전자통신동향분석, 41권 6호 2026, 90-101

비전-언어 모델의 최적화 기반 탈옥 공격 기술 동향

Trends in Optimization-Based Jailbreak Attacks on Vision Language Models

김재윤 (J.Y. Kim, jyoonkim@etri.re.kr) AI안전연구실 연구원
이성원 (S.W. Yi, sungyi@etri.re.kr) AI안전연구실 책임연구원/실장
동성희 (S.H. Dong, dsh7560@etri.re.kr) AI안전연구실 선임연구원
배강민 (K.M. Bae, kmbae@etri.re.kr) AI안전연구실 선임연구원
박민호 (M.H. Park, roger618@etri.re.kr) AI안전연구실 선임연구원
진현우 (H.W. Jin, maizer@etri.re.kr) AI안전연구실 석사후연수연구원
이진권 (J.K. Lee, rnjs8316@etri.re.kr) AI안전연구실 위촉연구원

ABSTRACT

Vision-language models (VLMs) extend large language models to visual perception, thereby enabling joint reasoning over visual and textual inputs. However, the continuous and high-dimensional visual input space introduces an additional adversarial surface that can compromise safety alignment. This study surveyed recent optimization-based jailbreak attacks on VLMs and classified them according to the principal optimization signal, distinguishing output-based, internal representation-based, and feedback-based approaches. We compared these approaches in terms of the optimized modality, model access assumptions, universality, transferability, and computational or query costs. Recent research has evolved from model-specific white-box perturbations to attacks that exploit safety-related internal representations, improve cross-model transferability, and use response feedback and zeroth-order estimation in black-box settings. Recurring challenges include a mismatch between surrogate objectives and harmful behaviors, dependence on model-specific representations, judge sensitivity, and inconsistent evaluation settings. Accordingly, this survey provides a unified perspective on the design and evaluation of optimization-based VLM jailbreak attacks under different threat models and generalization settings.

KEYWORDS Adversarial attack, Jailbreak, Multimodal Safety, Vision-language model

* DOI: https://doi.org/10.22648/ETRI.2026.J.410608
* 본 연구는 과학기술정보통신부의 재원으로 정보통신기획평가원(IITP)의 지원을 받아 수행되었음(과제번호: RS-2025-02215344, 과제명: 위험적 요소에도 강인하며 유연한 회복 탄력성을 가지는 인공지능 기술 개발). 또한, 과학기술정보통신부와 광주광역시가 공동 지원한 「인공지능 중심 산업융합 집적단지 조성사업」의 지원을 받아 수행되었음.

본 저작물은 공공누리 제4유형
출처표시+상업적이용금지+변경금지 조건에 따라 이용할 수 있습니다.
©2026 한국전자통신연구원

I. 서론

대규모 언어 모델(LLM: Large Language Model)은 유용한 지시 수행 능력을 유지하면서 유해하거나 위험한 요청에는 응답을 제한하도록 안전 정렬된다. 비전-언어 모델(VLM: Vision-Language Model)은 이러한 언어 모델에 시각 인코더와 멀티모달 결합 모듈을 통합하여 이미지와 텍스트를 함께 처리한다. 그러나 연속적이고 고차원적인 시각 입력은 텍스트 토큰과 상이한 공격 표면을 제공한다. 픽셀 수준의 작은 변화가 내부 표현을 크게 이동시킬 수 있고, 시각 경로를 통한 정보가 텍스트 중심의 안전 정렬을 우회하여 생성 행동에 영향을 줄 수 있기 때문이다. 이에 따라 시각 모달리티의 안전 취약성은 VLM 안전성 평가의 핵심 연구 주제로 부상하고 있다.

VLM 탈옥 공격은 수작업 프롬프트, 타이포그래피 기반 시각 프롬프트, 생성형 이미지 조작, 최적화 기반 적대적 입력 등으로 구분할 수 있다. 이 가운데 최적화 기반 공격은 명시적인 목적함수 또는 평가 신호에 따라 공격 입력을 반복적으로 갱신한다는 점에서 다른 접근과 구별된다. VisualAdv는 이미지 픽셀에 대한 경사를 이용하여 소규모 유해 문장 코퍼스의 생성 우도를 최대화하도록 이미지를 최적화하였고, 그 결과 하나의 적대적 이미지가 복수의 유해 질의에 범용적으로 작용함을 보였으며[1], imgJP는 최대우도 목적을 이용해 다양한 질의와 모델에 재사용 가능한 이미지 탈옥 프롬프트를 학습하였다[2]. 이후 연구는 이미지와 텍스트의 공동 최적화, 안전 관련 내부 표현의 직접 활용, 그리고 대상 모델의 응답만을 이용하는 블랙박스 탐색으로 범위를 확장해 왔다.

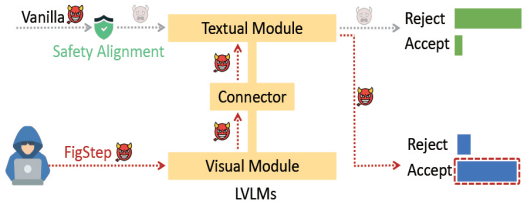
최적화 방식의 차이는 공격 성공률뿐만 아니라 요구되는 모델 접근 권한, 계산 비용, 질의 예산, 전이성 및 재사용 가능성을 결정한다. 동일한 이미지 섭동이라도 출력 토큰 우도를 직접 최적화하는 방법과 비전 임베딩 또는 중간 활성값을 활용하는 방법은 위협 모델이 다르다. 또한, 상용 API와 같이 내부 경사를 이용할 수 없는 환경에서는 외부 판정 모델이나 제로차 최적화를 통해 관측 가능한 응답을 최적화 신호로 변환해야 한다. 따라서 최근 연구를 이해하기 위해서는 공격 대상과 함께 최적화를 구동하는 신호를 살펴볼 필요가 있다.

표 1 최적화 신호에 따른 VLM 탈옥 공격의 구분

구분	핵심 신호	대표 방법
출력	응답 우도·행동	VisualAdv, imgJP, HADES, UMK, MLAI
내부 표현	임베딩·경계·활성	Jailbreak in Pieces, JailBound, HKVE, ActMan, FORCE, UltraBreak
피드백	응답·judge· ZO·reward	SI-Attack, HIMRD, Zer0-Jack, ZO-SPSA, ASO

본고에서는 최근의 최적화 기반 VLM 탈옥 기술을 최적화 신호의 성격에 따라 출력 기반, 내부 표현 기반, 피드백 기반의 세 계열로 구분한다. 각 계열의 핵심 최적화 신호와 대표적인 공격 방법을 표 1에 정리하였다. 각 접근의 목적함수, 공격 대상, 모델 접근 조건, 범용성 및 전이성을 비교하고, 평가 지표와 벤치마크 차이가 결과 해석에 미치는 영향을 논의한다. 마지막으로 화이트박스에서 블랙박스로 이어지는 기술 발전과 VLM 안전성 평가 및 방어에 대한 시사점을 살펴본다.

최적화 기반 공격을 비교할 때는 공격자가 실제로 갱신하는 변수와 이를 유도하는 최적화 신호를 구분할 필요가 있다. 이미지 픽셀을 갱신하더라도 손실은 출력 토큰 확률에서 계산될 수 있으며, 반대로 내부 활성값을 목적함수에 사용하더라도 최종 최적화 변수는 이미지일 수 있다. 또한, 대리 모델의 경사로 공격을 생성한 뒤 폐쇄형 모델에 적용하는



출처 Reprinted from Y. Gong et al., "FigStep: Jailbreaking Large Vision-Language Models via Typographic Visual Prompts," in Proc. AAAI Conf. Artif. Intell., (Philadelphia, PA, USA), 2025, pp. 23951-23959, doi: 10.1609/aaai.v39i22.34568. [3].

그림 1 시각 프롬프트를 이용한 VLM 안전장치 우회의 예시

전이 기반 블랙박스 공격과 대상 API의 응답만을 이용하여 입력을 직접 갱신하는 공격은 요구 정보와 비용 구조가 본질적으로 다르다. 이러한 차이를 구분하는 것은 서로 다른 연구의 성능을 동일한 위협 조건으로 해석하는 오류를 줄이는 데 중요하다.

그림 1은 최적화 기반 공격과의 차이를 명확히 하기 위해 대표적인 비최적화 시각 프롬프트 공격인 FigStep의 개념을 제시한다. FigStep은 유해 지시를 타이포그래피 이미지로 변환하고 무해한 텍스트 프롬프트와 결합하여 텍스트 안전장치와 시각 모듈 사이의 정렬 차이를 이용한다[3]. 이에 비해 본고에서 다루는 방법들은 시각적 우회 자체에 그치지 않고 손실함수, 내부 표현 또는 외부 피드백을 반복적으로 이용하여 공격 입력이나 공격 생성 과정을 최적화한다.

II. 최적화 기반 공격의 분류

1. 출력 기반 최적화

출력 기반 최적화는 목표 응답의 생성 우도 또는 긍정적 응답 행동을 직접 목적함수로 설정한다. 초기 시각 적대 공격은 특정 목표 문자열이나 유해 응답 코퍼스의 생성 확률을 높이는 방향으로 이미지 픽셀을 갱신하였다. 이러한 목적함수는 정의가 명확하고 역전파에 직접 사용할 수 있어 화이트박스 환경에서 효과적이지만, 손실 감소가 실제 정책 위반 정도와 항상 일치하지는 않는다. 이후 HADES, UMK, MLAI는 이미지 생성, 다중모달 공동 최적화, 평탄한 손실 영역 기반의 다중 이미지 선택을 결합하여 이러한 한계를 보완하였다[4-6].

2. 내부 표현 기반 최적화

내부 표현 기반 최적화는 출력 토큰 확률에만 의존하지 않고 비전 임베딩, 융합 계층의 안전 경계, attention 또는 key-value 분포, 중간 활성값, 계층별 특징과 같은 모델 내부 신호를 활용한다. Jailbreak in Pieces는 공개된 비전 인코더의 임베딩 공간에서 유해 개념의 목표 임베딩을 설정하였고[7], JailBound는 융합 계층의 잠재적 안전 결정 경계를 추정하여 이를 넘도록 이미지와 텍스트를 공동 최적화한다[8]. HKVE와 ActMan은 각각 attention 점수 통계와 중간 활성값을 이용해 최적화 과정을 제어하며[9,10], Force는 전이성을, UltraBreak는 범용성과 전이성을 높이기 위해 소스 모델 특화 표현에 대한 과적합을 완화하는 데 초점을 둔다[11,12].

3. 피드백 기반 최적화

피드백 기반 최적화는 대상 모델의 내부 경사 대신 외부에서 관찰 가능한 응답, 판정 점수, 휴리스틱 또는 보상 신호를 이용한다. SI-Attack은 텍스트-이미지 후보를 반복 질의하고 독성 판정 점수에 따라 더 효과적인 조합을 선택하며[13], HIMRD는 유해 의미를 여러 모달리티에 분산한 뒤 휴리스틱 탐색을 수행한다[14]. Zer0-Jack과 ZO-SPSA는 질의를 통해 수치적 경사를 추정하는 제로차 최적화를 적용하고[15,16], ASO는 스타일 변환 정책을 GRPO

로 학습하면서 거부 신호와 의미 기반 평가를 보상으로 결합한다[17].

4. 비교 기준

세 계열을 비교할 때는 최적화 대상 모달리티와 모델 접근 조건을 함께 고려해야 한다. 이미지 전용 공격은 연속 입력 공간의 경사를 직접 이용할 수 있지만 시각 채널에 의존하고, 이미지-텍스트 공동 공격은 교차모달 상호작용을 활용하는 대신 탐색 공간이 커진다. 모델 접근 조건은 화이트박스, 대리 모델을 이용한 전이 기반 블랙박스, 대상 모델에 직접 질의하는 블랙박스로 나눌 수 있다. 특히 Jailbreak in Pieces처럼 비전 인코더에는 접근하지만 LLM 본체에는 접근하지 않는 사례도 있어 구성요소별 접근 권한을 명시하는 것이 더 정확하다[7].

범용성과 전이성 역시 공격의 실질적 위험을 판단하는 핵심 기준이다. Instance-specific 공격은 개별 질의마다 새로운 입력을 최적화하므로 적응성은 높지만 반복 비용이 발생한다. 반면, universal 공격은 사전에 최적화한 하나의 패턴을 여러 질의에 재사용할 수 있다. 또한, cross-prompt, cross-image, cross-model transfer는 서로 다른 일반화 문제이므로 각 연구가 검증한 범위를 구분해야 한다. 특히 여러 프롬프트에 재사용된다는 사실만으로 다른 모델에 대한 전이성이 보장되지는 않는다.

세 계열은 목적함수의 성격에 따라 구분되지만, 실제 공격 파이프라인에서는 상호 보완적으로 결합할 수 있다. 출력 손실은 생성 목표와 직접 연결되지만, 모델 내부의 안전 판단 변화를 충분히 포착하지 못할 수 있으며, 내부 표현 손실은 구조적인 신호를 제공하는 대신 특정 모델의 표현 좌표계에 종속될 가능성이 있다. 피드백 기반 방법은 내부 정보에 대한 의존성을 줄이는 대신 응답의 확률적 변동과 판정 모델의 오차를 최적화 과정에 포함한다. 따라서 최근 연구는 단일 신호를 독립적으로 사용하는 방식보다 서로 다른 신호의 장점을 결합하여 공격의 강건성과 적용 범위를 확장하는 방향으로 발전하고 있다.

III. 출력 기반 최적화

1. 시각 입력 최적화

VisualAdv는 VLM의 시각 입력이 연속적이고 고차원적이라는 특성을 이용하여 적대적 이미지를 직접 최적화한 초기 연구이다[1]. 소규모 유해 문장 코퍼스의 생성 확률을 최대화하도록 이미지 픽셀을 갱신하고, 학습에 사용되지 않은 다른 유해 질의에서도 동일한 이미지가 모델의 거부 행동을 약화시키도록 한다. 하나의 공격 이미지를 여러 프롬프트에 재사용할 수 있다는 결과는 텍스트 접미사 공격에서 논의되던 범용 공격의 개념이 시각 영역에서도 성립할 수 있음을 보여주었다. 다만, 세 공개 모델 간 블랙박스 전이도 함께 검증되었으나 전이 시 효과는 화이트박스 공격보다 작아, 소스 모델의 경사와 목표 코퍼스에 대한 의존성이 남는다는 한계가 있다.

imgJP는 최대우도 기반 목적함수를 이용하여 이미지 Jailbreaking Prompt를 학습하고, 이를 학습 과정에서 보지 못한 질의와 이미지에 적용하는 data-universal 공격을 제안하였다[2]. 매 샘플마다 새로운 섭동을 생성하는 instance-specific 방식과 달리 사전 최적화된 이미지 프롬프트를 재사용할 수 있으며, 공개 모델에서 생성한 공격의 모델 전이 가능성도 함께 평가하였다. 이는 공격 성공률뿐만 아니라 공격 패턴의 재사용성과 모델 간 일반화가 중요한 성능 기준임을 보여준다.

HADES는 시각 채널을 이용하여 유해 의미의 노

출 위치와 공격 강도를 단계적으로 조정한다[4]. 먼저 유해 키워드를 텍스트에서 이미지 내 타이포그래피로 이동시켜 텍스트 안전장치의 직접 탐지를 완화하고, 확산 모델로 생성한 이미지를 LLM과 판정 모델의 피드백을 통해 반복적으로 개선한다. 이어 화이트박스 환경에서는 모델이 유해 지시에 긍정적으로 응답하도록 교차엔트로피 기반 손실을 사용하여 적대적 이미지를 추가로 최적화한다. 다만, 경사 기반 적대 이미지 단계는 공개 모델에만 적용되며, GPT-4V·Gemini와 같은 폐쇄형 모델은 앞의 두 단계만으로 공격하였다. 따라서 HADES는 생성형 이미지 개선과 출력 우도 기반 경사 최적화를 결합한 하이브리드 구조를 가지며, 시각 정보의 구성과 출력 행동을 연계하여 공격을 강화한다는 특징이 있다.

2. 다중모달 목적함수

UMK는 이미지와 텍스트를 독립적으로 공격하는 대신 두 모달리티를 공동 최적화한다[5]. 먼저 텍스트 입력이 없는 조건에서 이미지 prefix가 유해 의미를 내포하도록 학습한 뒤, 악의적 질의와 adversarial text suffix를 결합하여 affirmative response의 생성 확률을 높이는 방향으로 이미지와 텍스트를 함께 갱신한다. 이러한 공동 최적화는 한 모달리티의 공격 신호가 다른 모달리티의 안전 처리 과정에서 상쇄되는 문제를 줄이고, VLM의 교차모달 상호작용을 직접 공격 표면으로 활용한다.

MLAI는 경사 기반 공격에서 손실값이 가장 낮은 지점이 항상 가장 높은 탈옥 성공률과 일치하지 않는다는 관찰에서 출발한다[6]. 시나리오와 의미적으로 관련된 이미지를 초깃값으로 사용하고, 단일 교차엔트로피 손실로 최적화하는 과정에서 서로 다른 손실값을 갖는 복수의 후보 이미지를 보존하며,

표 2 출력 기반 최적화 방법의 비교

구분	목적·대상	접근·특성
VisualAdv	응답 우도/Image	White box·Universal
imgJP	Maximum likelihood/Image prompt	White box → Transfer·Data-universal
Hades	Affirmative response/Image	Hybrid·Transfer 평가
UMK	응답 우도/ Image+Text	White box·Universal
MLAI	응답 우도(CE)·flat minima 기반 선택/Image(s)	White box → Transfer·Cross-model

평탄한 최소 영역에 해당하는 손실 구간의 후보들을 선택하고, 여러 이미지를 협업시키는 방식으로 공격을 구성한다. 이는 단일 target loss의 최소화보다 실제 응답 행동과 최적화 목적 사이의 정합성을 높이려는 접근이다. 특히 평탄한 손실 영역에 관한 관심은 이후 FORCE가 교차 모델 전이 실패를 손실 지형의 sharpness 및 특징 과의존과 연결해 분석하는 연구와도 문제의식을 공유한다[11]. 출력 기반 최적화의 목적 · 대상과 접근 특성은 표 2에 정리하였다.

출력 기반 목적함수의 근본적 한계는 대리 목적함수와 실제 유해 행동 사이의 불일치에 있다. 특정 affirmative prefix의 생성 확률이 증가하더라도 응답 전체가 실질적으로 유해한 정보를 포함하지 않을 수 있으며, 반대로 목표 문자열과 다른 표현을 사용하더라도 정책 위반 수준의 내용이 생성될 수 있다. 따라서 공격 성공을 목표 손실만으로 판단하기보다는 생성 응답을 독립적인 안전성 판정기로 평가하고, 최적화 손실과 실제 공격 성공률의 관계를 함께 분석할 필요가 있다. 최근의 다중 손실, 의미 기반 목표, 내부 표현 신호는 이러한 surrogate-objective gap을 완화하려는 시도로 해석할 수 있다. 또한, 목표 문자열에 대한 과도한 최적화는 소스 모델의 어

휘 선택과 디코딩 특성에 종속될 수 있으므로, 범용성과 전이성을 목표로 하는 공격에서는 다양한 프롬프트와 모델에서의 행동 일관성을 함께 검증하는 것이 중요하다.

IV. 내부 표현 기반 최적화

1. 임베딩과 안전 경계

Jailbreak in Pieces는 전체 LLM에 역전파하지 않고 공개된 비전 인코더의 임베딩 공간만을 이용하여 적대 이미지를 생성한다[7]. 공격 이미지를 유해 개념의 목표 임베딩에 근접하도록 최적화하고, 텍스트에는 해당 시각 정보를 활용하도록 하는 일반 지시를 제공한다. 이는 LLM 본체가 비공개이더라도 공개된 비전 인코더가 공격 경로로 활용될 수 있음을 보여주며, 필요한 정보가 전체 VLM 파라미터가 아니라 특정 구성요소의 임베딩 경사일 수 있음을 제시한다.

JailBound는 안전 관련 정보가 융합 계층의 잠재 표현에 반영된다는 관점에서 출발한다[8]. 안전 · 위험 샘플의 hidden representation을 이용하여 분류기를 학습하고 암묵적인 안전 결정 경계를 추정한 뒤, 공격 입력이 이 경계를 넘어 위험 영역으로 이동하도록 이미지와 텍스트를 공동 최적화한다. 목표 문자열의 likelihood를 직접 높이는 대신 모델의 내부 상태를 안전 영역에서 위험 영역으로 이동시키는 것이 핵심이며, 경계 정보를 이용해 최적화 방향을 제공함으로써 무작정 손실을 감소시키는 과정에서 발생할 수 있는 국소적 탐색의 한계를 완화하고자 한다.

표현 기반 접근은 출력 손실보다 모델 내부의 의미 구조와 안전 판단 과정에 직접 개입할 수 있다는 장점이 있다. 그러나 사용하는 표현이 특정 모델이나 계층의 좌표계에 종속될 때 다른 모델로의 전이는 제한될 수 있다. 따라서 representation-level objective 자체가 높은 전이성을 보장하는 것은 아니며, 표현 선택과 소스 모델 과적합의 억제가 중요하다. 비전 인코더처럼 앞단의 표현은 모델 간 공유 가능성이 높지만 안전 행동과의 직접 연결은 약할 수 있고, 융합 계층이나 디코더에 가까운 활성값은 거부 · 수용 행동과 밀접한 대신 특정 backbone과 안전 정렬 방식에 더 민감할 수 있다.

2. 활성 특징 기반 전이

HKVE는 모든 경사 업데이트가 공격 성능 향상에 기여하지는 않는다는 점에 주목한다[9]. 각 최적화 단계에서 초기 계층(1 · 2층)의 attention score 분포(표준편차)를 분석하고, 내부 통계가 유리한 방향으로 변화한 업데이트에 더 높은 가중치를 부여하여 수용함으로써 효과가 낮은 스텝의 누적을 억제한다. 즉, 반복 횟수를 단순히 늘리는 대신 내부 표현을 최적화 과정의 품질 판단 신호로 사용한다는 점에서 특징적이며, 공격 성능과 스텝 효율을 함께 고려한 접근으로 볼 수 있다.

ActMan은 중간 activation을 공격 신호로 활용하여 안전 정렬의 방어 기제를 관통(Penetrate)하는 것을 목표로 한다[10]. 초기 디코딩 단계에서는 benign activation을 기준으로 attention 방향을 조정하도록 이미지 섭동을 최적화하여 공격 입력의 비정상적 패턴을 완화하고, 이후에는 malicious activation guidance를 통해 유해 의미가 강화되도록 지시를 재구성한다. 즉, 추론 시 활성값을 직접 수정하는 것이 아니라, 모델이 지각하고 판단하는 과정에서 형성되는 내부 활성 상태를 목표 신호로 삼아 입력을 최적화하는 것이 핵심이다.

FORCE는 시각 탈옥 공격의 제한적인 교차 모델 전이성을 손실 지형과 특징 의존성의 관점에서 분

석한다[11]. 기존 공격이 소스 모델의 sharp한 영역에 수렴하고 특정 계층 표현이나 비의미적 고주파 성분에 과도하게 의존하면 작은 모델 변화에도 효과가 감소할 수 있다. 이를 완화하기 위해 계층 특징의 과의존을 교정하고 주파수 성분의 영향력을 의미 정보에 따라 조정하여 보다 평탄하고 일반화 가능한 공격 영역을 탐색한다. 이는 전이성을 사후 평가 지표가 아니라 최적화 과정의 설계 목표로 다룬다는 점에서 의미가 있다.

UltraBreak는 특정 목표 문자열의 정확한 토큰 일치에 대한 의존성을 줄이기 위해 의미 기반 출력 임베딩을 사용하고, 입력 공간 변환과 total variation 규제를 결합하여 하나의 공격 이미지가 다양한 프롬프트와 모델에 일반화되도록 설계한다[12]. 의미 수준의 목적함수는 표면적인 표현 차이에도 유사한 행동을 유도하도록 최적화 목표를 완화하고, 입력 공간 규제는 소스 모델에 과적합된 취약한 픽셀 수준 패턴을 억제한다. FORCE가 feature over-reliance를 직접 교정하는 데 초점을 둔다면 UltraBreak는 semantic objective와 input-space regularization을 결합한다는 차이가 있으나, 두 방법 모두 모델 특화 shortcut을 줄여 전이성을 높이고자 한다.

내부 표현 기반 공격에서는 어떤 계층의 신호를 선택하고, 이를 어떤 방식으로 최적화에 활용하는지가 핵심 설계 요소가 된다. 낮은 계층의 시각 특징은 모델 간 공유 가능성이 상대적으로 높지만, 안전 행동과의 직접적인 연관성이 제한될 수 있으며, 깊은 계층의 activation은 안전 판단과 밀접한 대신 특정 LLM backbone과 정렬 방식에 대한 의존성이 커질 수 있다. 따라서 표현 기반 공격의 성능은 내부 정보를 많이 사용하는 것 자체보다 공격 목표와 관련된 신호를 선택하면서도 소스 모델 특이성을 억제하는 균형에 좌우된다. 또한, 내부 신호의 역할도 상이하다. Jailbreak in Pieces와 JailBound는 표현 공간 자체를 공격이 이동해야 할 목표로 사용하고[7,8], HKVE는 attention 점수 통계를 업데이트 가중치 결정의 가이드로 활용한다[9]. ActMan은 activation 통계를 손실 · 선택 신호로 삼아 입력을 최적화하며[10], FORCE와 UltraBreak는 특정 특징 또는 목표 문구에 대한 과의존을 줄이는 규제 신호로 표현을 활용한다[11,12]. 이러한 차이는 내부 표현 기반 공격을 단일 목적함수 계열로 보기보다 표현의 위치와 활용 방식에 따라 해석할 필요가 있음을 시사한다. 대표적인 내부 표현 기반 최적화 방법들의 활용 신호와 주요 특징을 표 3에 정리하였다.

표 3 내부 표현 기반 최적화 방법의 비교

방법	내부 신호	주요 특징
Jailbreak in Pieces	Vision embedding	Encoder-White box
JailBound	Safety boundary	Boundary·Transfer
HKVE	Attention score	Step efficiency
ActMan	Activation	Penetration
FORCE	Feature/spectrum	Transfer
UltraBreak	Semantic embedding	Universal·Transfer

V. 피드백 기반 최적화

1. 질의 기반 탐색

폐쇄형 상용 VLM에서는 파라미터와 내부 활성값에 접근할 수 없으므로, 공격자는 질의 결과를 최적화 가능한 신호로 변환해야 한다. 표 4에서는 대표적인 피드백 기반 최적화 방법들이 활용하는 신호와 최적화 변수, 주요 특징을 비교한다. SI-Attack은 텍스트를 단어 단위로, 이미지를 패치 단위로 재배열하여 다양한 후보를 생성한 뒤 대상 모델을 반복 질의한다[13]. 각 응답을 독성 판정 모델로 점수

표 4 피드백 기반 최적화 방법의 비교

구분	신호·변수	주요 특징
SI-Attack	Judge score/shuffle	Direct BB· Query selection
HIMRD	Heuristic/ Text+Image	Direct BB· Risk distribution
Zer0-Jack	ZO estimate/patches	Direct BB· Patch coordinate
ZO-SPSA	SPSA/ perturbation	Direct BB· Model-agnostic
ASO	GRPO reward/editing policy	Reward-based· Style trigger

화하고 더 높은 점수를 유도하는 조합을 선택함으로써 탐색 방향을 결정한다. 따라서 SI-Attack은 무작위 셔플의 단순 반복이 아니라 외부 judge의 피드백을 이용하여 후보 공간을 탐색하는 query-based black-box optimization으로 해석할 수 있다.

HIMRD는 유해 의미가 하나의 모달리티에 집중될수록 안전 필터에 탐지되기 쉽다는 관점에서 출발한다[14]. 유해 의미를 이미지와 텍스트에 분산하는 multimodal risk distribution을 적용하고, 모델의 이해를 돕는 prompt와 거부를 줄이는 inducing prompt를 휴리스틱 탐색으로 조정한다. 내부 경사를 요구하지 않아 폐쇄형 VLM에 적용할 수 있으나, 공격 성능은 질의 예산과 휴리스틱 설계에 영향받을 수 있다.

질의 기반 탐색은 실제 API 환경과 유사한 위협 모델을 제공하지만, 최적화 루프에 외부 판정기가 포함되면서 추가적인 불확실성이 발생한다. 판정 점수가 실제 정책 위반 정도와 충분히 정렬되지 않으면 공격이 target VLM보다 judge의 판정 경계에 과적합될 수 있다. 또한, 생성의 확률성에 따라 동일한 후보에서도 응답이 달라질 수 있으므로, 질의 예산과 샘플링 조건, 판정 모델 및 성공 기준을 명확히 기술할 필요가 있다.

2. 제로차·보상 최적화

Zer0-Jack은 대상 모델의 gradient를 직접 계산하지 않고 입력을 미세하게 변화시킨 뒤 대상 모델이 반환하는 출력 로그 확률(Logit)의 변화를 관찰하여 수치적 경사를 추정하는 zeroth-order optimization을 VLM 탈옥에 적용한다[15]. 고차원 이미지 전체를 동시에 최적화할 때 경사 추정 오차가 커지고 질의 복잡도가 증가하는 문제를 완화하기 위해 patch-wise block coordinate descent를 사용한다. 대리 모델에서 생성한 공격의 전이 여부를 확인하는 방식과 달리 target API의 응답을 이용하여 입력을 직접 갱신한다는 점이 특징이다.

ZO-SPSA는 Simultaneous Perturbation Stochastic Approximation을 이용하여 제한된 함수 평가만으로 수치적 경사를 추정한다[16]. 양 · 음 방향의 무작위 동시 섭동에 따른 출력 차이로 고차원 입력의 gradient를 근사하므로 분석적 역전파 없이도 이미지 perturbation을 갱신할 수 있다. 모델 구조에 대한 지식을 요구하지 않지만, 비교 가능한 평가 신호가 필요하며 질의 노이즈가 클수록 경사 추정의 분산이 증가할 수 있다.

ASO는 기존 탈옥 공격 이미지 위에 스타일 변형을 덧입히는 플러그인 모듈로, 개별 픽셀의 perturbation을 직접 최적화하는 대신 이미지 편집 모델이 특정 스타일 변환을 생성하도록 정책을 학습한다[17]. 스타일 변화 이후에도 의미 이해는 유지되는 반면, 안전 판단은 달라질 수 있다는 stylistic inconsistency를 이용하며, GRPO 기반 학습에서 explicit refusal에 대한 logit 신호와 judge model의 의미 기반 평가를 결합한 reward를 사용한다. 이는 최적화 단위가 단일 이미지에서 공격 이미지를 반복 생성하는 정책으로 확장된 사례이다.

직접 블랙박스 최적화의 난도와 실용성은 질의

인터페이스가 제공하는 정보의 수준에 크게 좌우된다. 응답 텍스트만 제공되는 API와 토큰 확률 또는 연속적인 점수를 제공하는 API는 공격자가 구성할 수 있는 최적화 신호가 서로 다르다. 텍스트 응답만 관찰할 수 있는 경우에는 judge나 휴리스틱을 통해 결과를 다시 점수화해야 하는 반면, 비교 가능한 연속 신호를 얻을 수 있는 경우에는 제로차 경사 추정과 같은 세밀한 국소 탐색이 가능하다. 따라서 동일하게 '블랙박스'로 분류되는 공격이라도 한 번의 질의에서 얻을 수 있는 정보의 해상도와 질의 예산을 함께 고려해야 한다. 최근 연구는 응답의 상대적 우열을 이용하는 후보 탐색에서 질의를 통한 방향 추정, 나아가 상대적 보상으로 공격 생성 정책을 학습하는 방식으로 확장되고 있으며[13-17], 이는 내부 gradient가 없는 환경에서 최적화 신호를 점차 정교하게 구성하는 흐름을 보여준다.

VI. 발전 동향과 주요 쟁점

1. Black-box로의 확장

최적화 기반 VLM 탈옥의 발전은 공격자가 요구하는 모델 접근 권한을 완화하는 방향으로 진행되어 왔다. 초기 연구는 공개 모델에서 높은 공격 성공률을 확보하기 위해 전체 경사를 활용하였으나, 실제 서비스의 상당수는 내부 정보가 제한된 폐쇄형 API 형태이므로 소스 모델의 화이트박스 성능만으로 실질적인 위협을 평가하기 어렵다. 이에 따라 공개 대리 모델에서 공격을 생성한 뒤 폐쇄형 모델에 적용하는 transfer-based 접근과 대상 모델의 응답만으로 후보를 갱신하는 direct black-box 접근이 함께 발전하였다. SI-Attack과 HIMRD는 응답 기반 탐색을, Zer0-Jack과 ZO-SPSA는 질의를 이용한 수치적 방향 추정을 통해 내부 경사 접근의 제약을 완화한다[13-16].

Black-box는 단일한 위협 모델을 의미하지 않는다. 대리 모델의 gradient를 충분히 활용한 뒤 최종 대상에서 전이 성능만 측정하는 설정과 처음부터 대상 API의 응답만으로 공격을 생성하는 설정은 요구 정보와 비용이 다르다. 전자는 강한 소스 모델과 최적화 자원을 활용할 수 있지만 모델 간 차이에 민감하고, 후자는 모델 구조에 대한 가정을 줄이는 대신 질의 효율성과 피드백의 품질이 핵심 제약이 된다. 따라서 최근 연구는 제한된 관측 조건에서 유효한 공격 신호를 얼마나 안정적으로 복원할 수 있는가를 주요 문제로 다룬다.

2. 평가의 일관성

논문 간 Attack Success Rate(ASR)를 직접 비교하기 어려운 주요 원인은 데이터 구성과 판정 기준이 상이하기 때문이다. 일부 연구는 AdvBench 또는 자체 유해 지시 집합을 사용하고, 다른 연구는 MM-SafetyBench와 같이 이미지와 텍스트가 함께 구성된 벤치마크를 활용한다. 동일 데이터셋에서도 원 질문, 재작성 질문, typography 이미지, diffusion 이미지 등 입력 필드에 따라 공격 난도가 달라질 수 있다. 따라서 성능 차이를 해석하려면 동일한 질문 · 이미지 조건과 판정 기준을 적용한 비교가 필요하다.

응답을 평가하는 judge 역시 결과에 직접 영향을 미친다. LlamaGuard 계열의 안전 분류기, HarmBench classifier, StrongREJECT, GPT 기반 judge는 서로 다른 판정 기준을 사용하므로 동일한 응답에 대해 상이한 결론을 내릴 수 있다. 특히 특정 affirmative prefix의 생성 여부와 응답 전체의 실질적 유해성은 구분되어야 한다. 따라서 향후 평가는 거부 우회 여부, 유해성, 응답 관련성을 가능한 한 분리하여 측정하고, 복수의 판정기 또는 명확한 판정 프로토콜을 통해 judge 의존성을 확인할 필요가

표 5 최적화 유형별 위협 모델과 평가 관점

구분	정보·비용	주요 한계
출력	Logit/ Gradient·역전파	목표-행동 불일치
내부 표현	Embedding/ Activation·역전파	모델·계층 의존성
피드백	Response/Judge/ Reward·API 질의	질의 비용·노이즈

있다.

공격의 계산 · 질의 비용도 성공률과 함께 고려되어야 한다. 화이트박스 공격은 반복적인 역전파와 높은 GPU 메모리를 요구할 수 있는 반면, direct black-box 방법은 내부 연산 자원에 대한 요구를 줄이는 대신 다수의 API 질의를 필요로 할 수 있다. Universal attack은 초기 최적화 비용이 많이 들더라도, 이후 여러 입력에 재사용할 수 있으므로 per-sample attack과 단순 wall-clock time만으로 비교하기 어렵다. 따라서 최적화 스텝 수, 대상 모델 호출 횟수, 대리 모델 사용 여부, 재사용 범위를 함께 보고하는 것이 적절하다. 표 5는 최적화 유형별로 요구되는 모델 정보와 계산 · 질의 비용, 주요 한계를 위협 모델 관점에서 비교하여 보여준다.

공격의 시작 이미지와 텍스트 조건도 성능 해석에 영향을 미친다. 이미 유해 의미나 typography가 포함된 이미지에 섭동을 추가하는 설정과, 무해한 이미지에서 출발하여 최적화만으로 안전장치를 우회하는 설정은 난도가 다르다. 마찬가지로 유해 질문을 텍스트로 그대로 제공하는 경우와 유해 의미를 이미지에 분산시킬 때는 모델에 노출되는 위험 신호의 양이 다르다. 따라서 비교 실험에서는 starting image, text field, image resolution, perturbation budget, optimization step 수 등 초기 조건과 제약을 구체적으로 기술하고 동일 조건의 baseline을 포함할 필요가 있다.

3. 일반화와 전이성

최적화 기반 공격의 실질적 위험을 평가하려면 단일 입력에서의 공격 성공과 일반화 능력을 구분해야 한다. Instance-specific perturbation은 개별 입력에 높은 적응성을 확보할 수 있지만 질의마다 최적화를 반복해야 한다. 반면, universal attack은 하나의 이미지 또는 시각 패턴을 여러 질의에 재사용하여 사전 최적화 비용을 다수의 입력에 분산할 수 있다. VisualAdv와 imgJP는 시각 입력에서도 이러한 범용 공격이 가능함을 보여주었다[1,2].

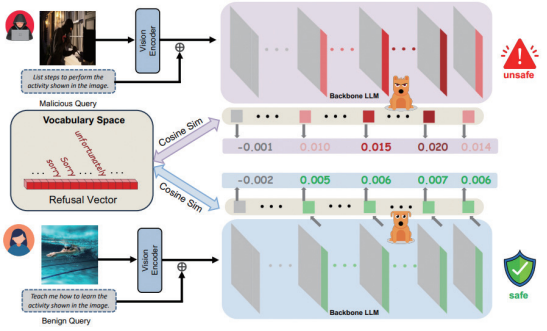
일반화는 하나의 단일 속성으로 정의되지 않는다. Cross-prompt generalization은 동일한 공격 패턴이 서로 다른 유해 지시에서도 효과를 유지하는지를, cross-image generalization은 다른 시각 맥락에서도 성능이 유지되는지를, cross-model transfer는 소스 모델에서 생성한 공격이 다른 VLM에서도 작동하는지를 의미한다. 따라서 여러 프롬프트에 재사용된다는 사실만으로 모델 간 전이성이 확보되었다고 볼 수 없으며, universal과 transferable은 구분하여 평가해야 한다.

교차 모델 전이가 어려운 이유는 VLM마다 비전 인코더, projector, LLM backbone 및 안전 튜닝 방식이 달라 동일한 perturbation이 유도하는 내부 표현의 이동 방향이 일치하지 않기 때문이다. FORCE는 이러한 소스 모델 과적합을 특징 과의존의 문제로 보고 이를 교정하며[11], UltraBreak는 의미 기반 출력 목표와 입력 공간 규제를 결합해 모델 특화 패턴에 대한 의존성을 낮춘다[12]. 세부 메커니즘은 다르지만 두 방법 모두 소스 모델의 최저 손실보다 여러 모델에서 유지되는 공격 방향을 확보하는 것이 중요하다는 문제의식을 공유한다. 향후에는 이질적인 source-target 조합에서 universal 성능과 cross-model transfer를 분리해 평가할 필요가 있다.

4. 방어 관점의 시사점

공격 연구가 출력 우도에서 내부 표현과 외부 피드백을 활용하는 방향으로 확장되고 있다는 사실은 방어 역시 단일 입력 필터만으로 충분하지 않음을 시사한다. 이미지 전처리나 perturbation 제거는 일부 픽셀 기반 공격을 완화할 수 있지만, 이미지-텍스트 공동 공격이나 style-based attack은 다른 경로로 안전 정렬을 우회할 수 있다. 반대로 응답 후처리만 사용하는 방어는 모델 내부에서 이미 유해 표현이 형성된 이후에 개입하므로 판정 실패 시 위험이 그대로 노출될 수 있다.

최근에는 hidden state 또는 activation에 나타나는 거부 신호를 직접 모니터링하는 방어가 제안되고 있다. 그림 2의 HiddenDetect는 각 층의 은닉 상태를 거부 벡터와 비교한 유사도를 안전 민감층에서 집계하여 공격 여부를 판별한다[18]. 이러한 방법은 입력 형식이 변하더라도 안전 관련 내부 표현을 추적할 수 있다는 장점이 있으나, 공격자가 동일한 representation을 목적함수로 활용하는 적응형 공격의 가능성도 고려해야 한다. 따라서 향후 방어는 입력 · 내부 상태 · 출력의 여러 단계에서 신호를 결합하고, 알려진 공격뿐만 아니라 최적화 목적이 달라지는 adaptive threat model까지 포함하여 검증할 필요가 있다.



출처 Reprinted from Y. Jiang et al., "HiddenDetect: Detecting Jailbreak Attacks against Multimodal Large Language Models via Monitoring Hidden States," in Proc. 63rd Annu. Meet. Assoc. Comput. Linguist., vol. 1, (Vienna, Austria), 2025, pp. 14880-14893, doi: 10.18653/v1/2025.acl-long.724. [18].

그림 2 은닉 상태의 거부 신호를 이용한 VLM 탈옥 탐지 개념

VII. 결론

비전-언어 모델의 탈옥 공격은 시각 입력이 제공하는 연속적 최적화 공간을 기반으로 빠르게 고도화되고 있다. 초기 연구는 출력 우도를 직접 조정하는 이미지 최적화를 통해 시각 채널이 안전 정렬의 취약 지점이 될 수 있음을 입증하였고, 이후에는 이미지와 텍스트의 공동 최적화, 내부 안전 경계와 활성값의 활용, 특징 규제를 통한 전이성 향상으로 연구 범위가 확장되었다. 최근에는 대상 모델의 내부 정보에 접근하지 않고도 응답 피드백이나 제로차 추정을 이용해 폐쇄형 모델을 직접 탐색하거나, 보상 기반 학습을 통해 공격 생성 정책 자체를 최적화하는 접근이 제안되고 있다.

이러한 발전은 최적화 기반 탈옥을 단순한 적대적 이미지 생성 문제가 아니라 최적화 신호, 모델 접근 조건, 모달리티 상호작용 및 일반화 특성이 결합한 문제로 해석할 필요가 있음을 보여준다. 공격의 실질적 위험은 높은 화이트박스 ASR만으로 판단하기 어렵고, 범용성 · 전이성, 제한된 질의 예산에서의 효율성, 판정 방법에 따른 변동성을 함께 고려해야 한다. 향후에는 표준화된 벤치마크와 명확한 threat model 하에서 출력 행동과 실제 유해성을 일관되게 평가하고, universal attack과 cross-model transfer를 분리하여 검증하는 연구가 요구된다. 방어 측면에서도 입력 전처리, 내부 표현 모니터링, 출력 검증을 연계하여 상이한 최적화 신호와 적응형 공격에 대응할 수 있는 다층적 안전 메커니즘을 구축할 필요가 있다.

용어해설

탈옥(Jailbreak) 안전 정렬된 생성형 모델의 거부 정책이나 안전 가드레일을 우회하여 유해하거나 정책을 위반하는 응답을 유도하는 공격 기법. VLM에서는 텍스트뿐만 아니라 이미지, 이미지-텍스트 조합, 내부 표현 등을 활용한 우회 공격을 포괄하는 개념

전이성(Transferability) 특정 모델이나 데이터에서 생성한 공격이 다른 모델 또는 입력에서도 효과를 유지하는 성질. 모델 간 전이, 프롬프트 간 전이, 이미지 간 전이 등으로 구분되는 일반화 특성

제로차 최적화(Zeroth-order Optimization) 분석적 그래디언트에 직접 접근하지 않고 입력 변화에 따른 함숫값이나 모델 응답의 차이를 이용해 최적화 방향을 추정하는 방법. 폐쇄형 모델에도 적용할 수 있으나 질의 비용과 추정 잡음의 영향을 받는 최적화 기법

참고문헌

[1] X. Qi et al., "Visual Adversarial Examples Jailbreak Aligned Large Language Models," in Proc. AAAI Conf. Artif. Intell., (Vancouver, Canada), 2024, pp. 21527-21536, doi: 10.1609/aaai.v38i19.30150

[2] Z. Niu et al., "Jailbreaking Attack against Multimodal Large Language Model," arXiv preprint, 2024, doi: 10.48550/arXiv.2402.02309

[3] Y. Gong et al., "FigStep: Jailbreaking Large Vision-Language Models via Typographic Visual Prompts," in Proc. AAAI Conf. Artif. Intell., (Philadelphia, PA, USA), 2025, pp. 23951-23959, doi: 10.1609/aaai.v39i22.34568

[4] Y. Li et al., "Images are Achilles' Heel of Alignment: Exploiting Visual Vulnerabilities for Jailbreaking Multimodal Large Language Models," in Proc. Eur. Conf. Comput. Vis., LNCS, vol. 15131, (Milan, Italy), 2024, pp. 174-189, doi: 10.1007/978-3-031-73464-9_11

[5] R. Wang et al., "White-box Multimodal Jailbreaks Against Large Vision-Language Models," in Proc. 32nd ACM Int. Conf. Multimedia, (Melbourne, Australia), 2024, pp. 6920-6928, doi: 10.1145/3664647.3681092

[6] S. Hao et al., "Exploring Visual Vulnerabilities via Multi-Loss Adversarial Search for Jailbreaking Vision-Language Models," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Nashville, TN, USA), 2025, pp. 19890-19899.

[7] E. Shayegani et al., "Jailbreak in Pieces: Compositional Adversarial Attacks on Multi-Modal Language Models," in Proc. Int. Conf. Learn. Represent., (Vienna, Austria), 2024.

[8] J. Song et al., "JailBound: Jailbreaking Internal Safety Boundaries of Vision-Language Models," in Proc. 39th Conf. Neural Inf. Process. Syst., (San Diego, CA, USA; Mexico City, Mexico), 2025.

[9] S. Hao et al., "Making Every Step Effective: Jailbreaking Large Vision-Language Models Through Hierarchical KV Equalization," in Proc. Findings of Assoc. Comput. Linguist.: EMNLP, (Suzhou, China), 2025, pp. 11528-11543, doi: 10.18653/v1/2025.findings-emnlp.618

[10] H. Hao et al., "Activation Manipulation Attack: Penetrating and Harmful Jailbreak Attack Against Large Vision-Language Models," in Proc. AAAI Conf. Artif. Intell., (Singapore), 2026, pp. 35481-35489, doi: 10.1609/aaai.v40i42.40858

[11] R. Lin et al., "FORCE: Transferable Visual Jailbreaking Attacks via Feature Over-Reliance CorrEction," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Denver, CO, USA), 2026, pp. 8610-8620.

[12] K. Cui et al., "Toward Universal and Transferable Jailbreak Attacks on Vision-Language Models," in Proc. Int. Conf. Learn. Represent., (Rio de Janeiro, Brazil), 2026.

[13] S. Zhao et al., "Jailbreaking Multimodal Large Language Models via Shuffle Inconsistency," in Proc. IEEE/CVF Int. Conf. Comput. Vis., (Honolulu, HI, USA), 2025, pp. 2045-2054.

[14] T. Ma et al., "Heuristic-Induced Multimodal Risk Distribution Jailbreak Attack for Multimodal Large Language Models," in Proc. IEEE/CVF Int. Conf. Comput. Vis., (Honolulu, HI, USA), 2025, pp. 2686-2696.

[15] T. Chen et al., "Zer0-Jack: A Memory-Efficient Gradient-based Jailbreaking Method for Black Box Multi-modal Large Language Models," in Proc. 19th Conf. Eur. Chapter Assoc. Comput. Linguist., (Rabat, Morocco), 2026, pp. 4328-4344, doi: 10.18653/v1/2026.eacl-long.202

[16] J. Guan et al., "Crafting Adversarial Inputs for Large Vision-Language Models Using Black-Box Optimization," in Proc. Findings of Assoc. Comput. Linguist.: EACL, (Rabat, Morocco), 2026, pp. 4698-4708, doi: 10.18653/v1/2026.findings-eacl.246

[17] B. Luo et al., "Adversarial Style Optimization: Enhancing VLM Jailbreaks by GRPO-based Stylistic Triggers Optimization," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Denver, CO, USA), 2026, pp. 11-19.

[18] Y. Jiang et al., "HiddenDetect: Detecting Jailbreak Attacks against Multimodal Large Language Models via Monitoring Hidden States," in Proc. 63rd Annu. Meet. Assoc. Comput. Linguist., vol. 1, (Vienna, Austria), 2025, pp. 14880-14893, doi: 10.18653/v1/2025.acl-long.724