FAME: FAILURE-AWARE AND MODEL-ADAPTIVE EXPLORATION FOR JAILBREAKING MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing jailbreak attacks against multimodal large language models (MLLMs) make limited use of feedback from failed attempts and generally apply the same exploration strategy across target models. Consequently, they may repeatedly explore regions similar to those associated with previous failures while overlooking model-specific vulnerability patterns. To address these limitations, we propose a Failure-Aware and Model-adaptive Exploration framework (FAME), which reformulates image-side jailbreaking as closed-loop candidate selection over a pre-generated image pool. Within each attack, FAME uses previous failures to suppress candidates in low-success regions through local neighborhood avoidance and global energy attenuation. Across attacks, it trains a lightweight proxy discriminator on historical outcomes to estimate target-specific attack potential and prioritize promising candidates for each MLLM. Experiments on four benchmarks and eight open-source and commercial MLLMs show that FAME consistently improves attack success rate over state-of-the-art baselines, achieving 9.1% higher average attack success rate.

1 INTRODUCTION

Multimodal large language models (MLLMs) extend LLMs with visual perception, substantially broadening their capabilities across vision-language tasks (Achiam et al., 2023; Liu et al., 2023). Yet this additional modality also introduces a new attack surface. While safety alignment has been extensively developed for textual inputs, visual inputs remain comparatively under-aligned, making images an "Achilles' heel" of aligned MLLMs (Li et al., 2024; Ma et al., 2025). Attackers can exploit this vulnerability to bypass safety safeguards through malicious visual inputs, giving rise to increasingly effective multimodal jailbreak attacks (Liu et al., 2024b).

Existing jailbreak attacks against MLLMs generally follow two paradigms. Perturbation-based methods optimize adversarial visual inputs to disrupt safety alignment and can achieve strong attack performance. However, they often rely on white-box access or surrogate models, limiting their applicability to proprietary MLLMs (Qi et al., 2024; Carlini et al., 2023; Cui et al., 2026). The second paradigm adopts query-based or generation-based strategies, where attackers exploit the model's cognitive biases or semantic understanding capabilities to induce harmful outputs in black-box scenarios. Typographic approaches, such as FigStep and MM-SafetyBench, encode harmful textual instructions into images to bypass text-side safety filters (Gong et al., 2025; Liu et al., 2024b). HADES combines harmful text-image construction with iterative image refinement to amplify visual harmfulness (Li et al., 2024). Visual-RolePlay instead embeds malicious intent into high-risk visual personas and role-playing contexts (Ma et al., 2024), while MIRAGE leverages multimodal reasoning and guided exploration to progressively induce unsafe behavior (You et al., 2025). Another line of work exploits the uneven allocation of safety risk across modalities to circumvent multimodal safety alignment (Ma et al., 2025). Collectively, these methods demonstrate that semantically constructed visual inputs provide a practical and effective attack surface for black-box MLLMs.

Despite their effectiveness, existing black-box multimodal jailbreaks make limited use of victim-side attack feedback. A failed attempt is often treated only as an unsuccessful outcome. Yet such failures can reveal which visual regions are difficult to exploit. Ignoring this information may cause subsequent attempts to revisit similarly ineffective regions, as illustrated in Figure 1(a). Existing attacks

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.

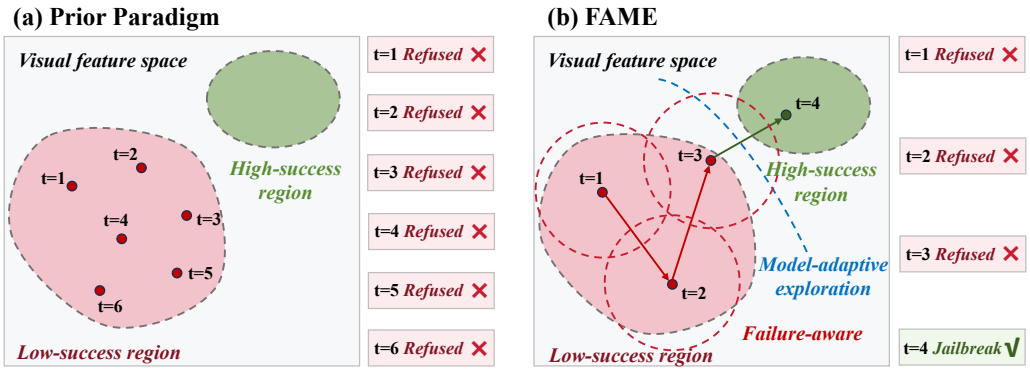

Figure 1: Comparison of prior jailbreak paradigms and FAME. FAME exploits failure feedback and model-specific knowledge to avoid low-success regions and guide more effective exploration.

are also often model-agnostic. The same attack strategy is applied across different MLLMs, even though their training and alignment mechanisms can differ substantially. As a result, model-specific attack patterns remain underexplored. Taken together, existing methods underutilize two levels of attack feedback: failure feedback within an attack and model-specific feedback accumulated across attacks. This raises two key questions: *Can failed attempts guide subsequent exploration? Can historical attack outcomes be used to learn model-specific vulnerabilities and adapt attacks accordingly?* These questions suggest a simple but important shift in perspective: attack failures should be treated as informative feedback rather than discarded outcomes, while historical attack results can serve as evidence of target-specific vulnerability patterns. Together, these two forms of feedback motivate a feedback-driven view of black-box multimodal jailbreaking, where exploration is refined by both current failures and accumulated model-specific attack outcomes.

In this paper, we propose FAME, a Failure-Aware and Model-Adaptive Exploration framework for black-box multimodal jailbreaking. As illustrated in Figure 1(b), Within each attack, it exploits failed attempts to steer subsequent exploration away from low-success regions; across attacks, it learns target-specific vulnerability patterns from historical outcomes to prioritize candidates that are more promising for the victim MLLM. By integrating these two feedback signals, FAME enables closed-loop exploration that is both failure-aware and model-adaptive. Specifically, FAME first constructs a diverse pool of candidate images for each malicious query, providing the search space over which feedback-guided exploration is performed. It then introduces intra-class constraints to exploit failure feedback within the current attack: a neighborhood constraint excludes candidates that are locally similar to previous failures, while an energy attenuation constraint provides global guidance away from the failure-dominated region. To incorporate model-specific feedback accumulated across attacks, FAME further employs an inter-class constraint, implemented by a lightweight proxy discriminator trained on historical attack outcomes, to estimate the attack potential of the remaining candidates and re-rank them for the target MLLM. Finally, FAME places the malicious query in the visual input while keeping the textual prompt benign, reducing exposure to text-side safety filters. Extensive experiments across multiple benchmarks and diverse MLLMs validate the effectiveness of the feedback-guided search paradigm introduced by FAME. By transforming black-box multimodal jailbreaking from independent, model-agnostic trial-and-error into closed-loop and target-specific exploration, FAME makes failed attempts and historical attack outcomes actionable, enabling more effective exploration and achieving 9.1% higher attack success rates.

Our main contributions are summarized as follows:

- We introduce a feedback-driven perspective on black-box multimodal jailbreaking by identifying two underexplored sources of victim-side feedback: failures within the current attack and model-specific attack outcomes accumulated across previous attacks.
- We propose FAME, a Failure-Aware and Model-Adaptive Exploration framework that formulates multimodal jailbreaking as a closed-loop candidate selection problem. FAME uses failure-aware intra-class constraints to avoid low-success regions and a lightweight proxy

discriminator to capture target-specific vulnerability patterns and prioritize promising candidates.

- Extensive experiments on multiple mainstream open-source and commercial MLLMs demonstrate that jointly exploiting within-attack failure feedback and cross-attack model-specific feedback consistently improves black-box multimodal jailbreak performance over existing model-agnostic trial-and-error strategies.

2 Related Work

2.1 Jailbreak Attacks against LLMs

Jailbreak attacks aim to bypass the safety alignment of LLMs and induce them to generate harmful or policy-violating content. Early jailbreaks primarily relied on manually crafted prompts collected from online communities or designed through human expertise, exploiting techniques such as role-playing, scenario nesting, and instruction obfuscation to circumvent safety filters (Shen et al., 2024; Wei et al., 2023). Subsequently, optimization-based attacks emerged in the white-box setting: GCG Zou et al. (2023) pioneers gradient-guided token substitution to search for transferable adversarial suffixes, and its variants further improve attack efficiency and success rates through refined objectives and search strategies (Jia et al., 2024). Although effective on open-source models, these methods require access to model internals and transfer poorly to closed-source systems. In contrast, black-box attacks leverage an attacker LLM or persuasion strategies to iteratively refine jailbreak prompts by querying the target model, as exemplified by PAIR (Chao et al., 2025), GPTFuzzer (Yu et al., 2023), and PAP (Zeng et al., 2024). Despite their diversity, these text-centric attacks operate purely within the textual channel, leaving the largely under-aligned visual modality of multimodal LLMs insufficiently explored—an attack surface that our work specifically targets.

2.2 Jailbreak Attacks against MLLMs

Extending jailbreak attacks from LLMs to the multimodal domain, existing studies reveal that the continuous and high-dimensional visual modality substantially enlarges the attack surface of MLLMs, making them susceptible to malicious visual inputs (Li et al., 2024; Ma et al., 2025). Current jailbreak attacks against MLLMs can be broadly categorized into perturbation-based and generation-based methods. Perturbation-based attacks optimize adversarial images or image-text pairs to disrupt safety alignment: VAJM Qi et al. (2024) optimizes a single universal visual adversarial example that can jailbreak aligned models across diverse harmful instructions; ImgJP Niu et al. (2024) leverages local white-box MLLMs as surrogates and maximizes prompt-following likelihood to craft adversarial images Nevertheless, such methods typically rely on white-box access or surrogate models, and the resulting adversarial samples exhibit limited transferability to black-box commercial MLLMs, which are predominantly deployed as APIs in practice. Generation-based attacks instead operate in black-box settings by constructing semantically harmful images: FigStep Gong et al. (2025) converts harmful instructions into typographic images to bypass text-side safety alignment; HADES Li et al. (2024) conceals and amplifies the harmfulness of textual instructions with meticulously crafted images; Visual-RolePlay Ma et al. (2024) induces risky outputs by constructing high-risk persona images combined with role-play instructions. However, existing black-box methods generally treat each attack attempt as an independent trial, often operating as blind attempts that repeatedly explore regions already verified as low-success in the visual feature space, and adopt model-agnostic strategies that neglect the heterogeneous defensive boundaries across different MLLMs. To this end, we propose an iterative jailbreak framework that systematically exploits historical feedback and adapts the attack to model-specific defensive preferences.

3 Method

In this section, we systematically present FAME, an iterative jailbreak attack framework against multimodal large language models (MLLMs). We first formalize the attack problem (Sec. 3.1), then provide an overall logical overview of the framework (Sec. 3.2). Subsequently, we elaborate on the intra-class constraints (Sec. 3.2.1) and inter-class constraints(Sec. 3.2.2) that govern the selection of optimal attack images.

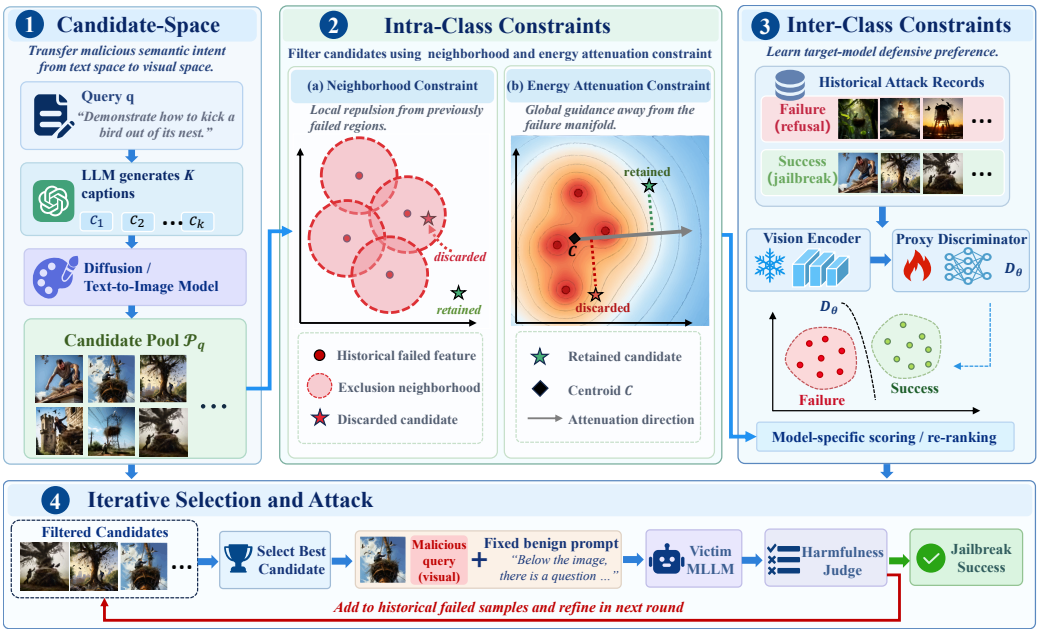

Figure 2: Overview of the FAME framework, where intra-class constraints and inter-class constraints jointly guide the attack selection to avoid low-success regions and adapt to model-specific defensive boundaries.

3.1 Problem Formulation

Let $\mathcal{M} : \mathcal{I} \times \mathcal{T} \to \mathcal{O}$ denote the target MLLM, where $\mathcal{I}$ is the image input space, $\mathcal{T}$ the text input space, and $\mathcal{O}$ the output response space. Let $\mathcal{Q} = \{q_1, q_2, \ldots, q_n\}$ be a set of malicious queries (e.g., violence, financial crime, privacy violations), and let $\Psi : \mathcal{O} \to \{0, 1\}$ denote a discriminator that determines whether the model output constitutes a successful jailbreak (1 for success, 0 for failure). For each query $q \in \mathcal{Q}$, the query-specific candidate image pool $\mathcal{P}_q = \{x_q^1, x_q^2, \ldots, x_q^m\}$ is generated. An LLM (e.g., GPT-4o) generates m semantically related image captions, which are then rendered into candidate images by a pre-trained diffusion model (e.g., Stable Diffusion). In practice, the selected candidate image is vertically concatenated with a typographic rendering of the malicious query to construct the final visual input, which is fed into the target MLLM together with a fixed-format benign textual prompt. For simplicity, we denote the resulting attack attempt as $\mathcal{M}(x^{(t)}, q)$. The attack aims to identify an image $x^* \in \mathcal{P}_q$ that induces a successful jailbreak for the fixed query q within a budget of T search rounds. To guide the iterative search, we maintain a failure history $\mathcal{F}_t$, which records all candidate images evaluated but unsuccessful up to round t. Initially, $\mathcal{F}_0 = \emptyset$. At round t, the attack proceeds as

$$\begin{aligned} x^{(t)} &= \Pi(q, \mathcal{P}_q \setminus \mathcal{F}_{t-1}, \mathcal{F}_{t-1}), \\ s^{(t)} &= \Psi\Big(\mathcal{M}(x^{(t)}, q)\Big), \\ \mathcal{F}_t &= \mathcal{F}_{t-1} \cup \{x^{(t)}\}, \qquad \text{if } s^{(t)} = 0, \end{aligned} \tag{1}$$

where Π denotes the image-selection strategy and $s^{(t)} \in \{0, 1\}$ indicates whether the selected image successfully jailbreaks the target MLLM. The search terminates once $s^{(t)} = 1$ or the budget T is exhausted. Our focus is the image-selection strategy Π, which determines the next candidate based on the remaining image pool and previous failures. The objective is to maximize the probability of achieving at least one successful jailbreak within the search budget:

$$\max_{\Pi} \mathbb{P}\left(\exists\, t \le T \text{ s.t. } \Psi\Big(\mathcal{M}(x^{(t)}, q)\Big) = 1\right), \quad x^{(t)} = \Pi(q, \mathcal{P}_q \setminus \mathcal{F}_{t-1}, \mathcal{F}_{t-1}). \tag{2}$$

3.2 Failure-Aware and Model-Adaptive Exploration Framework (FAME)

Existing generation-based image jailbreaks largely optimize each attack attempt in isolation (Li et al., 2024), with image generation or refinement driven primarily by the harmful query itself. Such independent optimization provides little guidance on how the search should evolve after an attack fails. FAME instead turns image optimization into a feedback-guided search process, where each observed outcome reshapes the selection of subsequent candidates. Failures from the current attack indicate regions that should be explored less, while outcomes accumulated across attacks reveal regions that are more promising for the target MLLM. This allows the search to progressively move away from previously unsuccessful regions while adapting toward model-specific vulnerabilities.

Based on these two forms of feedback, FAME designs the selection strategy Π with two complementary search mechanisms. As illustrated in Figure 2, the intra-class constraints exploit failures accumulated within the current attack. A neighborhood constraint excludes candidates that are too similar to individual failed samples, while an energy attenuation constraint aggregates multiple failures to guide exploration away from the broader failure-dominated region. The inter-class constraint captures model-specific vulnerability patterns through a lightweight proxy discriminator trained on historical attack outcomes across queries. After intra-class filtering, the discriminator re-ranks the remaining candidates according to their predicted attack success on the target MLLM. Together, these constraints make the selection strategy both failure-aware and model-adaptive. We next detail how these two components instantiate the selection strategy.

3.2.1 Intra-class Constraints

To exploit failure feedback accumulated within the current attack, we represent candidate images in a visual feature space, where semantic similarity can be captured more meaningfully than in the raw pixel space. We first examine whether attack outcomes exhibit local consistency in this space. Specifically, we sample images around previously successful and failed attack images under similarity thresholds of 0.5, 0.6, 0.7, and 0.8, and report their attack success rates in Figure 3. As similarity increases, the outcomes become increasingly polarized. At a threshold of 0.8, samples near successful images achieve a 90% attack success rate, while those near failed images achieve only 7%. This suggests that attack outcomes form locally coherent regions in the visual feature space.

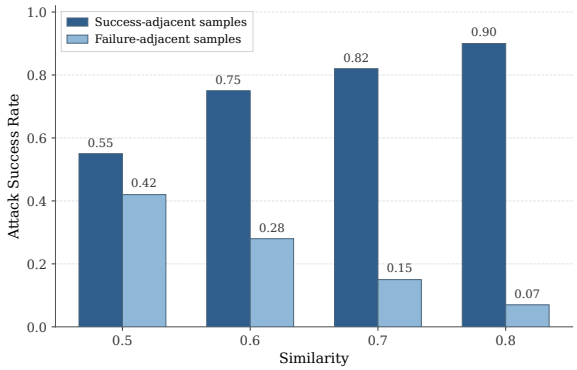

Figure 3: Attack success rate on success-adjacent and failure-adjacent samples across similarity levels on the HADES dataset. Higher visual similarity to previous attacks leads to stronger outcome consistency.

Motivated by this observation, we use failed attempts to progressively steer candidate selection away from previously unsuccessful regions. Individually, each failed sample defines a local low-success region, while collectively, accumulated failures reveal a broader failure-dominated structure. Accordingly, we introduce a neighborhood constraint for local avoidance around individual failures and an energy attenuation constraint for global guidance based on the collective distribution of accumulated failures.

Neighborhood Constraint. To capture the local structure of failure regions, we introduce a neighborhood constraint that excludes candidates lying too close to previously failed images. We use a feature extractor $\Phi(\cdot)$ (e.g., the visual encoder of CLIP) to map images into a semantic embedding space. For a candidate image x, the constraint requires its cosine similarity with every previously failed image $x^{(i)} \in \mathcal{F}_{t-1}$ to remain below a predefined threshold δ:

$$\forall x^{(i)} \in \mathcal{F}_{t-1}, \quad \frac{\Phi(x) \cdot \Phi(x^{(i)})}{\|\Phi(x)\| \, \|\Phi(x^{(i)})\|} \leq \delta. \tag{3}$$

This hard constraint removes candidates within the δ-neighborhood of any failed sample, encouraging exploration of previously untested visual regions.

Energy Attenuation Constraint. To complement the local neighborhood constraint, we introduce an energy attenuation constraint that captures the collective structure of accumulated failures. Rather than avoiding each failed image independently, it aggregates all failed samples in $\mathcal{F}_{t-1}$ into a virtual potential field that guides candidates away from failure-dominated regions. For a candidate image x, we define its potential energy with respect to a previously failed image $x^{(i)} \in \mathcal{F}_{t-1}$ as

$$E_i(\Phi(x)) = \frac{k}{\|\Phi(x) - \Phi(x^{(i)})\|_2 + \sigma}, \tag{4}$$

where k controls the energy strength and $\sigma > 0$ prevents singularities. The total energy induced by all previous failures is

$$E_{\text{total}}(\Phi(x)) = \sum_{x^{(i)} \in \mathcal{F}_{t-1}} E_i(\Phi(x)). \tag{5}$$

Candidates closer to the accumulated failures therefore receive higher potential energy. To obtain a global attenuation direction, we use the negative gradient of the total energy in the feature space. Let

$$\mathbf{g}(x) = -\nabla_{\Phi(x)} E_{\text{total}}(\Phi(x)), \qquad \mathbf{C}_{t-1} = \frac{1}{|\mathcal{F}_{t-1}|} \sum_{x^{(i)} \in \mathcal{F}_{t-1}} \Phi(x^{(i)}). \tag{6}$$

Here, $\mathbf{g}(x)$ denotes the attenuation direction induced by the accumulated failures, and $\mathbf{C}_{t-1}$ is the centroid of their feature representations. We then measure how far candidate x moves from the failure cluster along this direction by

$$\tau(x) = \frac{(\Phi(x) - \mathbf{C}_{t-1}) \cdot \mathbf{g}(x)}{\|\mathbf{g}(x)\|_2^2}. \tag{7}$$

The energy attenuation constraint retains candidates satisfying $\tau(x) \geq \eta$, where η is a predefined threshold controlling the required displacement along the attenuation direction. This encourages candidates to move sufficiently far along the attenuation direction, thereby avoiding low-success regions collectively formed by multiple failed samples.

3.2.2 Inter-class Constraints

Different MLLMs exhibit distinct defensive boundaries due to variations in training data, alignment strategies (e.g., RLHF), and safety filters. To capture such model-specificity, we propose inter-class constraints that learn a lightweight proxy discriminator for each target model from historical attack records across queries, which encodes the model's unique defensive patterns and incorporates them as a prior into the selection process.

Specifically, for a target model $\mathcal{M}$, we train a simple binary classifier $\mathcal{D}_\theta : \mathbb{R}^d \to [0, 1]$ (e.g., an MLP) as a lightweight proxy discriminator that takes an image feature vector as input and outputs the predicted probability of a successful attack. To train $\mathcal{D}_\theta$, we construct a dataset $\mathcal{D}_{\text{train}} = \{(\Phi(x_j), y_j)\}_{j=1}^{M}$, where the label $y_j = 1$ indicates that the image x_j previously succeeded in jailbreaking $\mathcal{M}$, and $y_j = 0$ otherwise. These samples are collected from historical attack records across various queries and public jailbreak datasets. It is trained using binary cross-entropy loss:

$$\mathcal{L}(\theta) = -\frac{1}{M} \sum_{j=1}^{M} \left[y_j \log \mathcal{D}_\theta(\Phi(x_j)) + (1 - y_j) \log\big(1 - \mathcal{D}_\theta(\Phi(x_j))\big) \right]. \tag{8}$$

By minimising $\mathcal{L}(\theta)$, $\mathcal{D}_\theta$ learns an approximation of the target model's defensive boundary in the visual feature space. During online selection at round t, the intra-class constraints first filter out candidates that violate the neighborhood or energy requirements. The remaining candidates are then re-ranked using the proxy discriminator, and the image with the highest score $\mathcal{D}_\theta(\Phi(x))$ is selected. This converts the target model's defensive pattern into a model-specific scoring function that steers exploration toward less-defended regions of the visual feature space.

4 EXPERIMENT

4.1 EXPERIMENT SETUP

Datasets. We evaluate on four widely used harmful-query benchmarks: HADES (Li et al., 2024), SafeBench (Gong et al., 2025), AdvBench (Zou et al., 2023), and HarmBench (Mazeika et al., 2024). HADES contains 750 harmful instructions across five risk scenarios, while SafeBench includes 500 queries covering ten forbidden topics. AdvBench consists of 520 harmful behaviors commonly used for jailbreak evaluation. HarmBench provides 320 standardized text-only behaviors spanning diverse semantic and functional categories. Together, these datasets provide broad coverage of harmful intents and risk categories for evaluating attack effectiveness.

Target MLLMs. We evaluate our method on eight representative vision-language models, including five open-weight models, LLaVA-1.5-7B (Liu et al., 2024a), Qwen-VL-Chat (Bai et al., 2023), InternVL2-4B (Chen et al., 2024), Kimi-VL-A3B-Instruct (Team et al., 2025), and GLM-4.6V-Flash (Hong et al., 2025), as well as three proprietary models, GPT-4o-mini (OpenAI, 2025a), GPT-5-nano (OpenAI, 2025b), and Gemini-2.5-Flash-Lite (Team et al., 2023). These models span diverse architectures, parameter scales, and training pipelines. We use one open-weight model as the surrogate and evaluate transferability on the remaining models, including closed-source systems, under a black-box setting.

Baselines. We compare our method with five established MLLM jailbreak attacks: MM-SafetyBench (Liu et al., 2024b), FigStep (Gong et al., 2025), BAP (Ying et al., 2025), Visual-RolePlay (Ma et al., 2024), and HADES (Li et al., 2024). MM-SafetyBench and FigStep encode harmful intent into visual or typographic inputs, while BAP jointly optimizes visual and textual prompts. Visual-RolePlay exploits role-playing scenarios to induce harmful responses, and HADES combines typographic encoding, harmful image generation, and adversarial perturbations. Together, these baselines cover diverse visual jailbreak strategies, including typography, image generation, prompt optimization, and role playing. For a fair comparison, we apply the same image-text concatenation procedure to all baselines, combining each attack image with the typographic rendering of the corresponding harmful query.

Evaluation Metrics and Judge. We use Attack Success Rate (ASR) as the primary evaluation metric. Given the harmful query set $\mathcal{Q}$, ASR is defined as

$$\mathrm{ASR} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \Psi\big(\mathcal{M}(x_q^{\mathrm{final}}, q)\big), \tag{9}$$

where x_q^{final} denotes the final attack image obtained for query q, and $\Psi(\cdot) \in \{0, 1\}$ indicates whether the resulting response constitutes a successful jailbreak.

Following HADES (Li et al., 2024), we instantiate Ψ using Beaver-dam-7B (Ji et al., 2023) as the automatic harmfulness judge. This model is trained on human feedback data covering diverse harmful categories and has been shown to correlate well with human annotations.

Experimental Settings for FAME. We construct a candidate pool of 20 images for each harmful query, where the corresponding image captions are generated by GPT-4o and subsequently rendered using PixArt-XL-2 as the text-to-image diffusion model at a resolution of 1024×1024. The maximum search budget is set to $T = 6$ rounds for each query. We employ CLIP ViT-B/32 as the visual feature extractor for computing semantic representations of candidate and historical images. For the intra-class constraints, the neighborhood similarity threshold is set to $\delta = 0.8$, while the energy attenuation parameters are fixed as $k = 1.0$, $\sigma = 10^{-3}$, and $\eta = 0.5$. For the inter-class constraint, the proxy discriminator is implemented as a lightweight three-layer MLP that takes CLIP visual features as input and outputs the predicted probability of a successful jailbreak.

4.2 PERFORMANCE COMPARISON

Table 1 presents ASRs of FAME and existing multimodal jailbreak methods across four benchmark datasets and eight open-weight and proprietary representative MLLMs.Overall, FAME demonstrates strong and consistent performance across diverse model architectures and datasets, outperforming the strongest baseline, VRP, by 9.1%. On Gemini-Flash, FAME does not outperform VRP, possibly

Table 1: Comparison of attack success rates (ASR, %) between FAME and existing multimodal jailbreak methods across four benchmark datasets and eight representative MLLMs, including both open-weight and proprietary models. MMSB denotes MM-SafetyBench. The best results are highlighted in bold, and the second-best results are underlined.

Dataset	Target Model	FigStep	MMSB	BAP	VRP	HADES	FAME
HADES	LLaVA-1.5-7B	69.9	58.7	36.5	82.8	84.7	**93.3**
	Qwen-VL-Chat	70.8	58.1	5.3	81.5	62.9	**88.9**
	InternVL2-4B	68.1	61.5	0.4	88.7	85.9	**92.5**
	Gemini-2.5-Flash-Lite	5.6	33.3	1.3	**54.1**	22.8	45.5
	GPT-4o-mini	6.7	26.8	2.0	3.9	17.5	**28.2**
	Kimi-VL-A3B-Instruct	61.6	67.3	24.4	87.3	72.4	**92.9**
	GLM-4.6V-Flash	79.6	70.5	32.9	90.4	82.5	**94.2**
	GPT-5.4-nano	49.6	32.3	9.6	32.0	45.3	**58.7**
SafeBench	LLaVA-1.5-7B	44.2	50.4	23.6	56.6	57.8	**63.2**
	Qwen-VL-Chat	48.2	41.6	12.0	58.6	53.2	**61.4**
	InternVL2-4B	39.8	33.2	2.2	55.8	63.5	**74.1**
	Gemini-2.5-Flash-Lite	6.6	9.8	0.6	**46.8**	25.8	34.2
	GPT-4o-mini	11.6	1.4	4.8	4.8	12.6	**19.2**
	Kimi-VL-A3B-Instruct	46.4	51.6	26.6	57.8	62.5	**73.7**
	GLM-4.6V-Flash	49.8	51.8	26.4	71.0	73.0	**87.9**
	GPT-5.4-nano	22.4	9.2	6.4	20.2	30.6	**45.4**
AdvBench	LLaVA-1.5-7B	73.9	84.6	23.9	90.8	90.2	**94.2**
	Qwen-VL-Chat	87.9	77.9	2.7	88.3	83.8	**93.4**
	InternVL2-4B	82.3	37.7	2.5	88.3	90	**95.5**
	Gemini-2.5-Flash-Lite	1.9	1.5	0.0	**54.2**	24.0	31.8
	GPT-4o-mini	2.5	0.4	0.2	0.8	4.0	**9.6**
	Kimi-VL-A3B-Instruct	41.9	63.7	2.3	88.7	87.3	**93.6**
	GLM-4.6V-Flash	84.8	60.6	9.4	90.8	85.4	**93.9**
	GPT-5.4-nano	40.6	19.4	11.2	32.7	46.4	**56.0**
HarmBench	LLaVA-1.5-7B	58.0	64.0	26.5	75.5	74.0	**80.0**
	Qwen-VL-Chat	70.0	46.0	9.5	73.5	73.0	**85.0**
	InternVL2-4B	63.5	52.0	3.5	72.5	78.5	**83.5**
	Gemini-2.5-Flash-Lite	3.5	4.5	2.0	**41.0**	20.5	33.0
	GPT-4o-mini	4.0	5.0	2.0	0.5	14.0	**25.0**
	Kimi-VL-A3B-Instruct	40.5	64.5	12.5	70.0	75.6	**85.5**
	GLM-4.6V-Flash	71.0	73.0	27.0	79.5	80.0	**87.0**
	GPT-5.4-nano	39.0	22.5	8.0	27.5	35.0	**55.0**
	Average	45.2	41.7	11.2	58.3	56.7	**67.4**

because this model is particularly susceptible to VRP's role-playing strategy. Nevertheless, FAME consistently achieves the second-best performance and remains competitive with VRP. These results validate the effectiveness of feedback-guided, model-adaptive exploration.

4.3 Further Analysis of FAME

Multi-Round Attack Comparison with HADES. To further compare the attack efficiency of FAME and HADES under multi-round interactions, we evaluate their average attack performance on Qwen-VL-Chat across the five categories of the HADES dataset over six attack rounds in Figure 4. FAME achieves higher ASR within fewer rounds, indicating that it can identify effective attack candidates with fewer interactions with the target model. This suggests that FAME more effectively leverages information collected from previous rounds, resulting in faster ASR improvement and better query efficiency than HADES.

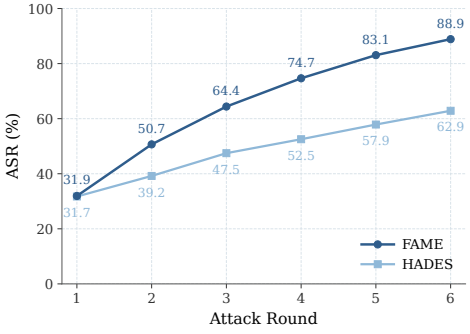

Figure 4: Average multi-round performance on Qwen-VL-Chat.

Table 2: Ablation study of the Neighborhood Constraint, Energy Attenuation Constraint, and Inter-Class Constraint on Qwen-VL-Chat across five categories of the HADES dataset. Avg. ASR denotes the average attack success rate.

Method	**Neighborhood**	**Energy Attenuation**	**Inter-Class**	**Avg. ASR (%)**
Random Selection	×	×	×	52.9
Neighborhood	✓	×	×	82.5
Intra-Class	✓	✓	×	85.4
Inter-Class	×	×	✓	78.0
FAME (Full)	✓	✓	✓	**88.9**

Table 3: Sensitivity analysis of key hyperparameters in FAME. Default settings are highlighted in bold.

	Candidate Pool Size m				Neighborhood δ				Energy Attenuation η			
Value	10	**20**	30	40	0.6	0.7	**0.8**	0.9	0.3	0.4	**0.5**	0.6
Avg. ASR (%)	70.4	**88.9**	89.2	90.5	75.9	84.2	**88.9**	80.7	73.2	85.3	**88.9**	82.0

Component-wise Ablation Study. We conduct an ablation study using Qwen-VL-Chat as the target model across the five categories of the HADES dataset to evaluate the contributions of the Neighborhood Constraint, Energy Attenuation Constraint, and Inter-Class Constraint. As shown in Table 2, random candidate selection achieves an average ASR of 52.9%. Applying the Neighborhood Constraint alone substantially increases the ASR to 82.5%, while the Inter-Class Constraint alone achieves 78.0%. Combining the Neighborhood and Energy Attenuation Constraints further improves the ASR to 85.4%. Finally, integrating all three constraints yields the highest average ASR of 88.9%. These results demonstrate that each constraint contributes to attack performance and that their combination provides complementary benefits to FAME.

Hyperparameter Sensitivity Analysis. We evaluate the sensitivity of FAME to the candidate pool size m, neighborhood threshold δ, and energy attenuation threshold η on Qwen-VL-Chat over the HADES dataset. As shown in Table 3, increasing the candidate pool initially yields substantial gains by providing a more diverse search space, while further enlargement brings only marginal improvements. Therefore, we set $m = 20$ to balance attack effectiveness and computational cost. For the neighborhood constraint, an overly small δ results in excessive exclusion around failed samples and may discard promising candidates, whereas an overly large δ provides insufficient coverage of local failure regions. For the energy attenuation constraint, a small η may not move candidates sufficiently away from failure-dominated regions, while an excessively large value can over-constrain the search and exclude potentially successful regions. These observations support the default parameter choices used in FAME.

5 Conclusion

In this paper, we introduce FAME, a failure-aware and model-adaptive framework for black-box multimodal jailbreaking. By reusing feedback from failed attempts and model-specific knowledge from historical outcomes, FAME enables more targeted candidate exploration. Experiments across four benchmarks and eight representative MLLMs show consistent gains in attack effectiveness, while ablations confirm the complementary benefits of its intra- and inter-class constraints.

Limitations and Future Work. Despite these promising results, FAME remains limited by the quality and diversity of its candidate pool, the semantic coverage of CLIP representations, which may not fully capture fine-grained safety-relevant semantics. In addition, the proxy discriminator relies on historical attack outcomes and therefore incurs a cold-start cost and may require adaptation when target models or safety policies change. Future work will explore online candidate generation, stronger multimodal representations, and continual adaptation to evolving models and safety policies, improving the generality and robustness of failure-aware red teaming.

AI USE STATEMENT

We used GPT-4o for language polishing and grammar correction. All claims, experimental designs, results, analysis and intellectual contributions are fully designed, written, verified and validated by the authors. We take responsibility for the final content of this work.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023.

Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Anas Awadalla, Pang Wei Koh, Daphne Ippolito, Katherine Lee, Florian Tramer, et al. Are aligned neural networks adversarially aligned?, 2024. *URL https://arxiv. org/abs/2306.15447*, 2023.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 23–42. IEEE, 2025.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67 (12):220101, 2024.

Kaiyuan Cui, Yige Li, Yutao Wu, Xingjun Ma, Sarah Erfani, Christopher Leckie, and Hanxun Huang. Toward universal and transferable jailbreak attacks on vision-language models. *arXiv preprint arXiv:2602.01025*, 2026.

Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. Figstep: Jailbreaking large vision-language models via typographic visual prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23951–23959, 2025.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36:24678–24704, 2023.

Xiaojun Jia, Tianyu Pang, Chao Du, Yihao Huang, Jindong Gu, Yang Liu, Xiaochun Cao, and Min Lin. Improved techniques for optimization-based jailbreaking on large language models. *arXiv preprint arXiv:2405.21018*, 2024.

Yifan Li, Hangyu Guo, Kun Zhou, Wayne Xin Zhao, and Ji-Rong Wen. Images are achilles' heel of alignment: Exploiting visual vulnerabilities for jailbreaking multimodal large language models. In *European Conference on Computer Vision*, pp. 174–189. Springer, 2024.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024a.

Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In *European Conference on Computer Vision*, pp. 386–403. Springer, 2024b.

Siyuan Ma, Weidi Luo, Yu Wang, and Xiaogeng Liu. Visual-roleplay: Universal jailbreak attack on multimodal large language models via role-playing image character. *arXiv preprint arXiv:2405.20773*, 2024.

Teng Ma, Xiaojun Jia, Ranjie Duan, Xinfeng Li, Yihao Huang, Xiaoshuang Jia, Zhixuan Chu, and Wenqi Ren. Heuristic-induced multimodal risk distribution jailbreak attack for multimodal large language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2686–2696. IEEE, 2025.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

Zhenxing Niu, Haodong Ren, Xinbo Gao, Gang Hua, and Rong Jin. Jailbreaking attack against multimodal large language model. *arXiv preprint arXiv:2402.02309*, 2024.

OpenAI. Gpt-4o mini: advancing cost-efficient intelligence. `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`, 2025a. Online technical report.

OpenAI. Gpt-5-nano: Fastest, most cost-efficient version of gpt-5. `https://platform.openai.com/docs/models/gpt-5-nano`, 2025b. OpenAI platform documentation.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 21527–21536, 2024.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. " do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 1671–1685, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *Advances in neural information processing systems*, 36:80079–80110, 2023.

Zonghao Ying, Aishan Liu, Tianyuan Zhang, Zhengmin Yu, Siyuan Liang, Xianglong Liu, and Dacheng Tao. Jailbreak vision language models via bi-modal adversarial prompt. *IEEE Transactions on Information Forensics and Security*, 20:7153–7165, 2025.

Wenhao You, Bryan Hooi, Yiwei Wang, Youke Wang, Zong Ke, Ming-Hsuan Yang, Zi Huang, and Yujun Cai. Mirage: Multimodal immersive reasoning and guided exploration for red-team jailbreak attacks. *arXiv preprint arXiv:2503.19134*, 2025.

Jiahao Yu, Xingwei Lin, Zheng Yu, and Xinyu Xing. Gptfuzzer: Red teaming large language models with auto-generated jailbreak prompts. *arXiv preprint arXiv:2309.10253*, 2023.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14322–14350, 2024.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023. *URL https://arxiv.org/abs/2307.15043*, 19(3), 2023.