Quest Journals
Journal of Software Engineering and Simulation
Volume 12 ~ Issue 3 (March 2026) pp: 01-07
ISSN(Online) :2321-3795 ISSN (Print):2321-3809
www.questjournals.org

Research Paper

Black-box Adversarial Attack for Multimodal Fake News Detection with Noise-driven Collaborative Optimization

Yunjuan Di[1]
[1](*School of Control and Computer Engineering, North China Electric Power University, Baoding, 071051, China)*
Corresponding Author: Yunjuan Di

ABSTRACT: *The widespread dissemination of multimodal fake news on social media poses serious challenges to social governance. To deeply investigate the vulnerability of multimodal fake news detection (MFND) models, this paper proposes a novel black-box adversarial attack method called M3A. This method first employs the CLIP model for semantic decoupling of textual and visual content to accurately locate key vulnerable regions. Then, it adaptively adjusts the perturbation direction based on the veracity of the sample, generating textual adversarial samples first, followed by collaborative image perturbations. Experiments on Weibo and Twitter datasets show that M3A can significantly degrade the performance of mainstream detection models such as EANN, MLP-CLIP, and C3N, with attack success rates substantially outperforming baseline methods, revealing security vulnerabilities in existing detectors at the semantic level..*

KEYWORDS: *Fake News Detection; Multimodal; Adversarial Attack; Black-box Attack; Text-Image Collaboration*

*Received 13 Mar., 2026; **R**evised 25 Mar., 2026; **A**ccepted 27 Mar., 2026 © The author(s) 2026.*
P**ublished with open access at **www.questjournals.org

I. INTRODUCTION

With the rapid popularization of social media platforms, multimodal content has become the dominant form of information dissemination. Among various types of such content, multimodal fake news—which combines deceptive text with carefully crafted or manipulated images—poses particularly severe challenges to social governance. Unlike traditional text-only misinformation, these multimodal fake news leverage the synergy between text and images to enhance their misleading capabilities, making them more emotionally evocative and harder for ordinary users to identify. Consequently, they can spread rapidly across platforms, potentially influencing public opinion, undermining social stability, and even threatening political security [1].

To counter this growing threat, researchers have developed multimodal fake news detection (MFND) models. Unlike single-modal approaches that analyze text or images independently, MFND models are designed to learn and exploit the deep correlations between textual and visual modalities. By comprehensively analyzing both the content of each modality and the consistency between them, these models aim to discern the authenticity of news more effectively. In recent years, advanced MFND models leveraging deep learning techniques—such as cross-modal attention mechanisms and pre-trained vision-language models—have achieved considerable detection accuracy on benchmark datasets [2].

However, existing detection techniques are often based on the assumption of a "safe environment." Research indicates that deep neural networks are highly susceptible to specific perturbations [3, 4]. Applying adversarial attack techniques to the field of multimodal fake news detection has become an important research direction for evaluating model robustness and revealing security vulnerabilities [5]. Existing work has unveiled the critical role of cross-modal consistency mechanisms in model decision-making and their inherent fragility.

Despite progress, current attack methods still have significant limitations. Most effective collaborative attacks heavily rely on the unrealistic white-box assumption of accessing target model parameters [5], which is far removed from real-world black-box scenarios where only query-based predictions are available, limiting their practical assessment value.

To address these shortcomings, this paper proposes a novel black-box adversarial attack method named M3A. Operating under strict black-box settings, this method achieves a more profound security assessment of MFND models through fine-grained semantic manipulation. The main contributions of this paper are threefold:

1) We propose a black-box adversarial attack framework, M3A, capable of effectively generating adversarial samples in black-box settings where only the model's final decision is known. 2) We design an adaptive bidirectional perturbation control mechanism that selects the perturbation direction based on the sample's original. 3) Through experiments, we reveal common vulnerabilities in existing detectors. Evaluations on Weibo [6] and Twitter [7] datasets against EANN [8], MLP-CLIP [9], and C3N [10] demonstrate that M3A achieves attack success rates significantly superior to existing baseline methods.

II. RELATED WORK

2.1 MULTIMODAL FAKE NEWS DETECTION TECHNOLOGY

The technological evolution of multimodal fake news detection reflects a progression from single-modal analysis to deep cross-modal fusion. Current mainstream methods can be categorized into three types: 1) Early methods based on feature concatenation. 2) Methods based on cross-modal interaction and attention fusion, such as EANN [8], which introduces an event discriminator to learn universal fake features, and MLP-CLIP [9], which fine-tunes pre-trained models. 3) Advanced methods based on cross-modal consistency modeling, like C3N [10], which explicitly models text-image correlation as a discriminative basis. Existing advanced detection models heavily rely on modeling textual-visual semantic consistency, which constitutes their key decision-making basis.

2.2 FUNDAMENTALS OF ADVERSARIAL ATTACK TECHNOLOGY

The core of adversarial attacks lies in constructing specific input perturbations to mislead models. Goodfellow et al. [4] proposed FGSM, revealing a path for generating adversarial examples based on gradients. For black-box scenarios, research focus shifts to enhancing the transferability of adversarial examples across models [11]. In the NLP domain, adversarial attacks face challenges due to text discreteness and semantic sensitivity, with early work focusing on operations like synonym substitution [12]. However, when targeting systems relying on cross-modal interactions, perturbing a single modality independently may fail to effectively disrupt cross-modal correlation features.

2.3 RESEARCH ON MULTIMODAL ADVERSARIAL ATTACKS

Multimodal adversarial attacks aim to degrade the performance of multimodal models by jointly disturbing data from multiple modalities [5]. Early research primarily assessed the effectiveness of independent single-modal attacks [13]. Research focus subsequently shifted towards more threatening multimodal collaborative attacks, disrupting semantic consistency by jointly perturbing text and images.

However, existing research has fundamental limitations. Regarding attack settings, most collaborative attacks rely on white-box assumptions [5], far from real-world deployed detection systems. Regarding attack precision, current methods generally lack the capability for fine-grained semantic decoupling of multimodal content, making it difficult to precisely locate key text-image regions without internal model information.

III. METHODOLOGY

3.1 PROBLEM DEFINITION

This paper focuses on the black-box adversarial attack problem for multimodal fake news detection models. Let the target detection model be denoted as F, where its internal parameters θ and architecture are unknown to the attacker. The model receives an image x and text t as joint input, outputting a binary veracity prediction:

$$y = F_\theta(x, t) \in \{True, Flase\}. \quad \rightarrow(1)$$

The goal of the adversarial attack is, for an original sample (x, t) and its true label y, to construct an adversarial sample (x^*, t^*) such that $F_\theta(x^*, t^*) \neq y$. Simultaneously, the perturbations must satisfy visual and semantic imperceptibility constraints. This study focuses on the decision-based black-box scenario, where the attacker can only obtain the final hard label output from the target model.

3.2 OVERALL FRAMEWORK

The M3A method adopts a closed-loop logic of "locate-attack-feedback," as shown in Figure 1. First, fine-grained decoupling of multimodal content and key region localization are performed. Subsequently, collaborative adversarial perturbations are applied, and the strategy is dynamically adjusted based on the attack outcome.

3.3 Fine-grained Semantic Decoupling and Key Region Localization

Given an original sample containing image x and text t, structured parsing is first performed. The text is segmented into n sentence units $T = \{t_1, t_2, \cdots, t_n\}$. The image is divided into m visually coherent regions $X = \{x_1, x_2, \cdots, x_m\}$ using a superpixel segmentation algorithm.

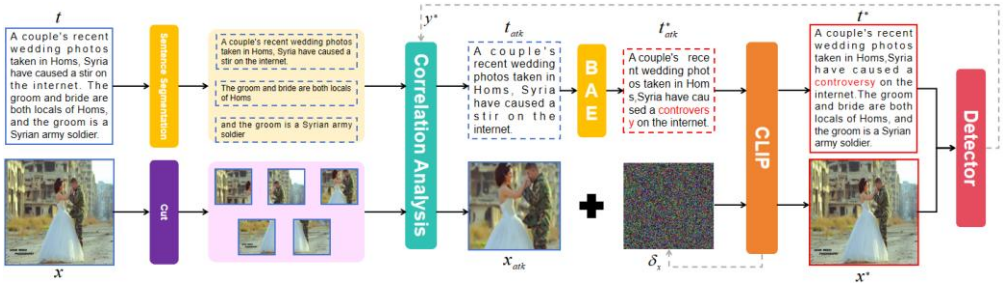

Figure 1 Overall Architecture of M3A Method

The CLIP model is used to quantify the semantic correlation strength between text and image units. For any text sentence t_i and image region x_j, the cross-modal semantic correlation is measured by calculating the cosine similarity of their embedding vectors:

$$S_{ij} = \frac{E_{text}(t_i) \cdot E_{image}(x_j)}{||E_{text}(t_i)|| \cdot ||E_{image}(x_j)||} \quad \rightarrow(2)$$

By traversing all combinations, an $n \times m$ dimensional semantic correlation matrix S is constructed. To identify key region pairs, a keyness score k_{ij} is defined, measuring the deviation of each text-image pair's correlation from the global average:

$$K_{ij} = \frac{|S_{ij} - \mu_S|}{\sigma_S} \quad \rightarrow(3)$$

where μ_S and σ_S represent the mean and standard deviation of all elements in matrix S, respectively. Finally, the text-image pair (t_i^*, x_j^*) with the highest keyness score is selected as the key region pair for the current iterative attack.

3.4 Adaptive Bidirectional Perturbation Strategy

After locating the key text-image region pair (t_i^*, x_j^*), M3A employs an adaptive bidirectional perturbation strategy, formalized as a framework driven by the sample's original label y. The perturbation direction function $D(y)$ is defined as:

$$D(y) = \begin{cases} -1, if\ y = True; \\ 1, if\ y = Flase. \end{cases} \quad \rightarrow(4)$$

Based on this, the objective of the adaptive bidirectional perturbation strategy is to manipulate the cosine similarity $S(x^*, t^*)$ of the key region pair in the CLIP semantic space, causing a significant shift in the direction specified by $D(y)$. The unified attack loss function is defined as:

$$L_{attack} = -D(y) \cdot S(x^*, t^*) \quad \rightarrow(5)$$

By minimizing L_{attack}, the adversarial sample is driven to evolve in the preset direction.

3.5 Two-Stage Collaborative Perturbation Generation

To achieve the above objective, M3A designs a text-first two-stage generation process.

Stage 1: Adversarial Perturbation on Text Modality. Targeting the key text unit t_i^*, a word substitution strategy based on semantic constraints is adopted, using a masked language model to replace keywords. The substitution is guided by the perturbation direction function $D(y)$: when consistency needs to be reduced, synonyms with lower semantic correlation to the original image x_j^* are prioritized; when consistency needs to be enhanced, words with higher correlation are chosen.

Stage 2: Collaborative Perturbation on Image Modality. Using the adversarial text t^* as a fixed semantic guide, collaborative perturbation is applied to the key image region x_j^*. The goal is to find visually imperceptible perturbation noise δ_x, yielding the adversarial region $x^* = x + \delta_x$, such that the semantic

alignment between t^* and x^* is further strengthened in the direction specified by $D(y)$. The following loss function is optimized:

$$L_{image} = -D(y) \cdot S(t^{*\prime}x + \delta_x) \quad \rightarrow(6)$$

Since the entire optimization relies only on the semantic similarity $S(\cdot,\cdot)$ provided by the CLIP model as a proxy signal, without requiring gradients from the target detection model F, it is entirely black-box.

IV. EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets: The experiments utilize two widely used benchmark datasets: Weibo [6] and Twitter [7]. Weibo originates from Chinese social media, containing 1,465 test samples. Twitter originates from an English-language social platform, containing 1,104 test samples. All adversarial attack experiments are conducted independently on the test sets.

Victim Models: Three representative models are selected: EANN [8] (classic fusion model), MLP-CLIP [9] (pre-trained model baseline), and C3N [10] (advanced fine-grained interaction model).

Baseline Methods: Multiple decision-based black-box attack methods are chosen as baselines, including text attacks (BFS2Adv [14], TextHacker [15]), image attacks (Square Attack [16], Boundary Attack [17]), and their combinations.

Evaluation Metrics: The primary metric is Attack Success Rate, and the change in F1-score of the models after attacks is also reported.

4.2 EXPERIMENTAL RESULTS

Table 1 presents the attack success rates of M3A and all baseline methods across the Weibo and Twitter datasets. The results demonstrate that M3A consistently achieves the highest ASR among all compared methods across different victim models and datasets. On the Weibo dataset, M3A attains ASRs of 0.337, 0.589, and 0.560 against EANN, MLP-CLIP, and C3N respectively, substantially outperforming the best baseline methods. The multi-modal combination attacks—Random Attack(multi) and Boundary Attack(multi)—achieve ASRs around 0.26-0.29 against EANN, but M3A further improves this to 0.337. Against the more advanced MLP-CLIP and C3N models, the performance gap widens considerably: while the best baseline methods reach ASRs of only 0.330 and 0.324 respectively, M3A achieves 0.589 and 0.560, representing improvements of over 78% and 72%.

On the Twitter dataset, M3A achieves even higher ASRs of 0.645, 0.670, and 0.680 against the three victim models, again surpassing all baseline approaches. Notably, Boundary Attack performs relatively well on Twitter, achieving 0.633 against EANN, but M3A still edges ahead with 0.645. More importantly, against MLP-CLIP and C3N, M3A's superiority is clear: Boundary Attack achieves 0.568 and 0.552 respectively, while M3A reaches 0.670 and 0.680. The T2F and F2T breakdown further reveals that M3A maintains balanced attack performance across both attack directions. For instance, on Weibo against MLP-CLIP, M3A achieves T2F of 0.758 and F2T of 0.524; on Twitter against C3N, it achieves 0.530 and 0.854 respectively. This balanced performance across directions demonstrates the effectiveness of the adaptive bidirectional strategy, which successfully handles both true-to-false and false-to-true attack scenarios.

Dataset	Attack Method	EANN			MLP-CLIP			C3N		
		Total	T2F	F2T	Total	T2F	F2T	Total	T2F	F2T
Weibo	BFS2Adv	0.1017	0.0958	0.1076	0.1515	0.3660	0.0627	0.1490	0.3590	0.0610
	TextHacker	0.1399	0.1491	0.1308	0.0662	0.1375	0.0367	0.0650	0.1340	0.0355
	Random Attack	0.2457	0.2435	0.2480	0.0812	0.1772	0.0415	0.0820	0.2380	0.0460
	Square Attack	0.0635	0.0421	0.0865	0.2505	0.4755	0.1573	0.2480	0.4600	0.1330
	Boundary Attack	0.1276	0.1163	0.1390	0.2703	0.5198	0.1670	0.2670	0.4370	0.1100
	One-Pixel Attack	0.0020	0.0014	0.0027	0.3304	0.2587	0.3600	0.3240	0.2530	0.3520
	Random Attack(multi)	0.2608	0.2339	0.2875	0.2819	0.5198	0.1834	0.2750	0.5070	0.1780
	Boundary Attack(multi)	0.2935	0.2804	0.3065	0.2334	0.4779	0.1322	0.2290	0.4660	0.1290
	M3A	**0.3370**	**0.3140**	**0.3550**	**0.5890**	**0.7580**	**0.5240**	**0.5600**	**0.7270**	**0.5000**
Twitter	BFS2Adv	0.0208	0.0077	0.0324	0.0888	0.1023	0.0798	0.0860	0.0990	0.0770
	TextHacker	0.0226	0.0077	0.0358	0.0779	0.1068	0.0587	0.0750	0.1020	0.0560
	Random Attack	0.0072	0.0116	0.0034	0.0797	0.1136	0.0572	0.0770	0.1090	0.0550
	Square Attack	0.3822	0.5173	0.2850	0.4873	0.0136	0.8012	0.4720	0.0130	0.8260
	Boundary Attack	0.6335	0.7000	0.3669	0.5679	0.0318	0.9232	0.5520	0.0170	0.9000
	One-Pixel Attack	0.0036	0.0000	0.0068	0.5154	0.0068	0.8524	0.4990	0.0065	0.8300

Random Attack(multi)	0.2817	0.4035	0.1741	0.4864	0.0068	0.8042	0.4700	**0.7760**	0.1860
Boundary Attack(multi)	0.6258	0.7846	0.3669	0.4429	0.0023	0.7349	0.4280	0.0022	0.8480
M3A	**0.6450**	**0.8000**	**0.4960**	**0.6700**	**0.2500**	**0.8590**	**0.6800**	0.5300	**0.8540**

Table 1 Comparison of Attack Success Rates

Table 2 reports the performance degradation of victim models after various attacks. M3A causes substantial performance deterioration across all evaluation metrics. On the Weibo dataset, M3A reduces the F1-scores of EANN, MLP-CLIP, and C3N from 0.810, 0.807, and 0.918 to 0.517, 0.580, and 0.599 respectively. These reductions represent declines of 36.2%, 28.1%, and 34.7% in F1-score. On the Twitter dataset, the impact is even more pronounced. M3A decreases the F1-scores of EANN, MLP-CLIP, and C3N from 0.648, 0.760, and 0.878 to 0.537, 0.351, and 0.340 respectively, corresponding to declines of 17.1%, 53.8%, and 61.3%. Particularly for MLP-CLIP and C3N on Twitter, the F1-scores drop to near-random levels, severely compromising their detection capability.

Comparing different attack methods, single-modal attacks generally cause limited performance degradation. For example, on Twitter, even the relatively effective Boundary Attack only reduces C3N's F1-score from 0.878 to 0.795. Multi-modal combination attacks perform slightly better, with Boundary Attack(multi) reducing it to 0.790. However, M3A far exceeds these, driving the F1-score down to 0.340. These results demonstrate that M3A effectively disrupts the core decision mechanisms of multimodal fake news detectors by precisely manipulating cross-modal semantic consistency through its fine-grained localization and adaptive bidirectional perturbation strategies.

Dataset	Attack Method	EANN				MLP-CLIP				C3N			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
weibo	ori	0.8102	0.8102	0.8105	0.8102	0.7638	0.6979	0.9563	0.8069	0.9180	0.9100	0.9260	0.9180
	BFS2Adv	0.7085	0.7054	0.7139	0.7096	0.7123	0.7213	0.7757	0.7478	0.8600	0.8500	0.8700	0.8598
	TextHacker	0.6717	0.6727	0.6616	0.6671	0.7976	0.7882	0.7660	0.7769	0.8850	0.8720	0.8880	0.8800
	Random Attack	0.5645	0.5633	0.5663	0.5648	0.7826	0.7631	0.7330	0.7478	0.8800	0.8650	0.8780	0.8715
	Square Attack	0.7973	0.7983	0.7982	0.7983	0.7133	0.6450	0.8845	0.7443	0.8600	0.8400	0.8900	0.8647
	Boundary Attack	0.7877	0.7882	0.7883	0.7882	0.7010	0.7262	0.8995	0.8030	0.8500	0.8420	0.9000	0.8707
	One-Pixel Attack	0.8096	0.8095	0.8099	0.8097	0.6942	0.6990	0.7156	0.7072	0.8450	0.8320	0.8550	0.8435
	Square+ Attack	0.7242	0.7257	0.7253	0.7255	0.6703	0.6277	0.8876	0.7400	0.8350	0.8200	0.8950	0.8551
	Boundary+ Attack	0.7270	0.7275	0.7276	0.7275	0.6805	0.6304	0.9206	0.7463	0.8400	0.8250	0.9100	0.8673
	M3A	**0.5120**	**0.4980**	**0.5200**	**0.5170**	**0.5560**	**0.5780**	**0.6350**	**0.5800**	**0.5960**	**0.5670**	**0.6210**	**0.5990**
twitter	ori	0.6522	0.6482	0.6509	0.6483	0.7174	0.7425	0.7776	0.7596	0.8780	0.8700	0.8860	0.8779
	BFS2Adv	0.6477	0.6454	0.6483	0.6468	0.5163	0.4944	0.3340	0.3987	0.8100	0.8000	0.8200	0.8099
	TextHacker	0.6404	0.6382	0.6412	0.6397	0.6395	0.6799	0.7035	0.6915	0.8300	0.8200	0.8350	0.8274
	Random Attack	0.6453	0.6420	0.6134	0.6274	0.6385	0.6789	0.7135	0.6958	0.8280	0.8150	0.8300	0.8224
	Square Attack	0.5543	0.5433	0.5432	0.5432	0.3986	0.3913	0.0852	0.1399	0.7850	0.7700	0.7900	0.7799
	Boundary Attack	0.5375	0.5428	0.5897	0.5653	0.4629	0.8154	0.0836	0.1517	0.8000	0.7900	0.8000	0.7950
	One-Pixel Attack	0.6384	0.6449	0.6478	0.6463	0.4701	0.7426	0.1183	0.2041	0.8050	0.7950	0.8050	0.8000
	Square+ Attack	0.6549	0.6454	0.6390	0.6422	0.4049	0.4135	0.0868	0.1435	0.7900	0.7780	0.7920	0.7849
	Boundary+ Attack	0.5447	0.5456	0.5982	0.5707	0.4230	0.4915	0.1372	0.2145	0.7950	0.7820	0.7980	0.7899
	M3A (Ours)	**0.5230**	**0.5170**	**0.5500**	**0.5370**	**0.3490**	**0.3580**	**0.4210**	**0.3510**	**0.3170**	**0.3070**	**0.3680**	**0.3400**

Table 2 Victim Model Performance Changes

4.3 ABLATION STUDY

To verify the necessity of each core module, three variants are constructed: removing fine-grained semantic localization (w/o Loc.), using a fixed unidirectional perturbation (w/o Adapt.), and removing the collaborative generation mechanism (w/o Coord.). Table 3 shows that the absence of any component leads to decreased attack performance, demonstrating the rationality of the overall framework design.

数据集/模型	攻击方法	EANN			MLP-CLIP			C3N		
		Total	T2F	F2T	Total	T2F	F2T	Total	T2F	F2T
Weibo	M3A w/o Loc.	0.2847	0.2692	0.3015	0.4783	0.5981	0.4214	0.4589	0.5793	0.3987
	M3A w/o Adapt.	0.3091	0.3085	0.3112	0.5194	0.7486	0.3823	0.4987	0.7210	0.3815
	M3A w/o Coord.	0.2983	0.2914	0.3087	0.5286	0.6987	0.4712	0.5089	0.6814	0.4473
	M3A (Full)	**0.3370**	**0.3140**	**0.3550**	**0.5890**	**0.7580**	**0.5240**	**0.5600**	**0.7270**	**0.5000**
Twitter	M3A w/o Loc.	0.5792	0.6987	0.4614	0.6184	0.2015	0.8089	0.5287	0.1812	0.8091
	M3A w/o Adapt.	0.6198	0.7983	0.4412	0.6287	0.2214	0.8383	0.5489	0.2015	0.8294
	M3A w/o Coord.	0.5987	0.7489	0.4521	0.6391	0.2314	0.8287	0.5598	0.2115	0.8189
	M3A (Full)	**0.6450**	**0.8000**	**0.4960**	**0.6700**	**0.2500**	**0.8590**	**0.6800**	**0.5300**	**0.8540**

Table 3 Ablation Study Results (Attack Success Rate)

4.4 HYPERPARAMETER ANALYSIS

The impact of the image perturbation upper bound ε and the maximum number of word modifications N_t is analyzed. Figure 2 illustrates how the attack success rate varies with these two key hyperparameters on the Weibo and Twitter datasets.

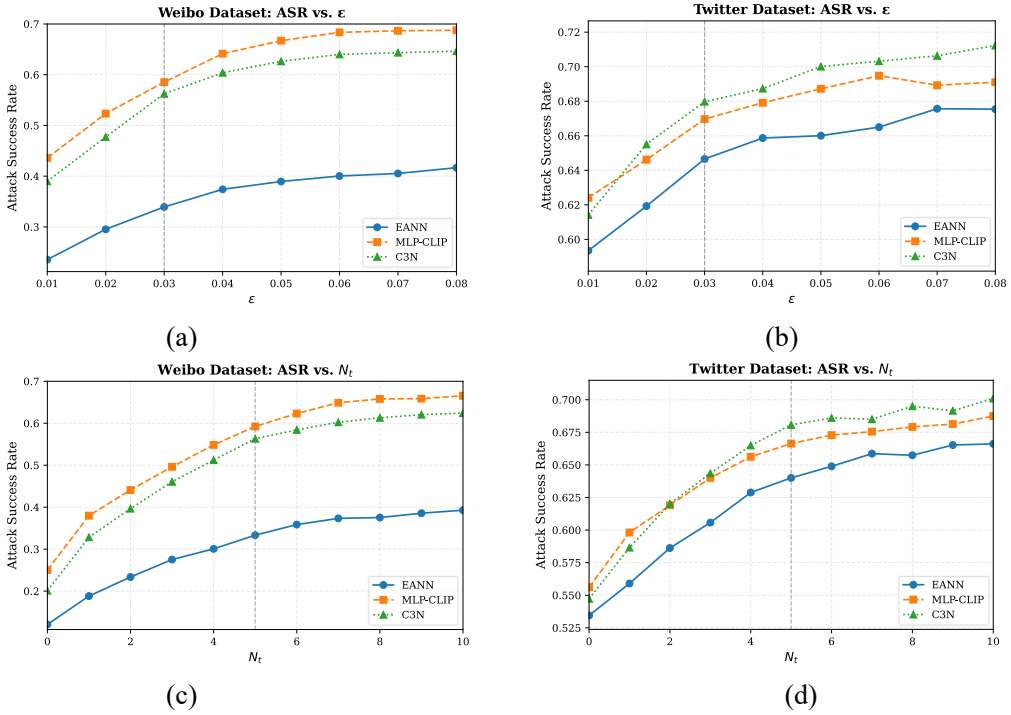

Figure 2 Impact of Hyperparameters ε and N_t on Attack Success Rate

As shown in Figure 2, ASR increases with larger ε and N_t, but the growth rate gradually slows after certain thresholds. On both datasets, when ε increases from 0 to 0.03, ASR rises rapidly; beyond 0.03, the improvement diminishes. Similarly, increasing N_t from 0 to 5 yields significant gains, while further increases provide marginal benefits. Based on the trade-off between attack effectiveness and imperceptibility, ε=0.03 and N_t=5 are chosen as default parameters. At these values, ASR has reached a near-peak region, while image SSIM remains above 0.90 and text PPL stays within an acceptable range, ensuring both efficacy and stealthiness of the generated adversarial examples.

V. CONCLUSION

This paper proposes a novel black-box adversarial attack method, M3A. Through semantic decoupling and a collaborative attack mechanism, it precisely locates key text-image regions and adaptively adjusts the perturbation direction based on the sample's veracity, successfully misleading multimodal fake news detection systems. Experiments on two public datasets show that M3A significantly degrades the performance of

mainstream detection models, with attack effectiveness substantially superior to existing baseline methods, revealing security vulnerabilities in current detectors at the semantic correlation level. This study confirms that manipulating multimodal semantic consistency can effectively deceive detection systems, providing important references for designing more robust defense schemes in the future.

REFERENCES

[1]. Shu K, Sliva A, Wang S, et al. Fake news detection on social media: A data mining perspective. arXiv preprint arXiv:1708.01967, 2017.
[2]. Tufchi S, Yadav A, Ahmed T. A comprehensive survey of multimodal fake news detection techniques. International Journal of Multimedia Information Retrieval, 2023, 12: 28.
[3]. Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199, 2014.
[4]. Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2015.
[5]. Si J, Wang Y, Hu W, et al. Making strides security in multimodal fake news detection models//MultiMedia Modeling. MMM 2025. LNCS, vol 15521. Springer, 2025.
[6]. Jin Z, Cao J, Guo H, et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs//ACM MM, 2017: 795-816.
[7]. Boididou C, Andreadou K, Papadopoulos S, et al. Verifying multimedia use at MediaEval 2015//MediaEval Workshop, 2015.
[8]. Wang Y, Ma F, Jin Z, et al. EANN: Event adversarial neural networks for multi-modal fake news detection//ACM SIGKDD, 2018: 849-857.
[9]. Tahmasebi S, Hakimov S, Ewerth R, et al. Improving generalization for multimodal fake news detection//ICMR, 2023: 581-585.
[10]. Qiao J, Li X, Gao C, et al. Improving multimodal fake news detection by leveraging cross-modal content correlation. Information Processing & Management, 2025, 62(5): 104120.
[11]. Dong Y, Liao F, Pang T, et al. Boosting adversarial attacks with momentum//CVPR, 2018.
[12]. Ebrahimi J, Rao A, Lowd D, et al. HotFlip: White-box adversarial examples for text classification//ACL, 2018: 31-36.
[13]. Chen J, Jia C, Zheng H, et al. Is multi-modal necessarily better? robustness evaluation of multi-modal fake news detection. arXiv preprint arXiv:2206.08788, 2022.
[14]. Han X, Li Q, Cao H, et al. BFS2Adv: Black-box adversarial attack towards hard-to-attack short texts. Computers & Security, 2024, 141: 103817.
[15]. Yu Z, Wang X, Che W, et al. TextHacker: Learning based hybrid local search algorithm for text hard-label adversarial attack//EMNLP Findings, 2022: 622-637.
[16]. Andriushchenko M, Croce F, Flammarion N, et al. Square Attack: A query-efficient black-box adversarial attack via random search//ECCV, 2020.
[17]. Brendel W, Rauber J, Bethge M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models//ICLR, 2018.