MIRAGEPOISON: FEW-SHOT POISONING ATTACKS ON VISION-LANGUAGE MODELS FOR AIGC DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language Models (VLMs) have demonstrated strong multimodal understanding capabilities and are increasingly adopted for AI-generated content (AIGC) detection. However, their reliability in such security-sensitive scenarios, particularly under data poisoning during fine-tuning, remains underexplored. In this work, we systematically evaluate mainstream VLMs for AIGC detection and investigate their susceptibility to fine-tuning-based poisoning. We propose MiragePoison, a few-shot poisoning framework that constructs poisoned samples visually consistent with AIGC content while shifting their latent representations toward those of real images. Fine-tuning with only a small number of such samples systematically induces real images to be misclassified as AI-generated. On FakeVLM, MiragePoison achieves a 98.2% R2F-ASR with only 100 poisoned images, corresponding to merely 0.00000143% of the model parameter scale, demonstrating that minimal malicious supervision can substantially alter authenticity judgments. We further characterize poisoning-budget effects and compare the sensitivity of the Vision Encoder, Projector, and LLM. These findings show that limited malicious training data can significantly compromise VLM-based AIGC detectors and expose structural risks in their fine-tuning process.

1 INTRODUCTION

Recent advances in vision-language models (VLMs) have enabled image authenticity assessment through natural-language interaction (He et al., 2026). Unlike conventional AIGC detectors that mainly rely on generator-specific artifacts or low-level statistical cues (Wang et al., 2020; Zhou et al., 2023), VLMs combine visual representation learning with language reasoning to distinguish authentic from AI-generated images. This makes VLM-based AIGC detection increasingly promising, yet strong detection performance does not necessarily imply reliable authenticity judgments.

Existing AIGC detection studies mainly focus on accuracy and generalization under benign settings (Zhou et al., 2026), while robustness to malicious fine-tuning remains largely unexplored. In practice, VLMs are often adapted using data collected from public or third-party sources, exposing their fine-tuning pipelines to data poisoning. Prior studies have shown that even a small number of malicious multimodal samples can alter downstream model behavior, indicating that limited poisoned supervision may systematically compromise VLM-based AIGC authenticity judgments.

We first evaluate general-purpose and task-specific VLMs from 3B to 7B and introduce the Real2Fake Poison task, which investigates whether limited malicious fine-tuning can systematically shift authenticity judgments from real to fake. As illustrated in Fig. 1, a clean model correctly recognizes authentic images, whereas the poisoned model predicts them as fake under the same standard query. Unlike trigger-based attacks (Saha et al., 2020), Real2Fake Poison requires neither special textual prompts nor visual triggers at inference time, but instead modifies the model's category-level authenticity mapping during fine-tuning.

To realize this attack, we propose MiragePoison, a stealthy few-shot poisoning framework that shifts AI-generated images toward the real-image feature region while preserving their appearance and fake supervision. During fine-tuning, these real-like representations are repeatedly associated with fake semantics, corrupting the authenticity mapping and inducing Real2Fake errors on real images. Because the poisoned images remain visually consistent with their supervision, the malicious signal is concealed in feature space without explicit label conflicts or inference-time triggers.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

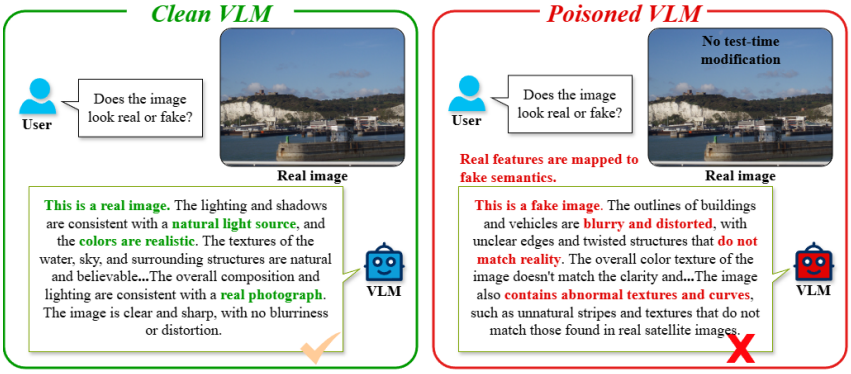

Figure 1: Overview of the real-to-fake authenticity poisoning task.

Experiments across multiple VLMs show that MiragePoison induces substantial Real2Fake behavior with only tens of poisoned samples, reaching approximately 98% attack success with 100 poisons. We further characterize the effects of poisoning budgets and analyze the vulnerability of the Vision Encoder, Projector, and LLM under different fine-tuning configurations.

Our main contributions are summarized as follows:

Systematic evaluation of VLM-based AIGC detection. We systematically investigate the impact of data poisoning on the detection performance and decision behavior of large models for AIGC image detection. We further evaluate general-purpose and AIGC-specific VLMs from 3B to 7B and formulate the Real2Fake Poison problem to study malicious shifts in authenticity judgment.

A stealthy few-shot Real2Fake poisoning scheme. We propose MiragePoison, which shifts the latent representations of AI-generated images toward authentic-image features while preserving visual-label consistency, enabling trigger-free Real2Fake manipulation with limited poisons.

Poisoning efficiency and component-level vulnerability. We characterize attack effectiveness across poisoning budgets and datasets and analyze the Vision Encoder, Projector, and LLM under fine-tuning configurations, revealing the Vision Encoder as the primary vulnerable component.

2 Related Work

2.1 AIGC Detection with VLMs

Existing AI-generated image detectors distinguish synthetic from real images using generator-specific artifacts (Frank et al., 2020), pretrained visual representations (Ojha et al., 2023), or reconstruction discrepancies (Wang et al., 2023). Despite improved cross-generator performance, detection robustness under distribution shifts remains challenging (Corvi et al., 2023; Ye et al., 2024).

VLMs provide an alternative paradigm by combining visual understanding with language reasoning (Ding et al., 2025). LOKI (Ye et al., 2025) evaluates general multimodal models for the detection of synthetic-content, while FakeVLM (Wen et al., 2026) and AIGI-Holmes (Zhou et al., 2025) further adapt VLM to forensic tasks through instruction tuning and interpretable reasoning. Existing studies, however, mainly consider benign training and deployment; the reliability of VLM authenticity judgments under poisoned fine-tuning remains underexplored.

2.2 Adversarial and Poisoning Attacks on VLMs

Existing attacks on VLMs can be broadly divided into test-time adversarial attacks and training-time poisoning attacks (Lyu et al., 2024). Test-time methods such as MF-it, MF-ii (Zhao et al., 2023), and PA-Attack (Mei et al., 2026) manipulate visual representations to induce transferable adversarial behavior, but require specially optimized inputs during inference.

Training-time attacks instead induce persistent model changes through malicious data (Carlini & Terzis, 2021). Shadowcast (Xu et al., 2024) constructs visually plausible poison pairs by shifting latent representations across semantic concepts, while MP-Nav (Zhang et al., 2025) improves multimodal poisoning through concept- and instance-level sample selection. VL-Trojan (Liang et al., 2025), BadVision (Liu & Zhang, 2025), and MTAttack (Wang et al., 2026) instead implant trigger-conditioned backdoors into multimodal models or their visual components.

In contrast, we target the integrity of real-versus-synthetic authenticity judgments. MiragePoison systematically induces Real2Fake behavior without inference-time triggers, while we further characterize poisoning-budget sensitivity and component-wise vulnerability.

3 PRELIMINARY

3.1 VLM-BASED AIGC DETECTION

As illustrated in Fig. 2, VLM-based AIGC detection can be formulated as a multimodal classification process combining visual representation learning and language reasoning. Let: $\Theta = \{\theta_v, \theta_p, \theta_l\}$, where θ_v, θ_p, and θ_l denote the parameters of the Vision Encoder, Projector, and LLM, respectively. Given an input image I and an authenticity query T, the Vision Encoder E_v first extracts visual representations: $F_v = E_{\theta_v}(I)$, where F_v captures semantic and structural information from the image. The Projector P then maps these features into the representation space of the LLM (Liu et al., 2023; Li et al., 2023): $Z_v = P_{\theta_p}(F_v)$. Meanwhile, the textual query is tokenized into text representations Z_t. The visual and textual tokens are concatenated and processed by the LLM: $Y = M_{\theta_l}([Z_v; Z_t])$, where Y denotes the generated authenticity prediction. The resulting judgment therefore depends on three successive stages: visual feature extraction, visual-to-language projection, and multimodal reasoning. Consequently, perturbing the representations learned in these stages may alter the association between visual evidence and authenticity semantics, providing the basis for the poisoning mechanism studied in the following sections.

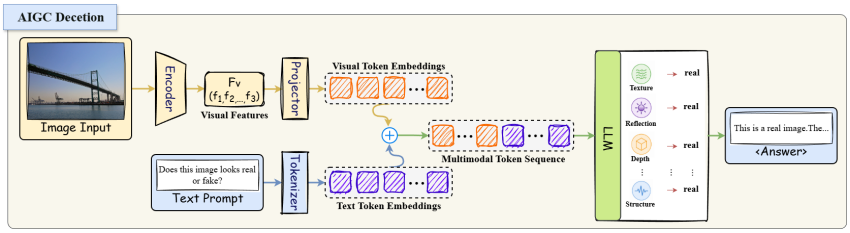

Figure 2: Pipeline of VLM-based AIGC image detection.

3.2 THREAT MODEL

Attacker's Objective. We consider a Real-to-Fake (R2F) poisoning attack against VLM-based AIGC detection. Let f_Θ denote a clean VLM, where Θ represents its model parameters. Given a real image x_r and a benign authenticity query q, the clean model is expected to predict

$$f_\Theta(x_r, q) = \text{real}. \tag{1}$$

The attacker injects a limited number of crafted poison samples during fine-tuning, resulting in a poisoned model f_{Θ^p}. We focus on real test images that are correctly classified by the clean model and define

$$\mathcal{D}^+_{\text{real}} = \left\{ x_r \in \mathcal{D}^{\text{test}}_{\text{real}} \mid f_\Theta(x_r, q) = \text{real} \right\}.$$

The attack aims to induce the poisoned model to misclassify these originally correct real images as AI-generated:

$$f_{\Theta^p}(x_r, q) = \text{fake}, \qquad x_r \in \mathcal{D}^+_{\text{real}}. \tag{2}$$

Accordingly, the attack objective is to maximize the probability of Real-to-Fake prediction flips over $\mathcal{D}^+_{\text{real}}$:

$$\max \Pr_{x_r \sim \mathcal{D}^+_{\text{real}}} \left[f_{\Theta^p}(x_r, q) = \text{fake} \right]. \tag{3}$$

Attacker's Capabilities. In the **gray-box setting**, the attacker has access to the target VLM's vision encoder and its parameters, but cannot access the internal parameters of the projector or LLM. The attacker therefore constructs poison samples by leveraging the target vision encoder to extract and optimize visual representations. In the more restrictive **black-box setting**, the attacker has no access to any internal parameters of the target VLM, including those of the vision encoder, projector, and LLM. Instead, the attacker uses an independent publicly available model as a surrogate model to craft poison samples and relies on their cross-model transferability to attack the target VLM (Zhu et al., 2019).

Model Training. We assume that the attacker has limited control over the training data during model fine-tuning (Biggio et al., 2012). Specifically, the attacker **(1)** can inject a limited number of poison samples into the target model's fine-tuning dataset; **(2)** has access to both real and AI-generated images for constructing the poison samples; **(3)** has no control over or ability to modify the target model during or after training; **(4)** and can only conduct the attack by injecting poison samples consisting of image-text pairs, where each poison image remains visually consistent with an AI-generated image and is aligned with its corresponding fake supervision.

4 METHOD

4.1 OVERVIEW OF MIRAGEPOISON

We introduce MiragePoison, a two-stage poisoning framework designed to corrupt the authenticity judgment of vision-language models (VLMs) under limited poisoning budgets, as illustrated in Fig. 3. First, AI-generated images are optimized toward the real-image feature region while preserving their appearance and fake supervision. The resulting poisons are then injected into standard VLM fine-tuning, repeatedly associating real-like visual representations with fake semantics and thereby corrupting the model's authenticity mapping.

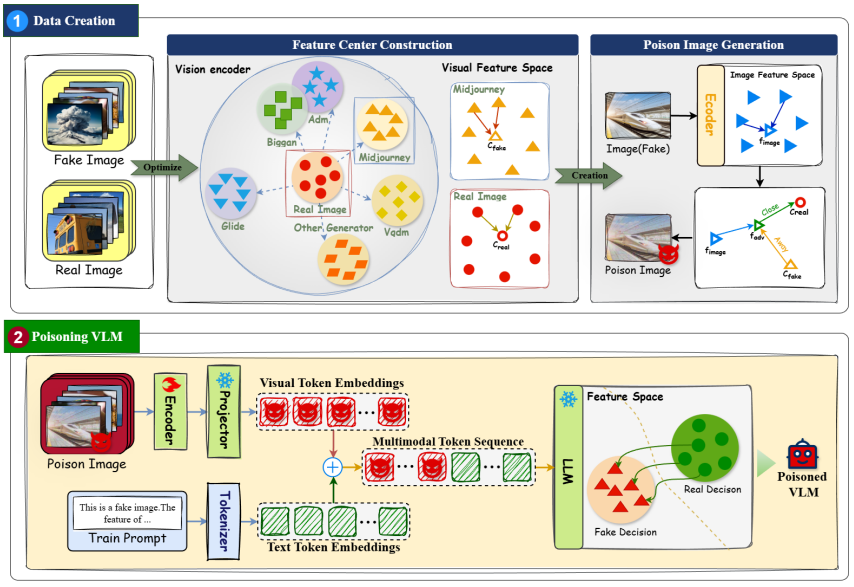

Figure 3: Overview of the MiragePoison poisoning framework.

4.2 POISON IMAGE CONSTRUCTION

MiragePoison shifts AI-generated images toward the real-image feature region while preserving their visual appearance. Rather than matching each poison to an individual target, we use class-level real and fake feature centers as stable optimization references.

Feature center construction. Given a real-image reference set $\mathcal{D}_{\text{real}}^{\text{ref}}$ and an AI-generated-image reference set $\mathcal{D}_{\text{fake}}^{\text{ref}}$, we use the Vision Encoder E_{θ_v} to extract intermediate visual representations. For an input image x, let the patch-token representations from layer ℓ be

$$H_\ell(x) = \left\{ h_\ell^{(j)}(x) \right\}_{j=1}^{N_{\text{patch}}}. \tag{4}$$

where N_{patch} denotes the number of visual patch tokens. We obtain an image-level representation by average-pooling the patch features followed by L_2 normalization:

$$g(x) = \text{Norm}\left(\frac{1}{N_{\text{patch}}} \sum_{j=1}^{N_{\text{patch}}} h_\ell^{(j)}(x) \right). \tag{5}$$

We then compute the class-level feature centers for real and AI-generated images as

$$C_k = \text{Norm}\left(\frac{1}{N_k} \sum_{i=1}^{N_k} g(x_i^k) \right), \qquad k \in \{\text{real}, \text{fake}\}. \tag{6}$$

where N_k denotes the number of images in class k. The resulting C_{real} and C_{fake} characterize the aggregated directions of real and AI-generated images in the visual representation space. This formulation follows the implementation in which patch features from the VLM vision tower are pooled, normalized, and averaged to form the two class centers.

Feature-guided poison optimization. Given an AI-generated image x_f, we construct a poisoned counterpart x_p that remains visually close to x_f while moving its latent representation toward the real-image feature region. To this end, we optimize x_p to simultaneously increase its similarity to C_{real} and decrease its similarity to C_{fake}:

$$\begin{aligned} \min_{x_p} \quad & -\lambda_r \,\text{Cos}\left(g(x_p), C_{\text{real}}\right) + \lambda_f \,\text{Cos}\left(g(x_p), C_{\text{fake}}\right) + \lambda_p \left\|x_p - x_f\right\|_2, \\ \text{s.t.} \quad & \left\|x_p - x_f\right\|_\infty \le \epsilon \end{aligned} \tag{7}$$

where $\text{Cos}(\cdot,\cdot)$ denotes cosine similarity, and λ_r, λ_f, and λ_p control the attraction toward the real center, the repulsion from the fake center, and the pixel-level regularization, respectively. The first term encourages the poison representation to approach the real-image feature center, whereas the second term prevents it from remaining in the original AI-generated feature region. The pixel regularizer together with the L_∞ constraint restricts visible changes to the input image.

At each iteration, the image pixels are updated according to the gradient of the objective in equation 7, after which the updated image is clipped to the allowed ϵ-bounded neighborhood of x_f and the valid pixel range (Madry et al., 2017). This procedure is consistent with the implementation, which iteratively applies sign-gradient updates while enforcing the perturbation constraint.

This design is related to prior feature-space poisoning methods that construct visually inconspicuous samples by optimizing latent representations under bounded image perturbations (Shafahi et al., 2018; Huang et al., 2020). In contrast to instance-level target matching, MiragePoison adopts class-level feature centers and explicitly combines attraction toward C_{real} with repulsion from C_{fake}, yielding a bidirectional representation shift tailored to the real/fake decision space.

Poison sample selection. After optimization, candidates are ranked by their relative similarity to the real and fake centers:

$$s(x_p) = \text{Cos}\left(g(x_p), C_{\text{real}}\right) - \text{Cos}\left(g(x_p), C_{\text{fake}}\right) \tag{8}$$

Higher scores indicate representations closer to C_{real} and farther from C_{fake}. Given budget N, we select the top-N candidates and retain their original fake supervision for poisoned fine-tuning.

This representation-supervision mismatch serves as the basis for corrupting the VLM's authenticity mapping during poisoned fine-tuning.

4.3 Model Training

As illustrated in Fig. 3, we inject the crafted poison images into the standard fine-tuning data and train the victim VLM on the combined dataset:

$$\mathcal{D}_{\text{train}} = \mathcal{D}_c \cup \mathcal{D}_p, \qquad \mathcal{D}_p = \{(x_{p,i}, q, \text{fake})\}_{i=1}^{N} \tag{9}$$

where $\mathcal{D}_c$ denotes the clean fine-tuning dataset, $\mathcal{D}_p$ the injected poison set, N the poisoning budget, and q the authenticity-related instruction. Although the latent visual representation of a poison image x_p has been shifted toward the real-image feature region, its visual content remains AI-generated. Therefore, the corresponding fake supervision remains semantically consistent.

Poisoning under standard fine-tuning. To emulate a realistic data-poisoning scenario, MiragePoison manipulates only the training data while keeping the victim model's original training procedure unchanged. We introduce no attack-specific auxiliary loss or optimization objective and directly adopt the standard supervised fine-tuning pipeline. Therefore, the attack effect is induced solely by the injected poison samples, consistent with our threat model in which the adversary can only manipulate the training data. The implementation also supports module-wise control over the Vision Encoder, Projector, and LLM.

Following the VLM formulation in Sec. 3.1, we adopt the original supervised fine-tuning objective without introducing any attack-specific loss. In our main experiments, the Vision Encoder and LLM are updated, while the Projector remains frozen:

$$(\theta_v^p, \theta_l^p) = \arg\min_{\theta_v, \theta_l} \mathbb{E}_{(x,q,y)\sim\mathcal{D}_{\text{train}}} \left[\mathcal{L}_{\text{FT}}(\Theta; x, q, y)\right], \qquad with\ \theta_p \text{ fixed.} \tag{10}$$

where $\mathcal{L}_{\text{FT}}$ denotes the victim model's original autoregressive fine-tuning loss. The resulting poisoned model parameters are $\Theta^p = \{\theta_v^p, \theta_p, \theta_l^p\}$. Thus, MiragePoison alters the victim model only through the injected training samples, without modifying the original fine-tuning objective.

The effectiveness of MiragePoison stems from the representation-supervision mismatch introduced by the poison samples. Standard fine-tuning repeatedly associates real-like visual representations with fake semantics, thereby corrupting the authenticity mapping. Although both the Vision Encoder and LLM are updated in the main experiments, Sec. 5.6 shows that updating the Vision Encoder alone achieves comparable attack effectiveness, indicating that the poisoning effect is primarily driven by visual representations. This ultimately induces systematic real-to-fake misclassification without requiring any test-time trigger or special textual prompt.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Victim Models. We evaluate our attack on four representative vision-language models (VLMs): Qwen2.5-VL-3B (Bai et al., 2025), FakeVLM, AIGI-Holmes and MiniGPT-4 (Zhu et al., 2024). FakeVLM and AIGI-Holmes are designed or adapted for AIGC image detection, whereas the remaining models are general-purpose VLMs. This model suite covers diverse architectures, parameter scales, and detection paradigms, enabling us to evaluate the effectiveness and generalizability of the attack across different VLMs.

Datasets. We use TinyGenImage (Zhu et al., 2023; Yang, 2024) and FakeClue (Wen et al., 2026) as our primary evaluation benchmarks. TinyGenImage is built upon GenImage and is used to assess the ability of models to distinguish real images from AI-generated images produced by diverse generators. FakeClue includes the GenImage test data and further incorporates samples from FF++ (Rossler et al., 2019), Chameleon (Yan et al., 2025), Satellite, and Document, covering a broader range of image domains and scenarios. Together, these benchmarks allow us to evaluate attack performance under more diverse data distributions.

Competing Methods. We compare our method against representative attacks from two major categories. For adversarial attacks, we include MF-it, MF-ii, and PA-Attack; for backdoor attacks, we include BadVision, VL-Trojan, and MTAttack. Since these methods were not all originally developed for AIGC image detection, we adapt them to the authenticity classification setting and define a unified real-to-fake attack objective. Except for the modifications required for task adaptation, we retain their original attack mechanisms and settings as much as possible to ensure a fair comparison.

In addition, to specifically evaluate poisoning attacks during the fine-tuning stage, we conduct a separate comparison with PGD, Shadowcast, and MP-Nav. These methods are adapted to the AIGC detection task under the same real-to-fake attack objective. The corresponding adaptation details and experimental configurations are provided together with the respective experiments.

Table 1: Baseline AIGC detection performance on TinyGenImage and FakeClue.

Model	TinyGenImage		FakeClue	
	Real Acc.	Fake Acc.	Real Acc.	Fake Acc.
Qwen-2.5	0.950	0.417	0.704	0.266
FakeVLM	0.992	0.985	0.982	0.991
AIGI-Holmes	0.922	0.999	0.978	0.459
MiniGPT-4	0.982	0.071	0.980	0.058

Evaluation Metrics. Our attack aims to induce a poisoned model to misclassify originally correctly recognized real images as AI-generated. We refer to this systematic real-to-fake misclassification behavior as detection hallucination, and use the Real-to-Fake Attack Success Rate (R2F-ASR) as the primary metric for attack effectiveness.

To exclude errors that already exist in the clean victim model, we only consider real test samples that are correctly classified before poisoning. Let $\mathcal{D}_r$ denote the real-image test set, f_θ the clean model, and f_{θ^p} the poisoned model. We first define

$$\mathcal{D}_r^+ = \{x \in \mathcal{D}_r \mid f_\theta(x) = \text{real}\} \tag{11}$$

The R2F-ASR is then defined as

$$\text{R2F-ASR} = \frac{\sum_{x_i \in \mathcal{D}_r^+} \mathbb{I}\left[f_{\theta^p}(x_i) = \text{fake}\right]}{|\mathcal{D}_r^+|} \tag{12}$$

where $\mathbb{I}(\cdot)$ is the indicator function. R2F-ASR measures the proportion of originally correct real-image predictions that are flipped to fake after poisoning. A higher R2F-ASR therefore indicates a stronger attack and more severe detection hallucination.

We additionally report the clean detection accuracy of each victim model before poisoning to characterize its baseline AIGC detection capability, but do not treat it as an attack success metric.

5.2 Clean AIGC Detection Performance

Before conducting the poisoning experiments, we first evaluate the AIGC detection capability of each base model on the TinyGenImage and FakeClue. As shown in Table 1, all models achieve high accuracy on real images, while their performance on AI-generated images varies substantially. Given this stable detection baseline on real images, we focus on the real-to-fake (R2F) attack, which aims to induce images that are originally classified correctly as real to be misclassified as AI-generated.

5.3 Attack Effectiveness and Poisoning Efficiency

We evaluate $N = \{25, 50, 75, 100, 200, 300, 400\}$ poisoning budgets using R2F-ASR. Poisons are constructed from TinyGenImage and evaluated on both TinyGenImage and FakeClue to measure attack effectiveness and cross-dataset transferability.

As shown in Table 2, MiragePoison achieves strong few-shot poisoning effectiveness across multiple VLMs. On TinyGenImage, the R2F-ASR of FakeVLM reaches 0.679 with only $N = 50$ poisoned samples and further increases to 0.982 at $N = 100$. AIGI-Holmes and MiniGPT-4 are even more vulnerable, both reaching an R2F-ASR of 1.000 with only 50 poisoned samples. In contrast, Qwen2.5-VL-3B remains around 0.15 across different poisoning budgets, indicating substantial differences in poisoning robustness among VLMs.

MiragePoison also exhibits strong cross-dataset transferability. Although the poisoned samples are constructed solely from TinyGenImage, the induced attack effect transfers effectively to FakeClue. For FakeVLM, the R2F-ASR on FakeClue increases from 0.034 at $N = 25$ to 0.987 at $N = 200$, and eventually reaches 1.000 at $N = 400$. These results indicate that the erroneous authenticity behavior induced by MiragePoison is not confined to the data distribution used for poison construction, but can generalize to test images from different data sources.

Table 2: Attack success rate under different poisoning budgets.

Dataset	Model	N=0	N=25	N=50	N=75	N=100	N=200	N=300	N=400	Params.
TinyGenImage	FakeVLM	0.000	0.005	0.679	0.690	0.982	0.997	0.977	0.997	7B
	Qwen2.5-VL-3B	0.000	0.148	0.151	0.149	0.147	0.151	0.148	0.147	3B
	AIGI-Holmes	0.000	0.597	1.000	1.000	1.000	1.000	1.000	1.000	7B
	MiniGPT-4	0.000	0.002	1.000	1.000	1.000	1.000	1.000	1.000	7B
FakeClue	FakeVLM	0.000	0.034	0.525	0.551	0.784	0.987	0.926	1.000	7B
	Qwen2.5-VL-3B	0.000	0.220	0.217	0.219	0.212	0.222	0.223	0.219	3B
	AIGI-Holmes	0.000	0.799	1.000	1.000	1.000	1.000	1.000	1.000	7B
	MiniGPT-4	0.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	7B

Table 3: Comparison with representative attacks under the unified Real-to-Fake setting.

Method	Attack Type	Threat Model	Test Modify	Trigger	Venue	ASR
FakeVLM	None	None	None	None	–	0.01
MF-it	Adversarial	Gray-box	✓	✗	NeurIPS 2023	0.27
MF-ii	Adversarial	Gray-box	✓	✗	NeurIPS 2023	0.68
PA-Attack	Adversarial	Gray-box	✓	✗	CVPR 2026	0.51
BadVision	Backdoor	White-box	✓	✓	CVPR 2025	0.96
VL-Trojan	Backdoor	Gray-box	✓	✓	IJCV 2025	0.68
MTAttack	Backdoor	Gray-box	✓	✓	AAAI 2026	0.92
Ours (N=100)	Poisoning	Gray-box	✗	✗	–	**0.98**
Ours (N=100)	Poisoning	Black-box	✗	✗	–	**1.00**

Overall, MiragePoison can substantially alter the authenticity judgments of multiple VLMs with a small poisoning budget, while exhibiting strong cross-dataset transferability. The results also reveal pronounced differences in robustness to fine-tuning poisoning across different VLMs.

5.4 Comparison with Existing Attacks

To evaluate the effectiveness of MiragePoison, we first compare it with three adversarial attack methods, including MF-it, MF-ii, and PA-Attack, as well as three backdoor attack methods, including BadVision, VL-Trojan, and MTAttack. We further compare MiragePoison with PGD, Shadowcast, and MP-Nav, both of which require model fine-tuning.

Comparison with adversarial and backdoor attacks. As shown in Table 3, the three adversarial attacks achieve ASRs ranging from 0.27 to 0.68, while the backdoor attacks generally exhibit stronger attack performance, with BadVision and MTAttack reaching 0.96 and 0.92, respectively. In comparison, MiragePoison achieves the highest ASR of 0.98 using only 100 poisoned samples, outperforming all competing methods. These results demonstrate that MiragePoison can effectively induce originally correctly classified real images to undergo real-to-fake prediction flips with a limited poisoning budget, yielding a higher attack success rate than existing adversarial and backdoor attacks.

Comparison with fine-tuning-based poisoning methods. As shown in Table 4, MiragePoison is substantially more effective than PGD, Shadowcast, and MP-Nav under low poisoning budgets, reaching 0.679 and 0.525 R2F-ASR on TinyGenImage and FakeClue at $N = 50$, respectively. PGD becomes competitive at larger budgets, whereas Shadowcast and MP-Nav remain relatively weak across the tested range. MiragePoison therefore exhibits stronger few-shot poisoning efficiency across both datasets.

5.5 Ablation Study on Feature Guidance

We evaluate two MiragePoison variants, Real-only and Fake-only, which retain only real-feature attraction or fake-feature repulsion, respectively. All variants use the same Top-N poison selection strategy and are evaluated on FakeClue with R2F-ASR.

Table 4: Comparison of different poisoning methods under varying poisoning budgets.

Dataset	Method	Poisoning Budget (N)							
		0	**25**	**50**	**75**	**100**	**200**	**300**	**400**
TinyGenImage	PGD	0.000	0.000	0.114	0.393	0.902	0.990	**0.989**	0.848
	Shadowcast	0.000	0.000	0.002	0.005	0.009	0.048	0.107	0.277
	MP-Nav	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.007
	Ours	**0.000**	**0.005**	**0.679**	**0.690**	**0.982**	**0.997**	0.977	**0.997**
FakeClue	PGD	0.000	0.008	0.040	0.106	0.560	0.759	**0.992**	0.955
	Shadowcast	0.000	0.022	0.025	0.030	0.036	0.066	0.104	0.235
	MP-Nav	0.000	0.009	0.009	0.009	0.008	0.010	0.015	0.029
	Ours	**0.000**	**0.034**	**0.525**	**0.551**	**0.784**	**0.987**	0.926	**1.000**

Table 5: Ablation study of different poison construction strategies under varying poisoning budgets.

Method	Real Attraction	Fake Repulsion	Poison Selection	R2F-ASR				
				0	**50**	**100**	**200**	**400**
Real-only	✓	✗	Top-N	0.000	0.493	0.778	0.932	0.967
Fake-only	✗	✓	Top-N	0.000	**0.575**	0.652	0.980	1.000
MiragePoison	✓	✓	Top-N	**0.000**	0.525	**0.784**	**0.987**	**1.000**

As shown in Table 5, the full method reaches 0.784 and 0.987 R2F-ASR at $N = 100$ and $N = 200$, outperforming both variants. Although Fake-only is slightly stronger at $N = 50$, the joint guidance is more stable as the budget increases, indicating complementary effects.

5.6 Component-wise Vulnerability Analysis

To investigate the vulnerability of different VLM components, we control the trainable states of the Vision Encoder, Projector, and LLM and evaluate each configuration on TinyGenImage and FakeClue. As shown in Fig. 4, both datasets exhibit a consistent pattern. When the Vision Encoder is trainable, R2F-ASR increases rapidly with the poisoning budget, while additionally updating the Projector or LLM provides limited improvement.

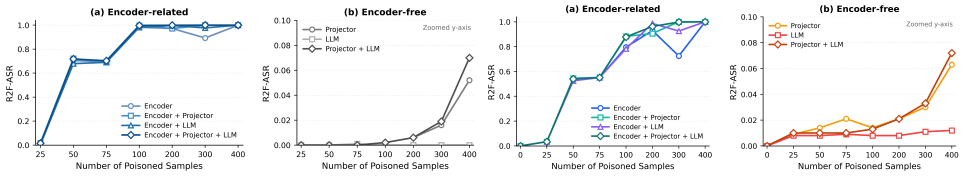

Figure 4: Component-wise vulnerability under different poisoning budgets on TinyGenImage and FakeClue. Encoder-related settings update the Vision Encoder, whereas Encoder-free settings keep it frozen. Encoder-free panels use a magnified y-axis range for clarity.

6 Conclusion

We investigate few-shot poisoning vulnerabilities in VLM-based AIGC detectors and propose MiragePoison, a trigger-free Real2Fake poisoning framework. Experiments demonstrate strong poisoning efficiency and cross-dataset transferability, while identifying the Vision Encoder as the primary vulnerable component. These findings motivate more robust fine-tuning and training-data verification for VLM-based authenticity detection.

AI use statement

In this work, we used generative AI tools to assist with English translation, experimental design feedback, code development and debugging, and interpretation of experimental results. Generative

AI tools were also used to support manuscript drafting and editing, improve readability, and assist with literature discovery and reference organization. All experiments, quantitative results, and final scientific conclusions were independently conducted, verified, and reviewed by the authors. We reviewed all AI-assisted content and take full responsibility for the final content of this work, including its text, claims, code, and experimental findings.

Ethics statement

This work investigates data poisoning vulnerabilities in VLM-based AIGC detection systems with the goal of improving their security and robustness. Although the proposed attack may have potential dual-use implications, our experiments are conducted solely in controlled research settings using publicly available models and datasets. We do not target deployed systems or real-world users. By exposing these vulnerabilities, we aim to facilitate the development of more robust training procedures and defenses against poisoning attacks.

Reproducibility Statement

We have provided sufficient methodological and experimental details to enable full reproduction of all results reported in this work, including the datasets, victim models, poison construction and selection procedures, fine-tuning configurations, poisoning budgets, and evaluation protocol. To further support reproducibility and independent verification, we will publicly release the complete source code, experiment configurations, and reproduction scripts after publication.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. *arXiv preprint arXiv:1206.6389*, 2012.

Nicholas Carlini and Andreas Terzis. Poisoning and backdooring contrastive learning. *arXiv preprint arXiv:2106.09667*, 2021.

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Feng Ding, Yue Zhang, Mengyao Xiao, Kangkang Wei, Zhangyi Shen, and Hong Rao. Vision-text interactive hybrid granularity proxy representation learning for ai-generated image detection. In *2025 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 102–107. IEEE, 2025.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pp. 3247–3258. PMLR, 2020.

Xinan He, Yue Zhou, Bing Fan, Bin Li, Guopu Zhu, and Feng Ding. Vlforgery face triad: Detection, localization and attribution via multimodal large language models. *Advances in Neural Information Processing Systems*, 38:163010–163044, 2026.

W Ronny Huang, Jonas Geiping, Liam Fowl, Gavin Taylor, and Tom Goldstein. Metapoison: Practical general-purpose clean-label data poisoning. *Advances in Neural Information Processing Systems*, 33:12080–12091, 2020.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023.

Jiawei Liang, Siyuan Liang, Aishan Liu, and Xiaochun Cao. Vl-trojan: Multimodal instruction backdoor attacks against autoregressive visual language models. *International Journal of Computer Vision*, 133(7):3994–4013, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Zhaoyi Liu and Huan Zhang. Stealthy backdoor attack in self-supervised learning vision encoders for large vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25060–25070. IEEE, 2025.

Weimin Lyu, Lu Pang, Tengfei Ma, Haibin Ling, and Chao Chen. Trojvlm: Backdoor attack against vision language models. In *European Conference on Computer Vision*, pp. 467–483. Springer, 2024.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Hefei Mei, Zirui Wang, Chang Xu, Jianyuan Guo, and Minjing Dong. Pa-attack: Guiding gray-box attacks on lvlm vision encoders with prototypes and attention. *arXiv preprint arXiv:2602.19418*, 2026.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24480–24489. IEEE, 2023.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1–11, 2019.

Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. Hidden trigger backdoor attacks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 11957–11965, 2020.

Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. *Advances in neural information processing systems*, 31, 2018.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8692–8701. IEEE, 2020.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22388–22398. IEEE, 2023.

Zihan Wang, Guansong Pang, Wenjun Miao, Jin Zheng, and Xiao Bai. Mtattack: Multi-target backdoor attacks against large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10440–10448, 2026.

Siwei Wen, Peilin Feng, Hengrui Kang, Zichen Wen, Yize Chen, Jiang Wu, Conghui He, Weijia Li, et al. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. *Advances in Neural Information Processing Systems*, 38:58972–59005, 2026.

Yuancheng Xu, Jiarui Yao, Manli Shu, Yanchao Sun, Zichu Wu, Ning Yu, Tom Goldstein, and Furong Huang. Shadowcast: Stealthy data poisoning attacks against vision-language models. *Advances in Neural Information Processing Systems*, 37:57733–57764, 2024.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. In *International Conference on Learning Representations*, volume 2025, pp. 70702–70720, 2025.

SanTai Yang. Tiny genimage. Kaggle Dataset, 2024. URL https://www.kaggle.com/datasets/yangsangtai/tiny-genimage. A reduced version of the GenImage dataset.

Junyan Ye, Baichuan Zhou, Zilong Huang, Junan Zhang, Tianyi Bai, Hengrui Kang, Jun He, Honglin Lin, Zihao Wang, Tong Wu, et al. Loki: A comprehensive synthetic data detection benchmark using large multimodal models. In *International Conference on Learning Representations*, volume 2025, pp. 70440–70522, 2025.

Wei Ye, Xinan He, and Feng Ding. Decoupling forgery semantics for generalizable deepfake detection. *arXiv preprint arXiv:2406.09739*, 2024.

Jingfeng Zhang, Prashanth Krishnamurthy, Naman Patel, Anthony Tzes, and Farshad Khorrami. MP-nav: Enhancing data poisoning attacks against multimodal learning. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=zy7VeNtSLM.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems*, 36:54111–54138, 2023.

Yue Zhou, Bing Fan, Pradeep K. Atrey, and Feng Ding. Exposing deepfakes using dual-channel network with multi-axis attention and frequency analysis. In *Proceedings of the 2023 ACM Workshop on Information Hiding and Multimedia Security*, pp. 169–174, 2023.

Yue Zhou, Xinan He, KaiQing Lin, Bing Fan, Feng Ding, and Bin Li. Breaking latent prior bias in detectors for generalizable aigc image detection. *Advances in Neural Information Processing Systems*, 38:30649–30679, 2026.

Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. Aigi-holmes: Towards explainable and generalizable ai-generated image detection via multimodal large language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18746–18758. IEEE, 2025.

Chen Zhu, W Ronny Huang, Hengduo Li, Gavin Taylor, Christoph Studer, and Tom Goldstein. Transferable clean-label poisoning attacks on deep neural nets. In *International conference on machine learning*, pp. 7614–7623. PMLR, 2019.

Deyao Zhu, Xiaoqian Shen, Xiang Li, Mohamed Elhoseiny, et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations*, volume 2024, pp. 18378–18394, 2024.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in neural information processing systems*, 36:77771–77782, 2023.