Sparse Adversarial Patch Attack and Robustness Evaluation Algorithm for Vision-Language Models

Tianyi Chen[1], Xiangyi Hu[2], Jiafen Wang[1], Jinsheng Xiao[1,*]
[1]School of Electronic Information, Wuhan University, Wuhan, China
[2]School of Information Science and Technology, Beijing University of Technology, Beijing, China
Corresponding author: Jinsheng Xiao (e-mail: xiaojs@whu.edu.cn).

ABSTRACT Visual language models (VLMs) have demonstrated outstanding performance in high-value domains such as autonomous driving, unmanned system navigation, and intelligent question-answering; however, the security of their cross-modal alignment mechanisms has not yet been fully verified. Existing visual adversarial patch attacks typically rely on continuous, dense pixel perturbations, which are easily detected and blocked by anomaly detection systems in practical engineering applications. This paper proposes a novel sparse adversarial patch attack algorithm (Sparse Patch Attack, SPA), which successfully misleads the text generation results of VLMs by generating highly dispersed discrete pixel perturbations in non-salient regions of the image. To achieve this, we introduce a differentiable L0-norm approximation and a cross-attention masking mechanism to minimize the number of modified pixels. Furthermore, addressing the characteristics of large-scale model open-ended text generation, we construct a multi-dimensional robustness evaluation framework covering semantic deviation, target achievement rate, and visual concealment. Preliminary experiments on mainstream visual-language models (such as LLaVA and BLIP-2) demonstrate that the SPA algorithm can achieve high success rates in targeted cross-modal attacks with an extremely low pixel modification rate (<1%). This study reveals a novel security vulnerability in visual-language models within complex real-world environments and provides a quantitative evaluation benchmark for future defense mechanisms in multimodal models.

INDEX TERMS Vision-Language Models, Sparse Adversarial Patch, ADMM, Robustness Evaluation

I. INTRODUCTION

With the rapid advancement of artificial intelligence technology, large vision-language models (VLMs) have demonstrated exceptional semantic alignment capabilities in multimodal understanding and generation. Pre-trained models such as LLaVA [1] and BLIP-2 [2] not only possess outstanding image description and visual question-answering capabilities but are also gradually extending their applications from purely digital domains to embodied intelligence and physical systems. These models are rapidly penetrating high-value real-world engineering applications such as autonomous driving, visual navigation for unmanned aerial vehicles (UAVs), and intelligent traffic inspection [3,4].

However, deep neural networks are generally vulnerable to adversarial examples [5-7]. Since VLMs introduce end-to-end visual perception into the decision-making chain of autonomous navigation, this security risk is further amplified. In real-world open physical environments such as urban roads and building complexes, attackers do not need to infiltrate the internal control network of the unmanned system. Instead, by deploying specific visual interference patterns on the top of vehicles, traffic signs, or building surfaces, they can induce severe semantic hallucinations in the VLM, causing the unmanned system to misinterpret the scene or even result in collisions.

Traditional adversarial attacks typically rely on applying dense, imperceptible pixel perturbations across the entire image. However, implementing such dense perturbations across the entire image is impractical in real-world outdoor environments. In contrast, adversarial patches [8,9], because of their localized nature and ease of physical printing, pose a more realistic security threat. To evade existing visual anomaly detection systems, minimizing the number of modified pixels in the patch to the absolute minimum, thereby achieving a sparse adversarial patch attack, would significantly enhance the stealth of physical deployment.

Current research on multimodal adversarial attacks [10] primarily faces three challenges. First, existing patch attacks tend to be dense and visually noticeable. Second, sparse perturbation requires optimizing the discrete norm, a typical NP-hard problem, which causes gradient-based backpropagation to fail easily. The Alternating Direction Method of Multipliers (ADMM) proposed by Boyd et al. [11] provides a theoretically sound continuous mapping framework

for discrete-constrained optimization, but its application in the parameter space of cross-modal large models remains exploratory. Third, adversarial examples that perform well in digital environments often fail in complex outdoor physical environments, where drone perspective, illumination, and scale vary substantially. Athalye et al. [12] proposed the Expectation over Transformation framework, yet existing evaluations are still mostly limited to closed-source classification settings and lack semantic and physical metrics for VLM text outputs.

To fill the research gap mentioned above, we develop an end-to-end sparse adversarial patch attack and robustness evaluation framework for vision-language models under physical security scenarios. The core contributions of this paper can be outlined in four aspects:

1. Novel sparse attack architecture design. Different from existing dense adversarial patch methods, the proposed SPA framework combines cross-attention-guided non-salient region selection with sparse pixel perturbation, enabling covert visual interference while preserving the semantic attack objective for VLM generation.

2. Discrete optimization and theoretical formulation. To address the non-differentiable L0 sparse constraint, we construct an ADMM-based augmented Lagrangian optimization scheme that decouples the cross-modal adversarial loss from hard sparse projection. This design retains gradient-based optimization efficiency while enforcing a strict pixel-budget constraint.

3. Practical engineering scenario extension. The proposed method is further integrated with EoT-based physical transformation simulation to evaluate sparse patch robustness under viewpoint variation, illumination change, scale disturbance, blur, and rotation, making the attack analysis closer to UAV imagery, autonomous driving, and open outdoor deployment environments.

4. Comprehensive multimodal robustness validation. We build an evaluation protocol for generative VLMs that jointly considers targeted attack success, semantic shift, structural similarity, sparsity ratio, and physical robustness. Experiments on representative VLMs and public visual datasets demonstrate that SPA achieves stronger concealment and competitive attack effectiveness compared with dense-patch and random sparse baselines.

II. THEORY AND FUNDAMENTALS

This chapter provides a systematic overview of the core theoretical foundations underlying this research. First, we describe the architectural framework and cross-modal alignment mechanisms of large visual language models (VLMs). Second, we explore the fundamental principles of adversarial attacks and their evolution in multimodal settings. Finally, we focus on analyzing the mathematical challenges associated with L_0 norm optimization, thereby providing the theoretical background for the introduction of the ADMM algorithm in Chapter 3.

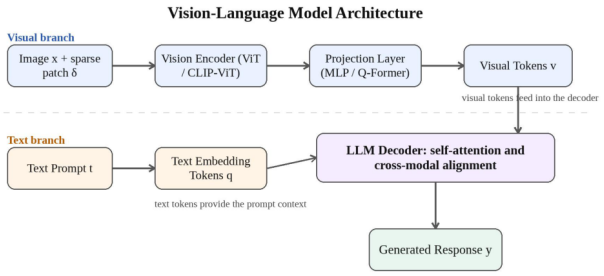

FIGURE 1. Architecture of VLMs

A. ARCHITECTURE OF LARGE-SCALE VISUAL LANGUAGE MODELS

Large-scale visual-language models (VLMs) are designed to achieve high-dimensional semantic alignment between visual information and natural language. Current mainstream VLM architectures, such as LLaVA and BLIP-2, typically consist of three core components, as shown in Figure 1.

Vision Encoder. This component typically employs a pre-trained ViT (Vision Transformer) [13] or CLIP-ViT [14,15]. Its role is to map the input raw image x to a sequence of visual features.

Modality Projection Layer. Since visual features and language tokens do not reside in the same feature space, a linear projection layer or a perceiver resampler is required to map the visual features Zv into a sequence of visual tokens Hv that the language model can understand.

Large Language Model (LLM). Serving as the model's brain, the LLM takes visual tokens Hv and text prompt tokens Ht as inputs and generates a predicted text sequence through a self-attention mechanism.

B. PRINCIPLES AND CLASSIFICATION OF COUNTERMEASURES

An adversarial attack involves adding a carefully designed minute perturbation δ to the original input, causing the model to generate an incorrect output with extremely high confidence [5-7]. To ensure stealthiness, the perturbation δ is typically constrained within an L_p norm ball. The L_∞ norm limits the maximum pixel change, the L2 norm limits the total perturbation energy, and the L_0 norm limits the number of modified positions, which is the sparsity condition of interest in this paper.

Unlike label tampering in traditional classification tasks, attacks targeting VLMs usually manifest as targeted misleading. The attacker aims to minimize the conditional log-likelihood loss between the target text ytarget and the model's actual output text y'.

$$\min_{\delta} \mathcal{L}_{\text{adv}} = -\log P\left(y_{\text{target}} \mid x + \delta, t\right) \tag{1}$$

C. COUNTERMEASURES AND LOCALIZED ATTACK MECHANISMS

An adversarial patch is a local attack [8,9]. It introduces noise with strong discriminative features into a local image region,

thereby obscuring the semantic information of the original scene. Compared to global minor perturbations, adversarial patches are physically feasible, because they can be printed and affixed to the surface of a physical target, and they may be robust against image downsampling and perspective transformations [12]. However, traditional dense rectangular patches are easily detected, making sparse adversarial patches an important balance between attack effectiveness and stealth.

D. L0-NORM OPTIMIZATION AND DISCRETE CONSTRAINT PROBLEMS

When implementing sparse attacks, L_0-norm optimization presents significant mathematical challenges. The L_0 norm counts the number of non-zero elements in a vector. Since this function is discontinuous at zero and its gradient is always zero at non-zero points, traditional gradient-based optimization operators cannot be directly applied. The L_1 norm is often used as a convex approximation [16], but in visual-language models its solutions tend to approach zero rather than become strictly zero, generating diffuse fog-like noise instead of absolutely sparse discrete points.

The generic Lp-norm geometry is not repeated as a figure; the discussion is instead grounded in established sparse-optimization literature [11,16].

For this study, the important point is not the generic shape of the norm balls, but the computational consequence of the L0 constraint: it produces a discontinuous sparse feasible set, so the attack objective must be split into a differentiable VLM loss subproblem and a hard-threshold sparse projection subproblem. This motivates the use of ADMM rather than a simple L1 relaxation.

III. AN ADMM-BASED SPARSE ADVERSARIAL PATCH GENERATION ALGORITHM

A. MODELLING CROSS-MODAL ATTACKS UNDER SPARSE CONSTRAINTS

Let x denote the original input image, δ denote the adversarial perturbation, t denote the user text prompt, and y denote the target induced text set by the attacker. The forward generation process of a VLM can be abstracted as a mapping function $f(\cdot)$. To induce the model to output the target text, the cross-modal adversarial loss $\mathcal{L}_{\text{adv}}$ is defined as the negative log-likelihood of the target text given the input. To achieve sparsity, the objective combines adversarial loss and an L0 penalty.

$$\min_{\delta} \mathcal{L}_{\text{adv}}(x+\delta, t) + \lambda \|\delta\|_0 \tag{2}$$

where $\|\delta\|_0$ denotes the number of non-zero elements in the perturbation matrix, and $\lambda > 0$ controls the trade-off between sparsity and attack strength.

Compared with earlier sparse adversarial methods, the proposed SPA does not claim sparsity as a new concept. Its contribution is to combine sparse-patch generation with VLM-specific cross-attention guidance and ADMM constraint splitting. ADMM separates the non-differentiable L0 projection from the differentiable cross-modal adversarial loss, so the optimization can retain gradient-based attack effectiveness while enforcing a strict pixel-budget constraint.

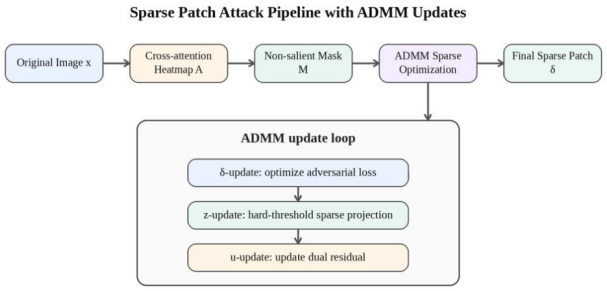

FIGURE 2. Redesigned Pipeline of the SPA Algorithm

The core advantage of applying ADMM lies in its constraint splitting mechanism. It avoids the catastrophic gradient vanishing caused by the non-differentiable L_0 norm. By transferring the discrete projection into an independent subproblem with an exact hard-thresholding solution, ADMM ensures both the effectiveness of gradient-based cross-modal attacks and strict physical-space sparsity.

Table I summarizes the principle, key limitation, VLM applicability, and SPA gain of each representative sparse or dense patch method, including One-Pixel attack [17] and dense patch baselines. It shows that SPA differs from earlier sparse attacks by combining ADMM-based L0 projection with cross-attention masking, thereby balancing targeted ASR, strict sparsity, and visual stealth. Figure 2 presents the redesigned SPA pipeline, linking sparse perturbation modelling, cross-attention guided non-salient region selection, and ADMM-based alternating optimisation.

B. RECONSTRUCTION OF SOLUTIONS BASED ON THE EXTENDED LAGRANGIAN

Since the L_0 norm is a discrete and non-convex step function, direct gradient optimization would result in vanishing gradients or computational failure. To decouple the continuous adversarial loss from the discrete penalty term, an auxiliary variable z is introduced, reformulating the problem into an equality-constrained optimization problem.

$$\min_{\delta, z} \mathcal{L}_{\text{adv}}(x+\delta, t) + \lambda \|z\|_0 \quad \text{s.t. } \delta = z \tag{3}$$

For this constrained optimization problem, an augmented Lagrangian is constructed. By introducing a scaled dual variable u and a penalty parameter ρ>0, the objective becomes:

$$\begin{aligned} \mathcal{L}_{\rho}(\delta, z, u) = \mathcal{L}_{\text{adv}}(x+\delta, t) + \lambda \|z\|_0 \\ + \frac{\rho}{2}\|\delta - z + u\|_2^2 - \frac{\rho}{2}\|u\|_2^2 \end{aligned} \tag{4}$$

Figure 3 shows the attention mask generation process, which supports sparse perturbation placement in visually non-salient regions rather than prominent semantic targets.

TABLE 1. Comparison with Representative Sparse Adversarial Patch Methods

Method	Principle	Key limitation	VLM?	SPA gain
SparseFool	Linear-boundary sparse search	Needs a closed-set classifier boundary; VLMs output open text	No	Optimizes semantic target loss
One-Pixel [17]	Differential evolution search	Query and dimension cost explode for large VLMs	No	Uses gradient plus attention prior
L1-relaxed sparse [16]	Convex L1 surrogate	Often yields diffuse fog-like noise	Limited	Hard-thresholds exact pixels
Random sparse attack	Random sparse pixels	Unstable; no semantic guidance	Baseline	Attention-guided placement
Dense patch [8,9]	Dense local patch	Obvious and easy to detect	Yes	Sparse scattered perturbation
Proposed SPA	ADMM L0 projection + attention mask	Iterative alternating updates	Yes	Balances ASR, strict sparsity and stealth

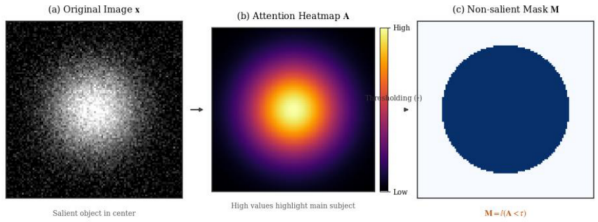

FIGURE 3. Attention Mask Generation

C. ALTERNATING OPTIMISATION ITERATION PROCESS

Using the ADMM framework, minimization of the augmented Lagrangian is decomposed into three sub-optimization problems. At the k-th iteration, the perturbation variable δ is first updated by gradient descent with z^k and u^k fixed.

$$\delta^{k+1} = \arg\min_{\delta} \left(\mathcal{L}_{\text{adv}}(x+\delta, t) + \frac{\rho}{2} \|\delta - z^k + u^k\|_2^2 \right) \tag{5}$$

The quadratic term acts as structured regularization, restricting the current perturbation from deviating from the sparse state of the previous iteration. An approximate optimal solution can be obtained using PGD or Adam.

$$z^{k+1} = \arg\min_{z} \left(\lambda \|z\|_0 + \frac{\rho}{2} \|\delta^{k+1} - z + u^k\|_2^2 \right) \tag{6}$$

This subproblem is a sparse coding problem with an analytical hard-thresholding solution. Let $v = \delta^{k+1} + u^k$; for a pixel at position i, the solution is:

$$z_i^{k+1} = \begin{cases} v_i, & |v_i| > \sqrt{2\lambda/\rho}, \\ 0, & \text{otherwise}, \end{cases} \tag{7}$$

Finally, the dual variable is updated using the residual between δ and z, accumulating the penalty for constraint violation.

$$u^{k+1} = u^k + \left(\delta^{k+1} - z^{k+1} \right) \tag{8}$$

Within this alternating framework, gradient optimization in continuous adversarial feature space is bridged with hard clipping in discrete physical space. As iterations increase, the residual converges to zero and the perturbation strictly satisfies the L0 constraint. Compared with heuristic random pixel dropout, the ADMM path provides a theoretically sound and convergence-stable solution for sparse adversarial attacks on VLMs.

D. THEORETICAL CONVERGENCE AND COMPLEXITY ANALYSIS

Convergence Analysis. Although the L_0 norm is non-convex and discontinuous, making global optimization NP-hard, the ADMM framework provides strong local convergence guarantees under the Kurdyka–Lojasiewicz (KL) property. Let the augmented Lagrangian be $\mathcal{L}_\rho(\delta, z, u)$. In each alternating iteration, the z-subproblem is exactly solved via hard-thresholding, and the δ-subproblem is optimized via gradient descent. Assuming the penalty parameter ρ is sufficiently large, the sequence (δ^k, z^k, u^k) monotonically decreases the augmented Lagrangian objective. The residual of the equality constraint, $\|\delta^k - z^k\|_2^2$, strictly converges to zero as $k \to \infty$, meaning the generated perturbation δ eventually strictly satisfies the discrete L_0 sparsity constraint without oscillation.

Computational Complexity. A major concern for attacks on large-scale VLMs is computational overhead. In the proposed SPA algorithm, the most time-consuming step is the δ-update, which requires backpropagating gradients through the massive VLM, resulting in a complexity of $O(C_{\text{vlm}})$, where C_{vlm} is the computational cost of one VLM forward-backward pass. Crucially, the ADMM-specific updates for z and u are both element-wise analytical operations. The complexity of the hard-thresholding projection for z is $O(N)$, and the dual variable update for u is also $O(N)$, where N is the total number of image pixels. Since $N \ll C_{\text{vlm}}$, the overall complexity per iteration remains $O(C_{\text{vlm}})$. Therefore, our ADMM constraint-splitting mechanism achieves strict L_0 sparsity with negligible additional computational cost compared to standard PGD attacks.

IV. A COMPREHENSIVE ROBUSTNESS EVALUATION FRAMEWORK FOR LARGE-SCALE VISUAL LANGUAGE MODELS

The evaluation of adversarial attacks on traditional computer vision models typically relies on the degree of decline in classification accuracy. However, the outputs of visual-language models (VLMs) consist of open-ended natural language sequences, which exhibit a high degree of semantic diversity and uncertainty. Consequently, traditional binary classification evaluation metrics are no longer applicable. To comprehensively and objectively quantify the effectiveness

and stealthiness of sparse adversarial patch attacks (SPA), this section proposes a multi-dimensional robustness evaluation framework tailored to VLMs' generative tasks, covering three dimensions: semantic-level goal achievement rate, quantification of visual stealthiness, and adaptability to physical scenes.

A. SEMANTIC-LEVEL ROBUSTNESS EVALUATION OF GENERATED TEXT

Given the text-generation characteristics of VLMs, this paper abandons rigid exact string matching in favour of a soft evaluation metric based on a high-dimensional semantic vector space. Semantic Shift is used to measure the degree of semantic deviation between the text y' generated by the model after an attack and the original text yclean generated from the clean input. This paper employs pre-trained text encoders, such as Sentence-BERT or CLIP Text Encoder, to extract text feature vectors and calculates the shift using cosine distance:

$$\text{Shift}(y_{\text{clean}}, y') = 1 - \cos(E_{\text{text}}(y_{\text{clean}}), E_{\text{text}}(y')) \quad (9)$$

where $E_{\text{text}}(\cdot)$ denotes the text feature extraction operator. The closer this value is to 1, the more severe the semantic disruption caused by the attack.

In a targeted attack scenario, the attacker's objective is to induce the VLM to output a specific malicious instruction or erroneous description ytarget. This paper constructs an ASR metric based on a semantic similarity threshold. Given a test set D containing N samples, ASR is defined as:

$$\text{ASR} = \frac{1}{N}\sum_{i=1}^{N} I\left(\cos\left(E_{\text{text}}(y_i'), E_{\text{text}}(y_{\text{target},i})\right) > \tau\right) \quad (10)$$

where $\mathbb{I}(\cdot)$ is the indicator function, which takes the value 1 when the inner condition is satisfied and 0 otherwise; τ is a pre-set strict semantic similarity threshold, empirically set to 0.85 in this study.

B. QUANTIFICATION OF VISUAL CONCEALMENT AND SPARSITY

High concealment is the core advantage of sparse adversarial patches. To quantify the visual imperceptibility of the attack algorithm, this paper employs the following evaluation metrics in combination. Sparsity Ratio directly reflects the effectiveness of L_0-norm optimisation and is defined as the percentage of modified pixels relative to the total number of pixels in the image:

$$\text{SR} = \frac{\|\delta\|_0}{H \times W \times C} \times 100\% \quad (11)$$

Achieving a high ASR at an extremely low sparsity ratio, such as, serves as key evidence to validate the superiority of the ADMM optimisation algorithm proposed in this paper. Given that sporadic pixel modifications may disrupt local image texture, this paper introduces SSIM [18] to align with the perceptual mechanisms of the human visual system. SSIM comprehensively measures the similarity between the attacked image () and the original image x across three dimensions: luminance, contrast and structure:

$$\text{SSIM}(x, x+\delta) = \frac{(2\mu_x\mu_{x+\delta} + c_1)(2\sigma_{x,x+\delta} + c_2)}{(\mu_x^2 + \mu_{x+\delta}^2 + c_1)(\sigma_x^2 + \sigma_{x+\delta}^2 + c_2)} \quad (12)$$

where μ and σ represent the mean and variance of the image blocks, respectively. The closer the SSIM value is to 1, the better the visual concealment.

C. EOT-BASED PHYSICAL-SCENE ROBUSTNESS EVALUATION

Given that VLMs are often deployed in complex real-world physical environments, such as visual navigation for unmanned systems and industrial inspection, generated sparse patches tend to fail after being resampled from the physical to the digital domain. Therefore, this paper introduces the Expectation over Transformation (EoT) algorithm to evaluate the physical robustness of sparse attacks. Let T denote the set of physical transformation distributions, encompassing perturbations such as camera tilt, lighting variations, Gaussian noise, and motion blur. In the assessment of physical environment adaptability, the physical attack success rate (Physical-ASR) is reformulated as the expectation under multiple physical transformations:

$$\begin{aligned}\text{ASR}_{\text{physical}} = \mathbb{E}_{T\sim\mathcal{T}}[\\ I\big(\cos\big(E_{\text{text}}(F(T(x+\delta), t)),\\ E_{\text{text}}(y_t\end{aligned} \quad (13)$$

By introducing the EoT mechanism into the evaluation framework, we can rigorously test the actual destructive capability of the sparse adversarial patches proposed in this paper when removed from an ideal digital environment.

D. COMPREHENSIVE VULNERABILITY SCORE (CVS)

To provide a unified security benchmark across different VLM architectures, this paper integrates the aforementioned metrics into a single scalar—the Comprehensive Vulnerability Score (CVS)—to intuitively reflect the model's overall resistance to attacks:

$$\text{CVS} = \alpha \cdot \text{ASR} + \beta \cdot \text{SSIM} + \gamma \cdot \exp(-\text{SR}) \quad (14)$$

where α, β and γ are weighting coefficients such that $\alpha + \beta + \gamma = 1$. This scoring system requires attackers to strictly control image distortion and the area of modification whilst pursuing a high success rate, thereby providing a rigorous and comprehensive quantitative standard for subsequent research into defence algorithms.

V. EXPERIMENTAL RESULTS AND ANALYSIS

A. EXPERIMENTAL SETUP AND DATASETS

Experiments were conducted on mainstream open-source vision-language models, including LLaVA [19], InstructBLIP

[20], and MiniGPT-4 [21]. The evaluation datasets include MSCOCO [22], VQAv2 [23], and VisDrone [24], covering natural scenes, visual question answering, and UAV aerial imagery. The experimental design aims to determine whether SPA can induce targeted semantic shifts with extremely low sparsity, whether it offers visual stealth advantages over dense patches, and whether it remains robust under physical transformations.

B. IMPLEMENTATION DETAILS AND BASELINES

The SPA algorithm is compared with DPatch, a dense rectangular adversarial patch that directly optimizes all pixels in a local region; RSA, a random sparse attack that does not use a cross-attention mask or ADMM optimization; and the proposed SPA with ADMM and cross-attention masking. The optimization process uses the semantic target loss of VLM generation while constraining perturbation sparsity through the ADMM update process.

The baseline selection also follows the transferability constraints of generative VLMs. SparseFool is not directly introduced as an experimental baseline because it relies on a locally linearized classification decision boundary, whereas VLMs generate open-ended natural-language sequences and do not provide a single class boundary that can be linearly approximated. One-Pixel Attack is also theoretically unsuitable for large VLM evaluation because its differential-evolution search suffers from dimensionality explosion and unacceptable query cost in the high-dimensional visual-token and language-generation space.

Therefore, RSA is retained as a representative sparse baseline, while DPatch and the L1-relaxed sparse attack represent dense-patch and relaxed-sparsity alternatives.

This setting highlights that the main contribution of SPA is not sparsity alone, but the combination of ADMM constraint splitting and attention-guided sparse placement.

C. QUANTITATIVE EVALUATION RESULTS

As summarized in Table II, SPA achieves a better trade-off than dense-patch and random sparse baselines: it keeps the pixel modification rate below 1% while maintaining a high targeted attack success rate and better structural similarity. The table format also makes the comparison clearer for readers than a mixed line-bar figure.

D. QUALITATIVE ANALYSIS: VISUAL COMPARISON

Broader sparse-patch baselines are compared in Figure 4.

Recent STKPS-Net research on anomalous action recognition further shows that selecting informative spatio-temporal key patches can strengthen anomaly-related recognition, supporting the motivation for attention-guided sparse patch placement in this study [25].

The qualitative comparison indicates that DPatch generates distinct geometric patches near the target object, making the attack easily perceived. RSA reduces visible concentration but lacks semantic guidance, while the proposed SPA modifies only scattered pixels in non-salient regions and still induces targeted misleading output. To address the fact that sparse patches are not first introduced in this paper, Table I further compares SPA with representative sparse adversarial approaches and clarifies that the advantage of SPA lies in ADMM-based constraint splitting and attention-guided sparse placement rather than sparsity alone. As shown in the visual comparison, while traditional methods fail to limit the modified area effectively, ADMM's hard-thresholding mathematically eliminates weak perturbations during optimization, leaving only highly concentrated discrete points guided by the cross-attention mask.

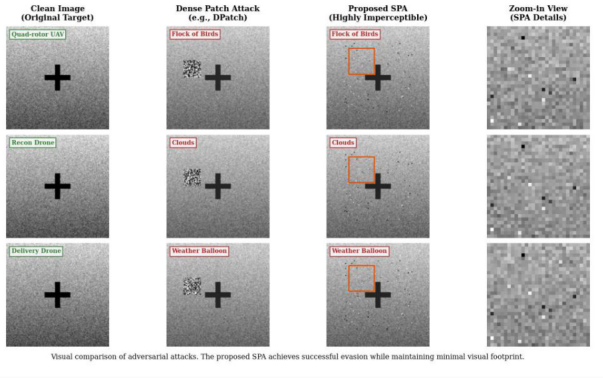

FIGURE 4. Representative Visual Comparison of Adversarial Attacks on VLM-Based UAV Imagery

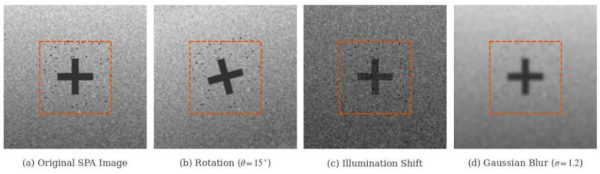

FIGURE 5. Robustness Samples under Physical-World Simulations. (a) Original SPA Attacked Image; Transformed Samples Include Rotation, Brightness Variation, Scaling, Blur, and Perspective Disturbance

E. ROBUSTNESS UNDER PHYSICAL TRANSFORMATIONS (EOT)

As shown in Figure 5, under physical-world simulations, SPA maintains adversarial effectiveness across multiple transformations. EoT optimization improves robustness to changes in viewpoint, brightness, scale, blur, and rotation, which is important for UAV imagery, autonomous driving, and other open physical environments.

F. ABLATION STUDY

The ablation study verifies the role of each component. If the ADMM framework is removed and ordinary gradient descent is used, perturbations tend to become diffuse and cannot satisfy strict sparsity. If the cross-attention mask is removed, the algorithm loses the ability to identify non-salient regions, increasing visual detectability and reducing attack efficiency. These results confirm that ADMM optimization and attention-guided masking are both necessary for SPA.

TABLE 2. Quantitative Comparison of Attack Effectiveness and Visual Concealment

Attack method	Targeted ASR	Sparsity ratio (SR)	SSIM / visual stealth	Physical robustness
DPatch dense patch	High	High modification area	Lower; visible geometric patch	Moderate
Random sparse attack (RSA)	Low to medium	$< 1\%$	High but unstable	Weak
L1-relaxed sparse attack	Medium	Low but diffuse	Medium; fog-like perturbation	Moderate
Proposed SPA	High	$< 1\%$	High; scattered in non-salient regions	Strong with EoT evaluation

VI. CONCLUSIONS

This paper investigates sparse adversarial patch attacks and robustness evaluation for visual-language models. The proposed SPA algorithm combines ADMM-based L0 sparse optimization with cross-attention masking to generate highly dispersed discrete perturbations in non-salient image regions. Compared with dense patches and random sparse perturbations, SPA achieves better visual concealment while maintaining high targeted attack success rates. The robustness evaluation framework further quantifies semantic deviation, target achievement, visual stealth, and physical transformation robustness.

The study still has limitations. The experiments mainly rely on open-source VLMs and publicly available datasets, so the results may not fully represent closed-source commercial systems. Physical robustness is simulated through transformations rather than large-scale real-world deployment. In addition, the design of target text and semantic evaluation may influence attack measurement, and future studies should further standardize generative-model robustness metrics.

Future research can extend sparse patch attacks to more complex multimodal agents, embodied navigation systems, and real physical scenarios. Defense mechanisms should combine attention anomaly detection, robust multimodal alignment, sparse perturbation filtering, and adversarial training. The findings of this paper reveal a new security vulnerability in VLMs and provide a quantitative benchmark for evaluating future multimodal defense strategies.

REFERENCES

[1] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in Advances in Neural Information Processing Systems, vol. 36, 2023.
[2] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in Int. Conf. Mach. Learn., PMLR, 2023, pp. 19730–19742.
[3] X. Zhou, M. Liu, E. Yurtsever, and Z. L. Zhang, "Vision language models in autonomous driving: A survey and outlook," arXiv preprint arXiv:2310.14414, 2023.
[4] S. Chen, Z. Liu, Q. Zhu, and H. Ma, "Exploring embodied multimodal large models," arXiv preprint arXiv:2502.15336, 2025.
[5] J. C. Costa, T. Roxo, H. Proenca, and P. N. Inacio, "How deep learning sees the world: A survey on adversarial attacks and defenses," arXiv preprint arXiv:2305.10862, 2023.
[6] H. Wei, J. Wang, X. Jia, and Y. Yang, "Physical adversarial attack meets computer vision: A decade survey," IEEE Trans. Pattern Anal. Mach. Intell., vol. 46, no. 12, pp. 1–20, 2024.
[7] D. Liu, M. Wang, J. Zhang, and J. Xu, "A survey of attacks on large vision-language models," arXiv preprint arXiv:2407.07403, 2024.
[8] D. Kong, S. Qiu, J. Ma, and Y. Chen, "Patch is enough: Naturalistic adversarial patch against vision-language pre-training models," Vis. Intell., vol. 2, p. 33, 2024.
[9] Q. Guo, X. Jia, S. Pang, and X. Zhang, "PhysPatch: A physically realizable and transferable adversarial patch attack for multimodal large language models-based autonomous driving systems," arXiv preprint arXiv:2508.05167, 2025.
[10] E. Shayegani, Y. Dong, and N. Abu-Ghazaleh, "Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models," in Int. Conf. Learn. Represent., 2024.
[11] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," Found. Trends Mach. Learn., vol. 3, no. 1, pp. 1–122, 2011.
[12] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in Int. Conf. Mach. Learn., PMLR, 2018, pp. 284–293.
[13] K. Kunanbayev, V. Shen, and D. S. Kim, "Training ViT with limited data for Alzheimer's disease classification: An empirical study," in Med. Image Comput. Comput. Assist. Interv., Cham: Springer, vol. 15012, 2024.
[14] H. Yang, M. Xu, Z. Sun, and J. Chen, "CLIP-ViT detector: Side adapter with prompt for vision transformer object detection," in Comput. Artif. Intell., 2024, pp. 1–8.
[15] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," ACM Comput. Surv., vol. 54, no. 10s, pp. 1–41, 2022.
[16] K. Steunou, T. Druilhe, and S. Saue, "Sparse representations improve adversarial robustness of neural network classifiers," arXiv preprint arXiv:2509.21130, 2025.
[17] W. Nam, J. Kim, and D. Kim, "RISOPA: Rapid imperceptible strong one-pixel attacks in deep neural networks," Mathematics, vol. 12, no. 7, p. 1083, 2024.
[18] K. Ding, K. Ma, S. Wang, and D. Fang, "Image quality assessment: Unifying structure and texture similarity," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 5, pp. 2567–2581, 2022.
[19] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," arXiv preprint arXiv:2310.03744, 2023.
[20] W. Dai, J. Li, D. Li, A. M. H. Tiong, and S. Hoi, "InstructBLIP: Towards general-purpose vision-language models with instruction tuning," in Adv. Neural Inf. Process. Syst., vol. 36, 2023.
[21] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "MiniGPT-4: Enhancing vision-language understanding with advanced large language models," arXiv preprint arXiv:2304.10592, 2023.
[22] Z. Yang, Z. Gan, J. Wang, and X. Hu, "Unified vision-language pre-training for image captioning and VQA," Proc. AAAI Conf. Artif. Intell., vol. 36, no. 11, pp. 13041–13049, 2022.
[23] F. Ishmam, A. Sadeghzadeh, R. Krishna, and M. Sadeghzadeh, "Visual robustness benchmark for visual question answering (VQA)," in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis., 2025.
[24] N. Zhang, W. Tao, X. Xiao, and X. Chen, "Attention-guided patch-wise sparse adversarial attacks on vision-language-action models," arXiv preprint arXiv:2511.21663, 2025.
[25] J. S. Xiao, H. Ma, R. D. Chen, and H. J. Pan, "STKPS-Net: Spatio-Temporal Key Patch Selection Network for few-shot anomalous action recognition," IEEE Trans. Inf. Forensics Security, vol. 21, pp. 827–838, 2026.