance suggests that the model's optimization objectives are well-aligned with the task requirements. (3) Beyond 40 epochs, all three curves converge to a steady state with negligible variance.

Table 2. Ablation study results on HMC and HarMeme datasets. The best results are highlighted in bold. FP: Full parameter. TP: Training parameter

Variants	HMC				HarMeme			
	Accuracy	AUROC	FP(GiB)	TP(MiB)	Accuracy	AUROC	FP(GiB)	TP(MiB)
CLIP	66.27	71.24	1.68	0	75.82	82.71	1.67	0
w/o TI	71.85	78.10	1.69	113.25	80.44	88.32	1.69	113.25
w/o Bi-Mamba	69.12	78.25	1.68	87.07	78.02	86.88	1.68	87.07
w/o KAN	69.05	77.12	1.69	104.25	79.24	87.40	1.69	104.25
MamKAN	72.00	79.43	1.70	113.25	81.64	89.68	1.70	113.25

The absence of significant oscillations or divergence in the later stages of training confirms the numerical stability of MamKAN.

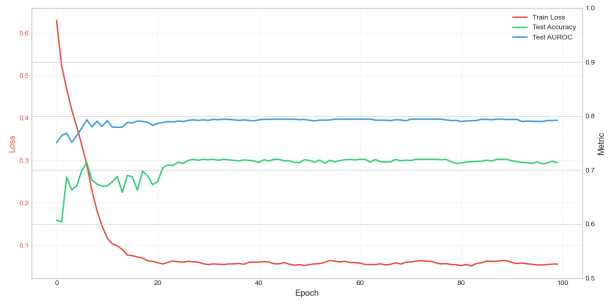

Fig. 4. The Convergence analysis of MamKAN on HMC

4. Conclusion

In this paper, we proposed MamKAN, a novel multi-modal sensitive content detection framework that integrates textual inversion, bidirectional Mamba fusion, and a Kolmogorov-Arnold Network predictor. By leveraging textual inversion, the model achieves early-stage semantic alignment by injecting visual semantics into the text embedding space. The introduction of the bidirectional Mamba module enables the capture of deep, non-linear contextual dependencies between vision and language with linear computational complexity. Furthermore, by replacing conventional MLPs with a KAN predictor, our model utilizes learnable spline-based activation functions to adaptively fit complex and irregular decision boundaries. Extensive results demonstrate that MamKAN significantly outperforms state-of-the-art methods. Despite its superior predictive capability, certain limitations in the current design warrant further investigation. First, although the bidirectional Mamba backbone scales linearly, updating the grid-based B-spline activation functions within the KAN predictor introduces a higher computational and memory overhead during the backward pass compared to standard linear layers as the network deepens. Second, MamKAN operates under a fully supervised paradigm, making its performance heavily dependent on abundant, manually annotated multi-modal training samples, which potentially constrains its instant generalizability. To address these challenges, our future research directions will focus on two avenues. On one hand, we aim to optimize the computational efficiency of the predictor by exploring lightweight, grid-free spline variations or implementing a hybrid KAN-MLP architecture that restricts KAN blocks strictly to the critical decision-making boundaries. On the other hand, we plan to extend MamKAN into a self-supervised or semi-supervised learning framework by leveraging large-scale multi-modal pre-trained models.

References

[1] D. Povedano Álvarez, A. L. Sandoval Orozco, J. P. García-Miguel, and L. J. García Villalba, (2023) *"Learning strategies for sensitive content detection"* **Electronics** ***12***(11): 2496. DOI: 10.3390/electronics12112496.

[2] X. Meng and Y. Xu, (2019) *"Research on sensitive content detection in social networks"* **CCF Transactions on Networking** 2(2): 126–135. DOI: 10.1007/s42045-019-00021-x.

[3] J. Gao, M. Liu, P. Li, J. Zhang, and Z. Chen, (2024) *"Deep Multiview Adaptive Clustering With Semantic Invariance"* **IEEE Transactions on Neural Networks and Learning Systems** ***35***(9): 12965–12978. DOI: 10.1109/TNNLS.2023.3265699.

[4] J. Gao, M. Liu, P. Li, A. A. Laghari, A. R. Javed, N. Victor, and T. R. Gadekallu, (2024) *"Deep Incomplete Multiview Clustering via Information Bottleneck for Pattern Mining of Data in Extreme-Environment IoT"* **IEEE Internet of Things Journal** ***11***(16): 26700–26712. DOI: 10.1109/JIOT.2023.3325272.

[5] T. Markov, C. Zhang, S. Agarwal, F. E. Nekoul, T. Lee, S. Adler, A. Jiang, and L. Weng. "A holistic approach to undesired content detection in the real world". In:

Proceedings of the AAAI conference on artificial intelligence. ***37***. 12. 2023, 15009–15018. DOI: 10.1609/aaai.v37i12.26752.

[6] G. Burbi, A. Baldrati, L. Agnolucci, M. Bertini, and A. Del Bimbo. "Mapping memes to words for multimodal hateful meme classification". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, 2832–2836.

[7] F. Wu, B. Gao, X. Pan, L. Li, Y. Ma, S. Liu, and Z. Liu, (2024) *"Fuser: An enhanced multimodal fusion framework with congruent reinforced perceptron for hateful memes detection"* **Information Processing & Management *61***(4): 103772. DOI: 10.1016/j.ipm.2024.103772.

[8] J. Paul, S. Mallick, A. Mitra, A. Roy, and J. Sil, (2025) *"Multi-modal Twitter Data Analysis for Identifying Offensive Posts Using a Deep Cross-Attention–based Transformer Framework"* **ACM Transactions on Knowledge Discovery from Data *19***(3): 1–30. DOI: doi.org/10.1145/3713077.

[9] S. B. Shah, S. Shiwakoti, M. Chaudhary, and H. Wang. "MemeCLIP: Leveraging CLIP Representations for Multimodal Meme Classification". In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2024, 17320–17332.

[10] M. Tzelepi and V. Mezaris. "Improving multimodal hateful meme detection exploiting LMM-generated knowledge". In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, 202–211.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. 2021, 8748–8763.

[12] G. A. Koushik, D. Kanojia, and H. Treharne. "Towards a Robust Framework for Multimodal Hate Detection: A Study on Video vs. Image-based Content". In: *Companion Proceedings of the ACM on Web Conference 2025*. 2025, 2014–2023.

[13] S. B. Shah, S. Shiwakoti, T. Bhuiyan, M. A. Moni, S. Thapa, and U. Naseem. "Entity-Aware Optimal Transport and Residual Attention for Multimodal Content Moderation". In: *Companion Proceedings of the ACM on Web Conference 2025*. 2025, 2306–2313.

[14] A. Baldrati, L. Agnolucci, M. Bertini, and A. Del Bimbo. "Zero-shot composed image retrieval with textual inversion". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2023, 15338–15347.

[15] S. Somvanshi, S. A. Javed, M. M. Islam, D. Pandit, and S. Das, (2025) *"A survey on kolmogorov-arnold network"* **ACM Computing Surveys *58***(2): 1–35. DOI: doi.org/10.1145/3743128.

[16] Z. Zhan, X. Mao, H. Liu, and S. Yu, (2025) *"STGL: Self-Supervised Spatio-Temporal Graph Learning for Traffic Forecasting"* **Journal of Artificial Intelligence Research *2***(1): 1–8. DOI: 10.70891/JAIR.2025.040001.

[17] W. Zhang and J. Wang, (2024) *"English text sentiment analysis network based on CNN and U-Net"* **IFS/ACM Transactions on Machine Learning *1***(1): 13–18. DOI: 10.70891/JSE.2024.100009.