EVOLUTIONARY ADVERSARIAL ATTACKS MEET LLMS: SEMANTIC-AWARE ADAPTIVE OPTIMIZATION AGAINST VISION-LANGUAGE PRE-TRAINED MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Targeted attacks against black-box vision-language models remain challenging because transfer-based optimization must simultaneously achieve precise semantic steering, cross-model transferability, and bounded image perturbations without access to victim gradients or internal representations. Although evolutionary algorithms provide a natural gradient-free solution, their search is typically semantically blind and relies on fixed optimization strategies. We introduce **LEA**, an LLM-driven evolutionary adversarial framework that turns conventional evolutionary search into a *semantic-aware adaptive optimization process*. LEA first constructs original and target semantic anchors to explicitly characterize the desired semantic transition. More importantly, rather than directly generating adversarial perturbations, LEA employs the LLM as a high-level controller that interprets the evolving search state and adaptively coordinates semantic region priors, operator selection, and search parameters. A heterogeneous evolutionary operator pool then performs the low-level gradient-free perturbation search, enabling complementary global exploration and local refinement. Experiments on SMALLCAP, ViECap, and LLaVA-13B demonstrate consistent improvements over representative black-box and transfer-based baselines. On SMALLCAP, LEA improves Target ASR from 39.7% to 53.2% over the strongest targeted baseline while increasing the average CLIP score from 63.78 to 68.76. These results show that high-level LLM reasoning can effectively control evolutionary adversarial search and expose the vulnerability of vision-language systems to targeted semantic manipulation under transfer-based black-box settings.

1 INTRODUCTION

*V*ision-*L*anguage *P*re-trained (VLP) models have become a fundamental paradigm for cross-modal intelligence. By learning aligned visual and textual representations from large-scale image–text corpora, they have achieved strong performance across a broad range of tasks, including image-text retrieval, image captioning, visual question answering, and multimodal reasoning. With the rapid development of multimodal large language models, VLP-based systems are increasingly deployed in real-world applications such as search engines, accessibility assistance, content moderation, and interactive agents (Chen et al., 2023). Their growing deployment, however, also raises important security and reliability concerns.

In practical applications, VLP systems are commonly accessible only through public APIs or closed-source interfaces, where users can observe model outputs but cannot access model architectures, parameters, gradients, or intermediate representations. We therefore consider a *transfer-based black-box setting*, in which adversarial examples are optimized using a publicly available surrogate model and are subsequently transferred to unseen victim models for output-only evaluation. This restricted-access setting poses a realistic yet challenging threat model for adversarial evaluation. Existing studies have shown that carefully crafted visual perturbations can induce incorrect cross-modal matching, misleading captions, or abnormal multimodal responses (Li et al., 2025; Liu et al., 2026). For targeted attacks, the challenge is even more demanding: the adversarial examples must not only disrupt the original prediction, but also steer the victim model toward a user-specified target semantic description. Achieving such controllable semantic manipulation under restricted black-box access

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.

is difficult because the attack must be optimized without victim-side gradients or internal feedback while simultaneously maintaining semantic controllability and bounded visual distortion.

Evolutionary computation provides a natural optimization paradigm for such black-box attacks because it does not rely on gradient information and can directly search over complex, non-differentiable, and multi-objective spaces (Huang et al., 2026). However, conventional evolutionary attacks are typically driven by low-level operators, such as mutation, crossover, or population-based updates, without explicitly understanding the semantic relationship between the original input, the target concept, and the intermediate adversarial states. Consequently, the search process is often semantically blind, leading to inefficient exploration, unstable convergence, and limited target controllability in high-dimensional cross-modal spaces. Moreover, fixed operator schedules may not be equally effective at different optimization stages, making it difficult to balance global exploration, local refinement, semantic alignment, and perturbation stealthiness.

Promisingly, *L*arge *L*anguage *M*odels (LLMs) offer a way to overcome this limitation. Beyond language generation, modern LLMs exhibit strong semantic reasoning, abstraction, and decision-making capabilities, which make them suitable as high-level controllers for heuristic optimization. Intuitively, an LLM can analyze the semantic gap between the current adversarial state and the target, identify semantically relevant regions, and dynamically adjust the search strategy according to optimization feedback. This guides us to combine LLMs with evolutionary computation in a different fashion: an LLM acts as a *semantic-aware adaptive controller*, while evolutionary operators remain responsible for low-level perturbation search. Such cooperation can empower gradient-free global searching with semantic reasoning.

Motivated by this insight, we propose **LEA**, an **LLM-driven Evolutionary Adversarial Attack** framework for targeted black-box attacks against VLP models. LEA first constructs an original semantic anchor from the clean image–text pair, and another anchor from the pair of the refined target text and its diffusion-generated image. These two semantic anchors present an explicit optimization direction: the adversarial representation is encouraged to approach the target semantics while moving away from the original semantics. In the semantic-based schema, LEA introduces an LLM-guided adaptive evolutionary controller that dynamically provides region-of-interest priors, operator-selection probabilities, parameter recommendations, and objective-balancing weights according to the optimization state. A hybrid operator pool, including GA, MFO, PSO, TLBO, and local refinement, is further employed to combine complementary exploration and exploitation behaviors. In this way, LEA facilitates conventional blind evolutionary search into an optimization process that jointly considers semantic alignment, original-semantic separation, and perturbation stealthiness.

Our main contributions are summarized as follows:

- We propose **LEA**, a targeted black-box adversarial attack framework that enhances evolutionary algorithms with a Large Language Model (LLM), empowering heuristic search with semantic-aware reasoning.
- We build semantic-preserving anchors to guide image-space perturbations toward attacker-specified targets with bounded visual distortions. We further treat an LLM as an adaptive controller that provides region priors, operator scheduling, parameter adjustment, and objective balancing to reform evolutionary algorithms.
- We conducts extensive experiments on multiple VLP/VLM models to validate LEA in terms of attack success, semantic alignment, cross-model transferability, and computational trade-offs.

2 RELATED WORK

2.1 BLACK-BOX AND TRANSFERABLE ATTACKS ON VLP/VLMS

Recent surveys have systematically documented the rapidly expanding attack surface of large vision-language models, covering adversarial visual perturbations, multimodal manipulation, and transferable black-box attacks (Liu et al., 2025). Co-Attack (Zhang et al., 2022) investigated joint visual-textual attacks, while SGA (Lu et al., 2023) and VLATTACK (Yin et al., 2023) focused on improving cross-model transferability. More recent methods such as AnyAttack (Zhang et al., 2025) and CoA (Xie et al., 2025) further explored targeted transferable attacks, whereas SceneTAP (Cao

et al., 2025) demonstrated that high-level semantic planning can facilitate realistic VLM attacks. Robustness evaluations on large and closed-source VLMs further confirm their susceptibility to transferred adversarial inputs (Dong et al., 2023; Zhao et al., 2023; Tu et al., 2024).

Despite these advances, most existing attacks rely on predefined perturbation transformations or static semantic-update rules. They provide limited mechanisms for adapting the search strategy according to semantic progress and intermediate optimization states. LEA instead introduces a semantic-aware controller to dynamically coordinate black-box evolutionary search.

2.2 LLM-guided Evolutionary Optimization

Evolutionary computation is well suited to gradient-free and multi-objective black-box optimization, but conventional evolutionary algorithms primarily rely on predefined low-level operators and lack explicit semantic reasoning. Recent studies have explored integrating LLMs with evolutionary computation for prior injection, heuristic generation, strategy selection, and adaptive optimization control (Wu et al., 2025; Liu et al., 2024a; Huang et al., 2026).

LEA extends this emerging paradigm to targeted adversarial optimization of vision-language models. Rather than directly generating perturbations, the LLM serves as a semantic-aware controller that coordinates region priors, operator selection, parameter adaptation, and objective balancing throughout the evolutionary process.

3 Problem Setup

We consider a targeted black-box attack against a vision-language model M. Given a clean image I, its description T, and a target description T_{ref}, the attacker aims to construct

$$I_{\mathrm{adv}} = I + \delta_I, \qquad \|\delta_I\|_\infty \le \epsilon_I, \tag{1}$$

such that the victim output approaches the target semantics. The attacker has no access to the architecture, parameters, gradients, or other intermediates of M. Nevertheless, the attacker can use a publicly available vision-language surrogate, while M is used only for black-box evaluation.

Let E_v and E_t denote the surrogate image and text encoders. The original and adversarial semantic representations are

$$F = \alpha E_v(I) + (1-\alpha)E_t(T), \qquad F_{\mathrm{adv}} = \alpha E_v(I_{\mathrm{adv}}) + (1-\alpha)E_t(T), \tag{2}$$

where α controls modality fusion. Given a target semantic reference F_{ref} constructed in Section 4.2, we consider the attack objective

$$\max_{\delta_I} \ \mathrm{sim}(F_{\mathrm{adv}}, F_{\mathrm{ref}}) - \gamma\,\mathrm{sim}(F_{\mathrm{adv}}, F), \quad \text{s.t. } \|\delta_I\|_\infty \le \epsilon_I. \tag{3}$$

Instead of evading M directly, Eq. (3) describes the semantic attack objective at the problem level. In the actual evolutionary optimization, LEA further decomposes this objective into target-semantic alignment, original-semantic separation, and perturbation magnitude, enabling explicit multi-objective population evaluation as described in Section 4.3.

4 Method

4.1 Overview

We propose LEA, an LLM-guided evolutionary framework for targeted black-box attacks against vision-language models. As illustrated in Fig. 1, LEA follows a cross-modal semantic-guided image-space attack paradigm. Instead of perturbing both modalities, LEA modifies only the visual input, while the textual information is used to define semantic references and guide the optimization process.

The key idea of LEA is to transform conventional semantically blind evolutionary search into an adaptive, target-aware optimization process. To this end, LEA first constructs two semantic anchors: an *original semantic anchor* that characterizes the clean image–text pair, and a *target semantic anchor* that specifies the desired adversarial semantic direction. The perturbation is then optimized by a

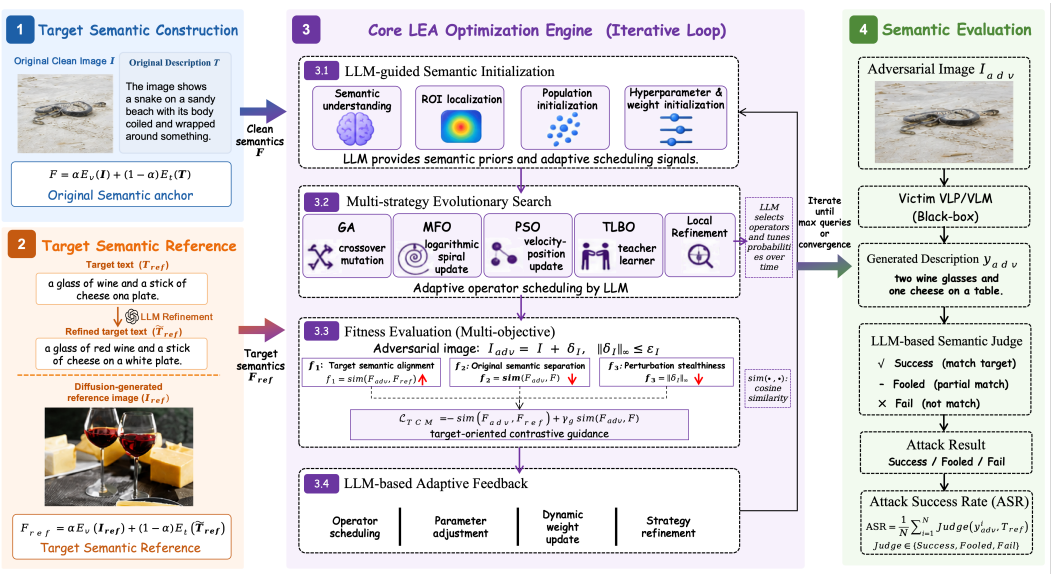


Figure 1: The LEA framework contains four phases: i&ii. constructing original and target semantic anchors, iii. performing LLM-guided multi-strategy evolutionary optimization with adaptive operator scheduling and multi-objective fitness evaluation; iv. using an LLM-based semantic judge to determine the attack result under the black-box VLP setting.

multi-strategy evolutionary search process, where an LLM acts as a high-level semantic controller. Rather than directly generating perturbations, the LLM analyzes the semantic and optimization states and dynamically provides region priors, operator-selection probabilities, parameter recommendations, and objective-balancing signals. The low-level perturbation search is performed by complementary evolutionary operators, enabling LEA to combine global exploration, local exploitation, and semantic-aware adaptation within a unified framework. The LLM-based semantic judge is used only for post-hoc attack evaluation and does not provide victim-model feedback to the surrogate-based evolutionary optimization.

4.2 Semantic Anchor Construction

Original semantic anchor. Let I denote the clean image and T its corresponding textual description. We employ a publicly available vision-language surrogate model with an image encoder $E_v(\cdot)$ and a text encoder $E_t(\cdot)$ to characterize their cross-modal semantics. Following the modality-aware representation used in prior transferable attacks (Xie et al., 2025), we define the original semantic anchor as

$$F = \alpha E_v(I) + (1-\alpha)E_t(T), \tag{4}$$

where $\alpha \in [0,1]$ controls the relative contribution of visual and textual information. The resulting representation F describes the original semantic state and is subsequently used as a reference from which the adversarial representation is encouraged to depart.

Target semantic anchor. Given a user-specified target description T_{ref}, we first employ an LLM-based semantic refinement function to retain the visually relevant objects, attributes, and relations,

$$\widetilde{T}_{\text{ref}} = \Phi_{\text{refine}}(T_{\text{ref}}; p_{\text{ref}}), \tag{5}$$

where p_{ref} denotes the refinement prompt. A text-to-image generator $G_{T\to I}$, instantiated by Stable Diffusion (Rombach et al., 2022), then synthesizes a reference image

$$I_{\text{ref}} = G_{T\to I}(\widetilde{T}_{\text{ref}}). \tag{6}$$

The target semantic anchor is constructed using the same surrogate encoders,

$$F_{\text{ref}} = \alpha E_v(I_{\text{ref}}) + (1-\alpha)E_t(\widetilde{T}_{\text{ref}}). \tag{7}$$

Combining the refined target text with its synthesized visual reference provides a joint cross-modal specification of the desired target semantics. The resulting target anchor explicitly defines the semantic region toward which the adversarial representation is optimized.

For a perturbation parameterized by $\delta_I = g_\theta(I)$, the adversarial representation is

$$F_{\text{adv}}(\theta) = \alpha E_v(I + g_\theta(I)) + (1-\alpha)E_t(T), \tag{8}$$

where the original text T remains unchanged. Therefore, LEA optimizes only the image modality while using cross-modal semantics as guidance.

4.3 LLM-Guided Multi-strategy Evolutionary Optimization

Multi-objective adversarial search. The perturbation parameters θ are optimized under the feasible set $\Theta(\epsilon_I)$ such that $\|g_\theta(I)\|_\infty \le \epsilon_I$. We consider three complementary objectives:

$$\begin{aligned} f_1(\theta) &= \text{sim}\left(F_{\text{adv}}(\theta), F_{\text{ref}}\right), \\ f_2(\theta) &= \text{sim}\left(F_{\text{adv}}(\theta), F\right), \\ f_3(\theta) &= \|g_\theta(I)\|_\infty . \end{aligned} \tag{9}$$

Here, f_1 measures target-semantic alignment and is maximized, whereas f_2 measures residual similarity to the original semantics and is minimized. The third objective f_3 measures the perturbation magnitude and is also minimized. Accordingly, the evolutionary search jointly promotes target alignment, original-semantic separation, and bounded image-space perturbations.

In addition to the vector-valued objectives used for population-level multi-objective evaluation, LEA adopts a generation-dependent targeted contrastive matching (TCM) score to explicitly balance target alignment and original-semantic separation:

$$\mathcal{L}_{\text{TCM}}^{(g)}(\theta) = -\,\text{sim}\left(F_{\text{adv}}(\theta), F_{\text{ref}}\right) + \gamma^{(g)}\,\text{sim}\left(F_{\text{adv}}(\theta), F\right), \tag{10}$$

where $\gamma^{(g)} > 0$ is adaptively adjusted by the LLM controller according to the current evolutionary state. The three-objective representation is maintained for population evaluation and archive updating, while $\mathcal{L}_{\text{TCM}}^{(g)}$ provides an adaptive semantic-balancing signal during evolutionary optimization.

At generation g, the population and non-dominated elite archive are denoted by

$$\mathcal{P}^{(g)} = \left\{\theta_i^{(g)}\right\}_{i=1}^{N}, \qquad \mathcal{A}^{(g+1)} = \text{UpdateArchive}\left(\mathcal{P}^{(g)}, \mathcal{A}^{(g)}\right). \tag{11}$$

where $\mathcal{P}^{(g)}$ denotes the population at generation g, and $\mathcal{A}^{(g)}$ denotes the non-dominated elite archive accumulated from previous generations. The archive preserves high-quality non-dominated candidates and provides a compact summary of the evolutionary state.

LLM as a semantic-aware adaptive controller. The central design of LEA is to use an LLM as a high-level controller rather than a direct perturbation generator. At each generation, we summarize the current optimization state using the population, elite archive, objective statistics, population diversity, and the source/target semantics. The controller is formulated as

$$\left(p_{\text{ROI}}^{(g)}, \boldsymbol{\pi}^{(g)}, \boldsymbol{\lambda}^{(g)}, \gamma^{(g)}\right) = \Phi_{\text{ctrl}}\left(\mathcal{S}^{(g)}, T, \widetilde{T}_{\text{ref}}\right), \tag{12}$$

where $\mathcal{S}^{(g)}$ denotes the current evolutionary state. $p_{\text{ROI}}^{(g)}$ is a semantic region prior that biases perturbation sampling toward potentially relevant image regions; $\boldsymbol{\pi}^{(g)}$ specifies the selection probabilities of different evolutionary operators; $\boldsymbol{\lambda}^{(g)}$ denotes their adaptive parameters; and $\gamma^{(g)}$ controls the trade-off between approaching the target semantics and departing from the original semantics.

This formulation makes the LLM responsible for *what to search and how to search*, while the evolutionary operators remain responsible for *how candidate perturbations are generated*. In this way, semantic reasoning and gradient-free optimization are explicitly decoupled.

Algorithm 1: Framework of LEA

Input: Clean image I, original text T, target text T_{ref}, surrogate encoders (E_v, E_t), budget ϵ_I, maximum generations $G_{\max}$, operator pool $\mathcal{O}$, and an LLM-based controller Φ_{ctrl}.
Output: Adversarial image I_{adv}

Initialize: Refine T_{ref}, build the reference embedding F and target embedding F_{ref}; initialize the population $\mathcal{P}^{(0)}$ and the elite archive $\mathcal{A}^{(0)} \leftarrow \varnothing$;
for $g = 0$ **to** $G_{\max} - 1$ **do**
  Evaluate: Score candidates in $\mathcal{P}^{(g)}$ under the multi-objective criteria and update $\mathcal{A}^{(g+1)}$;
  Perceive: Construct the state $\mathcal{S}^{(g)}$ from the population, archive, and objective statistics;
  Decide: Φ_{ctrl} maps $\mathcal{S}^{(g)}$ to an operator-selection policy $\boldsymbol{\pi}^{(g)}$ and its hyper-parameters;
  Mutate: Sample operators from $\boldsymbol{\pi}^{(g)}$ and generate offspring under the ϵ_I-ball projection;
  Select: Form $\mathcal{P}^{(g+1)}$ by multi-objective fitness and TCM-based semantic progress in Eq.(10);
  if *converged* **then break**;
Output: Pick the elite θ^* from $\mathcal{A}_{\text{final}}$ that maximizes the attack objective;
return $I_{\text{adv}} \leftarrow I + g_{\theta^*}(I)$;

Multi-strategy operator pool. To provide complementary exploration and exploitation behaviors, LEA employs

$$\mathcal{O} = \{\text{GA}, \text{MFO}, \text{PSO}, \text{TLBO}, \text{LocalRefine}\}. \tag{13}$$

The operators play different roles during optimization. GA maintains population diversity through crossover and mutation; MFO performs spiral-based updates around elite solutions; PSO exploits individual and population-level best states through velocity–position updates; TLBO refines solutions through teacher and learner interactions; and the local refinement operator performs small-scale adjustments in semantically sensitive regions to improve stealthiness. Additional descriptions of the evolutionary operators and their functional roles are provided in the supplementary file

For each parent $\theta_i^{(g)}$, an operator is sampled according to the LLM-predicted distribution,

$$o_i^{(g)} \sim \text{Categorical}(\boldsymbol{\pi}^{(g)}), \tag{14}$$

and the offspring is generated by

$$\theta_i' = \Pi_{\Theta(\epsilon_I)} \left[\text{Apply} \left(o_i^{(g)}, \theta_i^{(g)}; \boldsymbol{\lambda}^{(g)}, p_{\text{ROI}}^{(g)} \right) \right], \tag{15}$$

where $\Pi_{\Theta(\epsilon_I)}(\cdot)$ projects the candidate back to the feasible perturbation space.

After evaluating the offspring using the multi-objective criteria in Eq. (9) and computing the TCM score in Eq. (10), the elite archive and population are updated, and the resulting optimization statistics are fed back to Φ_{ctrl}. This produces a closed-loop process

$$\mathcal{S}^{(g)} \to \Phi_{\text{ctrl}} \to \left(p_{\text{ROI}}^{(g)}, \boldsymbol{\pi}^{(g)}, \boldsymbol{\lambda}^{(g)}, \gamma^{(g)} \right) \to \mathcal{P}^{(g+1)} \to \mathcal{S}^{(g+1)}. \tag{16}$$

through which LEA adaptively changes its search behavior as optimization proceeds. The process terminates when the maximum generation budget is reached or the evolutionary population converges. The final adversarial example is constructed from the selected elite solution θ^* as

$$I_{\text{adv}} = I + g_{\theta^*}(I). \tag{17}$$

Algorithm 1 summarizes the whole iterative loop from perceiving, decision, to selection. The complete optimization procedure and implementation details are provided in the supplementary file.

5 Experiments

We evaluate LEA from four complementary perspectives. First, we examine whether LEA can improve targeted attack effectiveness across different black-box vision-language models. Second, we

analyze the contribution of LLM-based semantic guidance and multi-strategy evolutionary search. Third, we investigate the computational trade-off between attack performance and optimization cost. Finally, we verify whether the LLM-based semantic evaluation protocol provides stable and reliable attack judgments.

5.1 Experimental Setup

Datasets and victim models. Clean images are uniformly sampled from the ImageNet-1K validation set (Deng et al., 2009), while target descriptions are selected from MS-COCO Captions (Chen et al., 2015) after removing overly short or non-descriptive captions. We evaluate transferability on three representative vision-language models: SMALLCAP (Ramos et al., 2023), ViECap (Fei et al., 2023), and LLaVA-13B (Liu et al., 2024b). For the prompt-sensitive LLaVA-13B model, we use the fixed query "What is the content of this image?" for all clean and adversarial inputs to ensure consistent evaluation.

Black-box setting and baselines. Following the threat model in Section 3, the attacker has no access to the architecture, parameters, gradients, or intermediate representations of the victim models. Adversarial examples are optimized using the public surrogate model and then transferred to unseen victims for black-box evaluation. We compare LEA with four representative baselines: BR (Dong et al., 2023), UB (Tu et al., 2024), VRE (Zhao et al., 2023), and CoA (Xie et al., 2025). CoA is the most directly related targeted transfer-based attack, as it progressively updates semantic targets during adversarial generation.

Evaluation metrics. We evaluate attack effectiveness using Avg. CLIP, Target ASR, and Fooled Rate. CLIP similarity is computed as the cosine similarity between the adversarial-image embedding and the target-text embedding. We report the average result over RN50, RN101, ViT-B/16, ViT-B/32, and ViT-L/14, with cosine similarities multiplied by 100 and reported in percentage form. CLIP ViT-B/32 is used as the optimization surrogate. RN50, RN101, ViT-B/16, and ViT-L/14 serve as additional semantic evaluators. For consistency with the original evaluation protocol, Avg. CLIP is reported as the average over all five backbones.

For open-ended victim outputs, the semantic evaluator assigns each output to one of three categories: *Target*, *Fooled-only*, or *Fail*. Target denotes that the generated output successfully expresses the specified target semantics; Fooled-only indicates that the output substantially deviates from the original semantics without sufficiently matching the target; and Fail indicates that it remains consistent with the original semantics or does not match the target. Accordingly,

$$\text{Target ASR}(\%) = 100 \times \frac{N_{\text{target}}}{N}, \tag{18}$$

and

$$\text{Fooled Rate}(\%) = 100 \times \frac{N_{\text{target}} + N_{\text{fooled}}}{N}. \tag{19}$$

CLIP-based semantic similarity (Radford et al., 2021) and LLM-based attack success are evaluated independently: CLIP encoders are used only for embedding similarity, whereas Target ASR and Fooled Rate are determined by the semantic evaluator. Complete backbone-wise results are reported in the supplementary file

5.2 Main Results

As shown in Table 1, LEA consistently achieves the strongest target-oriented attack performance across all three victim models. On SMALLCAP, LEA improves Avg. CLIP from 63.78 for CoA to 68.76 and Target ASR from 39.7% to 53.2%, while increasing the Fooled Rate from 49.3% to 65.3%. On ViECap, LEA achieves 72.3% Target ASR, outperforming CoA by 7.3 percentage points, together with the highest Avg. CLIP of 70.12.

LLaVA-13B is considerably more resistant to precise target-semantic steering; nevertheless, LEA increases Target ASR from 8.2% to 10.7% and Fooled Rate from 10.8% to 20.4%. These consistent improvements across architecturally different victim models demonstrate the cross-model transferability of the generated adversarial examples. Full results for individual CLIP encoders are provided in the supplementary file.

Table 1: Targeted transfer-based black-box attack results. Avg. CLIP denotes the average score over five CLIP backbones. Higher values are better for all metrics.

Victim	Method	Avg. CLIP (%)	Target ASR (%)	Fooled Rate (%)
SMALLCAP	BR	49.86	8.3	10.6
	UB	49.32	7.3	8.2
	VRE	57.02	19.8	24.9
	CoA	63.78	39.7	49.3
	LEA	**68.76**	**53.2**	**65.3**
ViECap	BR	45.22	14.3	22.5
	UB	45.78	10.8	18.7
	VRE	64.24	60.1	75.6
	CoA	68.74	65.0	80.4
	LEA	**70.12**	**72.3**	**82.2**
LLaVA-13B	BR	45.44	1.9	3.2
	UB	45.00	0.1	0.8
	VRE	45.22	3.9	6.1
	CoA	46.44	8.2	10.8
	LEA	**47.44**	**10.7**	**20.4**

5.3 Component Analysis

To examine the contribution of the major components, we construct three variants on SMALLCAP. **GA** retains only the conventional genetic evolutionary pipeline; **GA+LLM** introduces LLM-based semantic guidance while retaining the GA operator; and **LEA** further incorporates the complete multi-strategy operator pool.

Table 2: Component analysis on SMALLCAP.

Variant	Avg. CLIP (%)	Target ASR (%)	Fooled Rate (%)
GA	49.00	7.5	9.8
GA+LLM	63.00	36.6	49.5
LEA	**68.76**	**53.2**	**65.3**

Table 2 shows a clear progressive improvement. Conventional GA achieves only 7.5% Target ASR and an Avg. CLIP of 49.00, indicating that unguided evolutionary search is insufficient for efficient semantic steering. Introducing LLM guidance increases Target ASR to 36.6% and Avg. CLIP to 63.00, demonstrating the benefit of semantic priors and adaptive optimization guidance. The complete LEA framework further raises Target ASR to 53.2% and Avg. CLIP to 68.76, indicating that LLM-based semantic control and heterogeneous evolutionary search provide complementary benefits. The complete backbone-wise ablation results are reported in the supplementary file.

5.4 Efficiency and Evaluation Reliability

Table 3: Reliability of semantic evaluation on SMALLCAP.

Evaluator	Target ASR (%)	Fooled Rate (%)	Agreement (%)
GPT-4o	53.2	65.3	–
Claude-3.7	52.8	64.9	96.8%
Gemini-2.5	53.6	65.8	97.4%
Human	53.0	65.1	96.5%

Reliability of semantic evaluation. Since open-ended VLP outputs cannot be reliably evaluated by exact string matching, we examine whether the reported ASRs depend on a specific semantic evaluator. Although GPT-4o is used as both the default optimization controller and the default

semantic evaluator, the two roles are strictly separated: the evaluator is invoked only after adversarial generation and does not access the evolutionary state, elite archive, controller outputs, or fitness values.

The resulting attack rates differ by less than one percentage point across the three LLM judges, while human majority voting closely matches the automatic evaluation with 96.5% agreement. These results indicate that the reported attack effectiveness is largely insensitive to the choice of semantic evaluator. Detailed human-annotation settings and prompt-robustness results are provided in the supplementary file.

Runtime–performance trade-off. We further analyze the computational trade-off between attack performance and optimization cost. Figure 2 compares the average CLIP score with the average runtime per optimization step.

BR and UB require less than one second per step but achieve relatively low Avg. CLIP scores of approximately 49. VRE requires substantially higher computational cost while reaching an Avg. CLIP of 57.02. The conventional GA requires approximately 126 s/step but achieves only 49.00, indicating inefficient evolutionary search. Introducing LLM guidance improves Avg. CLIP to 63.00 while reducing the runtime to approximately 96 s/step, and LEA further reaches 68.76 at approximately 68 s/step.

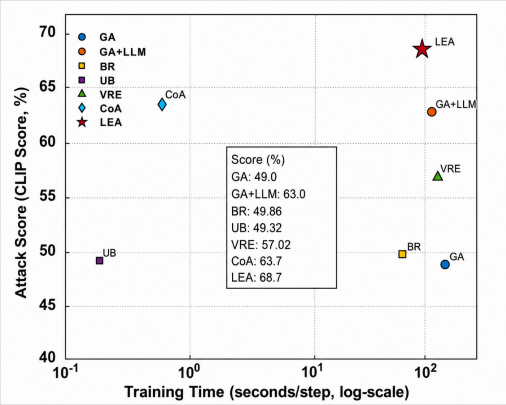


Figure 2: Attack performance versus computational cost. The horizontal axis denotes average runtime per optimization step, and the vertical axis denotes Avg. CLIP.

CoA remains substantially faster than LEA, requiring less than one second per step while achieving an Avg. CLIP of 63.78. Therefore, the result should be interpreted as a performance–computation trade-off rather than an absolute efficiency advantage: LEA introduces additional computational overhead, but provides the strongest target semantic alignment among the compared methods.

6 Conclusion

In this work, we introduced LEA, an LLM-guided evolutionary framework for targeted transfer-based black-box attacks against vision-language models. Experiments across multiple victim models show that high-level semantic guidance can substantially improve evolutionary adversarial search, while component analyses confirm the complementary benefits of LLM control and multi-strategy optimization.

More broadly, our results indicate that LLMs can serve as effective high-level controllers for gradient-free adversarial optimization, and that current vision-language systems remain vulnerable to targeted semantic manipulation under restricted black-box access. The main limitation of LEA is the additional computational overhead introduced by iterative evolutionary search and LLM interaction. Future work will focus on improving search efficiency and scalability, extending the framework to broader vision-language tasks and models, and developing defenses against semantic-guided transferable attacks.

AI Use Statement

Generative AI tools were used to assist with Chinese-to-English translation, language editing, manuscript organization, and feedback on the presentation of the methodology and experimental analysis. All mathematical formulations, experimental results, citations, and technical claims were independently verified by the authors. The authors take full responsibility for the final content of the paper.

REFERENCES

Yue Cao, Yun Xing, Jie Zhang, Di Lin, Tianwei Zhang, Ivor Tsang, Yang Liu, and Qing Guo. Scenetap: Scene-coherent typographic adversarial planner against vision-language models in real-world environments. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 25050–25059, 2025.

Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. VLP: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1):38–56, 2023.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

Yinpeng Dong, Huanran Chen, Jiawei Chen, Zhengwei Fang, Xiao Yang, Yichi Zhang, Yu Tian, Hang Su, and Jun Zhu. How robust is Google's Bard to adversarial image attacks? *arXiv preprint arXiv:2309.11751*, 2023.

Junjie Fei, Teng Wang, Jinrui Zhang, Zhenyu He, Chengjie Wang, and Feng Zheng. Transferable decoding with visual entities for zero-shot image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3136–3146, 2023.

Yuxiao Huang, Shenghao Wu, Wenjie Zhang, Jibin Wu, Liang Feng, and Kay Chen Tan. Autonomous multi-objective optimization using large language model. *IEEE Transactions on Evolutionary Computation*, 30(2):594–608, 2026.

Fengling Li, Tianshi Wang, Lei Zhu, Jingjing Li, and Heng Tao Shen. Attack as defense: Proactive adversarial multi-modal learning to evade retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4717–4733, 2025.

Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Yu Cheng, and Wei Hu. A survey of attacks on large vision-language models: Resources, advances, and future trends. *IEEE Transactions on Neural Networks and Learning Systems*, 36(11):19525–19545, 2025. doi: 10.1109/TNNLS.2025.3592935.

Daizong Liu, Xiaowen Cai, Pan Zhou, Xiaoye Qu, Lichao Sun, and Wei Hu. Are large vision-language models robust to adversarial visual transformations? *IEEE Transactions on Information Forensics and Security*, 21:2564–2579, 2026.

Fei Liu, Xialiang Tong, Mingxuan Yuan, Xi Lin, Fu Luo, Zhenkun Wang, Zhichao Lu, and Qingfu Zhang. Evolution of heuristics: Towards efficient automatic algorithm design using large language model. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32201–32223. PMLR, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024b.

Dong Lu, Zhiqiang Wang, Teng Wang, Weili Guan, Hongchang Gao, and Feng Zheng. Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 102–111, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. Smallcap: lightweight image captioning prompted with retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2840–2849, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Haoqin Tu, Chenhang Cui, Zijun Wang, Yiyang Zhou, Bingchen Zhao, Junlin Han, Wangchunshu Zhou, Huaxiu Yao, and Cihang Xie. How many are in this image? a safety evaluation benchmark for vision LLMs. In *European Conference on Computer Vision*, pp. 37–55. Springer, 2024.

Xingyu Wu, Sheng-hao Wu, Jibin Wu, Liang Feng, and Kay Chen Tan. Evolutionary computation in the era of large language model: Survey and roadmap. *IEEE Transactions on Evolutionary Computation*, 29(2):534–554, 2025.

Peng Xie, Yequan Bie, Jianda Mao, Yangqiu Song, Yang Wang, Hao Chen, and Kani Chen. Chain of attack: On the robustness of vision-language models against transfer-based adversarial attacks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14679–14689, 2025.

Ziyi Yin, Muchao Ye, Tianrong Zhang, Tianyu Du, Jinguo Zhu, Han Liu, Jinghui Chen, Ting Wang, and Fenglong Ma. Vlattack: Multimodal adversarial attacks on vision-language tasks via pre-trained models. *Advances in Neural Information Processing Systems*, 36:52936–52956, 2023.

Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 5005–5013, 2022.

Jiaming Zhang, Junhong Ye, Xingjun Ma, Yige Li, Yunfan Yang, Yunhao Chen, Jitao Sang, and Dit-Yan Yeung. Anyattack: Towards large-scale self-supervised adversarial attacks on vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19900–19909, June 2025.

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.