XSPA: CRAFTING IMPERCEPTIBLE X-SHAPED SPARSE ADVERSARIAL PERTURBATIONS FOR TRANSFERABLE ATTACKS ON VLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) rely on a shared visual-textual representation space to support tasks such as zero-shot classification, image captioning, and visual question answering (VQA). Although this shared space enables strong cross-task generalization, it also creates a critical vulnerability: subtle visual perturbations may propagate through the common embedding space and induce correlated semantic failures across tasks. This issue is especially concerning in interactive and decision-support scenarios, yet it remains unclear whether VLMs are still fragile under highly constrained, sparse, and geometrically fixed perturbations. We propose X-shaped Sparse Pixel Attack (XSPA), an imperceptible structured attack that restricts perturbations to two intersecting diagonal lines. Compared with dense perturbations and flexible localized patches, XSPA operates under a much stricter attack budget, offering a stringent test of VLM robustness. Within this sparse support, XSPA jointly optimizes a classification-oriented objective, cross-task semantic guidance, and regularization on perturbation magnitude and linewise smoothness, inducing both transferable misclassification and semantic drift in captioning and VQA while preserving visual subtlety. Under the default setting, XSPA modifies only about 1.04% of image pixels. Experiments on the COCO dataset show that XSPA consistently degrades performance across zero-shot classification, image captioning, and VQA. Zero-shot accuracy drops by 52.33 points on OpenAI CLIP ViT-L/14 and 67.00 points on OpenCLIP ViT-B/16, while GPT-4-evaluated caption consistency decreases by up to 58.60 points and VQA correctness by up to 44.25 points. These results show that even highly sparse and visually subtle perturbations with fixed geometric priors can substantially disrupt cross-task semantics in VLMs, revealing an important robustness gap in current multimodal systems.

1 INTRODUCTION

Recent advances in multimodal learning, driven by models such as CLIP and BLIP-2, have substantially improved cross-modal understanding in zero-shot classification, visual question answering, and open-ended generation Radford et al. (2021); Li et al. (2023). A key reason is the shared embedding space aligning visual and textual representations across downstream tasks. Yet this same mechanism also introduces a robustness concern. Perturbations affecting visual encoding may propagate through the shared semantic space and cause not only isolated prediction errors, but also cross-task semantic failures across classification, captioning, and question answering. This matters because a single corrupted input can affect multiple decisions at once.

Adversarial attack research has consistently shown that deep models are highly vulnerable to perturbations that are nearly imperceptible to humans Goodfellow et al. (2015); Madry et al. (2018). Existing attacks on visual models mainly fall into two categories: dense full-image perturbations under L_p constraints, which prioritize attack strength, and localized or physically motivated interventions, such as patches, shadows, or optical manipulations, which emphasize locality and real-world realizability Brown et al. (2018); Zhong et al. (2022); Hsiao et al. (2024). Although these studies provide valuable insights, they usually retain substantial freedom in perturbation support, shape, or

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

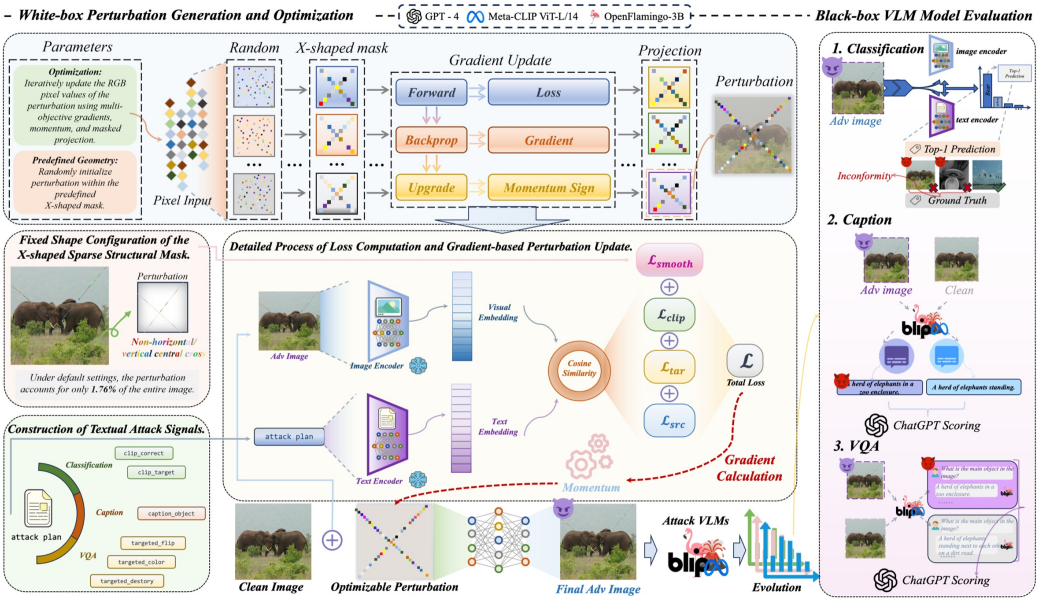

Figure 1: Overall framework of XSPA. XSPA optimizes perturbations within a fixed X-shaped sparse mask on a surrogate CLIP model, integrating momentum-based optimization, input diversity, and joint semantic objectives to promote target-oriented semantic attraction while suppressing source semantics. The resulting adversarial image is then transferred to downstream vision-language models for evaluation across zero-shot classification, image captioning, and visual question answering tasks.

optimization space. Such settings demonstrate attack effectiveness, but are less informative for a more fundamental question: how vulnerable are VLMs when perturbation support, geometry, and optimization freedom are all restricted?

This question is particularly important for vision-language models. Unlike unimodal classifiers, VLMs must not only produce correct visual predictions, but also preserve semantic alignment across multiple downstream tasks. As a result, robustness evaluation under extremely flexible perturbation settings may obscure whether cross-task semantic behavior can already be destabilized by visual signals that are both extremely low-budget and strongly structured. Characterizing this vulnerability boundary is valuable in two respects: scientifically, it clarifies how much geometric freedom is required to disrupt multimodal alignment; practically, it provides a controlled stress test for future robustness evaluation and defense design.

Motivated by this perspective, we study whether a perturbation with extremely limited support and fixed geometry is already sufficient to induce significant cross-task semantic disruption in VLMs. Rather than treating perturbation shape as a free variable, we impose a stringently constrained geometric prior and use it as a controlled probe of multimodal robustness. Concretely, we propose X-shaped Sparse Pixel Attack (XSPA), which confines perturbations to a sparse X-shaped structure formed by two intersecting diagonal line segments. Within this restricted region, XSPA integrates transfer-enhancing optimization strategies Xie et al. (2019) with joint semantic objectives for zero-shot classification, image captioning, and visual question answering. Beyond inducing label errors, the optimization promotes target-oriented semantic attraction and suppresses source semantics, thereby encouraging semantic drift across downstream generative tasks. Figure 1 illustrates this pipeline, from sparse perturbation optimization on the surrogate CLIP model to transfer-based evaluation on downstream vision-language tasks.

We evaluate XSPA on four CLIP-style visual encoders and six downstream VLMs, using LLM-as-a-Judge style automatic evaluators for open-ended tasks Zheng et al. (2023); OpenAI (2024). The results show that XSPA markedly degrades performance on zero-shot classification, image captioning, and visual question answering under a severely restricted perturbation budget. These findings

suggest that even when perturbations are confined to an extremely small region with a fixed geometric prior, VLMs remain vulnerable to cross-task semantic disruption. More broadly, our results show that multimodal robustness cannot be understood only through dense noise or flexible local attacks; it must also account for the fragility of shared semantic alignment under strongly constrained structured perturbations. In this sense, XSPA serves not only as an effective attack, but also as a controlled diagnostic tool for probing how little geometric freedom is needed to destabilize multimodal behavior. This perspective helps connect attack evaluation with a more precise characterization of vulnerability boundaries in shared visual-textual representation spaces. The main contributions of this work are as follows.

- To the best of our knowledge, we propose the first X-shaped sparse pixel attack (XSPA) that constrains perturbations to a fixed geometric structure with extremely limited support. The method unifies momentum-based optimization, input diversity, cross-task semantic guidance, and source-semantic suppression within a highly restricted two-diagonal-line perturbation region.
- We conduct extensive experiments across zero-shot classification, image captioning, and visual question answering tasks, demonstrating that XSPA consistently achieves substantial cross-task semantic disruption on four CLIP-style encoders and six downstream vision-language models under minimal perturbation budgets.
- We perform comprehensive ablation studies that reveal how highly constrained structured perturbations with fixed geometry can still induce strong multimodal misalignment, providing insights into the vulnerability boundaries of shared visual-textual representation spaces.

2 Related Work

2.1 Sparse and Structured Attacks on Vision Models

Adversarial attacks on vision models provide the methodological basis for studying attacks on vision-language models. Early work established that deep networks remain vulnerable under diverse threat models Goodfellow et al. (2015); Madry et al. (2018); Moosavi-Dezfooli et al. (2016); Xie et al. (2019); Andriushchenko et al. (2020). Most relevant to our work are sparse and structured attacks that modify only a few pixels or constrained regions Su et al. (2019); Modas et al. (2019); Zhong & Liu (2025), and physically inspired attacks using patches, shadows, or natural illumination Brown et al. (2018); Ran et al. (2023); Zhong et al. (2022); Huang & Ling (2022); Hsiao et al. (2024). However, most existing methods still optimize support size, patch area, or physical realizability rather than testing highly constrained geometric priors; by contrast, we fix an X-shaped sparse structure and test whether it alone suffices to induce cross-task semantic deviation.

2.2 Attacks and Robustness Studies on Vision-Language Models

With the rise of CLIP-style vision-language pre-training, robustness in shared cross-modal embedding spaces has become a central issue Radford et al. (2021). Prior work shows that adversarial perturbations can transfer across retrieval, classification, and generation through the shared representation space, and that large VLMs remain vulnerable in both classification-oriented and open-ended settings Zhang et al. (2022); Guo et al. (2023); Zhao et al. (2023); Cui et al. (2024). Recent work has improved attack transferability and unified cross-task design Zhang et al. (2025); Xie et al. (2025); Liu et al. (2025b); Zhao et al. (2025); Liu et al. (2025a); Li et al. (2026a;b), while robustness studies explore certification, adaptation, and test-time tuning Wang et al. (2024); Schlarmann et al. (2024); Tong et al. (2025); Sheng et al. (2025); Dong et al. (2025); Mirza et al. (2026).

2.3 Automatic Evaluation for Open-Ended Tasks

For zero-shot classification, adversarial effectiveness can be measured directly by accuracy or attack success rate, but for open-ended tasks such as image captioning and visual question answering, exact-match and lexical-overlap metrics often fail to capture semantically meaningful changes Zhao et al. (2023); Cui et al. (2024). We therefore adopt a GPT-4-based automatic evaluator that assesses semantic correctness and factual consistency, providing a more faithful measurement of whether

perturbations alter the model's understanding rather than merely changing surface wording. This distinction is essential in cross-task attack evaluation: a perturbation may preserve surface fluency while changing the predicted concept, answer, or factual relation. Combining task-specific metrics with semantic judging therefore allows classification, captioning, and VQA outcomes to be compared under a unified evaluation protocol Zheng et al. (2023); OpenAI (2024).

3 METHOD

3.1 PROBLEM DEFINITION

Given an input image $x \in [0,1]^{3\times H\times W}$, XSPA optimizes a perturbation δ restricted to a fixed X-shaped binary mask $M \in \{0,1\}^{H\times W}$:

$$x^{adv} = \mathrm{clip}(x + M \odot \delta, 0, 1), \tag{1}$$

where $\odot$ denotes element-wise multiplication. We consider a white-box surrogate / black-box downstream setting: the perturbation is optimized on a CLIP-style surrogate and transferred to downstream models for zero-shot classification, image captioning, and visual question answering. Thus, unlike full-image L_p-bounded attacks, XSPA modifies only a highly sparse predefined support.

3.2 X-SHAPED SPARSE STRUCTURAL MASK

XSPA uses two oblique line segments crossing at $c_{\mathrm{ctr}} = (\rho_{\mathrm{col}}W, \rho_{\mathrm{row}}H)$. Given their angles $\mathcal{A} = \{\theta_1, \theta_2\}$, relative length r, and width b, the mask support is

$$\Omega_M = \bigcup_{\theta\in\mathcal{A}} \mathcal{N}_b\big(\ell(c_{\mathrm{ctr}}, \theta, r)\big), \tag{2}$$

where $\ell(\cdot)$ is a line segment whose length is proportional to the shorter image side and $\mathcal{N}_b(\cdot)$ expands it to width b. The binary mask and its sparsity ratio are

$$M_{h,w} = \mathbf{1}[(h,w) \in \Omega_M], \qquad \eta_M = \frac{|\Omega_M|}{HW}. \tag{3}$$

The mask is obtained by rasterizing the two centerlines and expanding their sampled coordinates locally. Under the default 384×384 setting, the 1,532-pixel support gives $\eta_M \approx 1.04\%$. Further construction details are provided in the Appendix.

Fixing the support before optimization separates geometric effects from unconstrained pixel selection and prevents the optimizer from allocating perturbations to visually convenient regions. Consequently, transferability must arise from the learned values within the prescribed X structure rather than from adaptive support search.

3.3 CROSS-TASK JOINT OBJECTIVE

On the fixed support, XSPA jointly optimizes classification attack signals, cross-task semantic guidance, and structural regularization:

$$\mathcal{L} = \lambda^{(t)}_{clip}\mathcal{L}_{clip} + \lambda^{(t)}_{tar}\mathcal{L}_{tar} + \lambda^{(t)}_{src}\mathcal{L}_{src} + \mathcal{L}_{smooth}. \tag{4}$$

Let $f_v(\cdot)$ and $f_t(\cdot)$ be the surrogate CLIP image and text encoders, $v = f_v(x^{adv})$ the normalized image feature, and $z_k = \langle v, f_t(s_k)\rangle/\tau$ the logit for class prompt s_k. We use a margin loss for untargeted attacks and cross-entropy for targeted attacks:

$$\mathcal{L}_{clip} = \begin{cases} z_y - \max_{k\neq y} z_k, & \text{untargeted,} \\ \mathrm{CE}(z, y_t), & \text{targeted,} \end{cases} \tag{5}$$

where y and y_t denote the source and target classes.

To extend the attack beyond label prediction, we construct a target text pool $\mathcal{T}_{tar} = \{p_u\}_{u=1}^{N_t}$ and a source text pool $\mathcal{T}_{src} = \{q_w\}_{w=1}^{N_s}$ from target labels, caption-side semantic drifts, and VQA-side

answer shifts. Their losses are

$$\mathcal{L}_{tar} = -\frac{1}{N_t}\sum_{u=1}^{N_t}\langle v, f_t(p_u)\rangle,$$
$$\mathcal{L}_{src} = \frac{1}{N_s}\sum_{w=1}^{N_s}\langle v, f_t(q_w)\rangle. \tag{6}$$

The first term attracts the adversarial representation toward target semantics, whereas the second suppresses source semantics. Combining both pools avoids relying on a single class label and exposes the surrogate representation to task-relevant concepts shared by classification, captioning, and VQA. Because all prompts are encoded by the same frozen text encoder, these signals remain comparable in one embedding space and define a task-agnostic semantic direction that can transfer beyond the surrogate classifier. To avoid sharp local oscillations on the thin support, we further use

$$\mathcal{L}_{smooth} = \lambda^{(t)}_{mag}\mathcal{L}_{mag} + \lambda^{(t)}_{line}\mathcal{L}_{line}, \tag{7}$$

where

$$\mathcal{L}_{mag} = \frac{1}{|\Omega_M|}\sum_{(h,w)\in\Omega_M}\|\delta_{h,w}\|_2^2,$$
$$\mathcal{L}_{line} = \frac{1}{|\mathcal{P}|}\sum_{P\in\mathcal{P}}\frac{1}{|P|-1}\sum_{n=2}^{|P|}\|\delta_{P_n} - \delta_{P_{n-1}}\|_2^2. \tag{8}$$

The magnitude term controls isolated perturbation spikes, whereas the line term couples neighboring pixels along each ordered centerline path $P \in \mathcal{P}$. Their combination preserves the sparse geometry while discouraging visually abrupt changes. For each $\nu \in \{clip, tar, src, mag, line\}$, the corresponding weight follows

$$\lambda^{(t)}_{\nu} = \begin{cases}\lambda^{A}_{\nu}, & t < \gamma N,\\ \lambda^{B}_{\nu}, & t \geq \gamma N.\end{cases} \tag{9}$$

Early iterations emphasize classification guidance to stabilize the attack direction, whereas later iterations strengthen semantic misalignment to promote cross-task transfer.

3.4 Mask-constrained Optimization

The perturbation is restricted to the feasible set

$$\mathcal{C}(M,\epsilon) = \{\delta \mid (1-M)\odot\delta = 0,\ \|\delta\|_\infty \leq \epsilon\}. \tag{10}$$

At iteration t, we apply random resize-and-pad input diversity $D(\cdot)$, compute the joint gradient, accumulate MI-FGSM-style momentum, and project the update onto $\mathcal{C}(M,\epsilon)$:

$$g^{(t)} = \nabla_\delta\mathcal{L}\Big(D(x^{adv,(t)})\Big),$$
$$m^{(t+1)} = \mu m^{(t)} + \frac{g^{(t)}}{\text{mean}(|g^{(t)}|) + 10^{-12}}, \tag{11}$$
$$\delta^{(t+1)} = M\odot\text{clip}\Big(\delta^{(t)} - \alpha\ \text{sign}(m^{(t+1)}), -\epsilon, \epsilon\Big).$$

Input diversity reduces surrogate overfitting, while momentum stabilizes the update direction. The projection keeps every modification inside the X-shaped support. The complete procedure and pseudocode are included in the Appendix.

4 Experiments

4.1 Experimental Setup

Dataset. Following ITA's evaluation protocol, we select 300 images from COCO-80 Lin et al. (2014) as a unified evaluation set. All three tasks are evaluated on the same dataset to enable consistent comparison across attack methods. For zero-shot classification, we use the COCO-80 category set as the semantic label space.

Table 1: Zero-shot classification results under different adversarial attacks. Lower ACC and higher naturalness are better. Numbers in parentheses denote the absolute change relative to clean samples. Bold indicates the best result.

	OpenCLIP ViT-B/16		Meta-CLIP ViT-L/14		EVA-CLIP ViT-G/14		OpenAI CLIP ViT-L/14	
Method	ACC(%)	GPT-4o	ACC(%)	GPT-4o	ACC(%)	GPT-4o	ACC(%)	GPT-4o
Clean	97	3.65	98	3.65	98	3.65	93	3.65
NaturalLightAttack	94(↓3)	2.32(↓1.33)	97(↓1)	2.14(↓1.51)	97(↓1)	2.16(↓1.49)	93(-)	2.07(↓1.58)
ShadowAttack	84(↓13)	2.11(↓1.54)	82(↓16)	1.91(↓1.74)	95(↓3)	2.03(↓1.62)	79(↓14)	1.78(↓1.87)
ITA	46(↓51)	2.11(↓1.54)	64(↓34)	2.15(↓1.50)	84(↓14)	2.17(↓1.48)	51(↓42)	2.15(↓1.50)
Ours	**30(↓67)**	**2.71(↓0.94)**	**60.67(↓37.33)**	**2.89(↓0.76)**	**63.71(↓34.29)**	**2.78(↓0.87)**	**40.67(↓52.33)**	**2.76(↓0.89)**

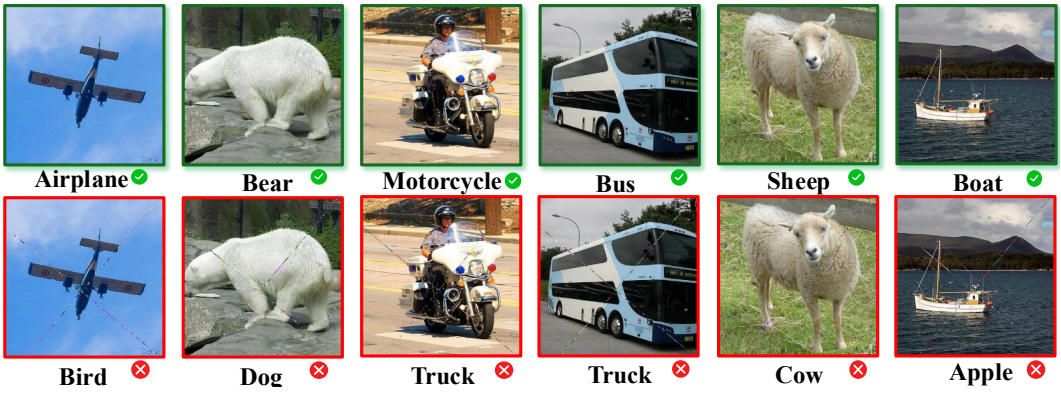

Figure 2: Qualitative zero-shot classification examples. The top row shows clean images, and the bottom row shows XSPA adversarial images. Despite the fixed X-shaped sparse perturbation, clear prediction shifts are observed.

Models. For zero-shot classification, we evaluate four representative CLIP-style visual encoders: OpenCLIP ViT-B/16 Cherti et al. (2023), Meta-CLIP ViT-L/14 Xu et al. (2024), EVA-CLIP ViT-G/14 Sun et al. (2023), and OpenAI CLIP ViT-L/14 Radford et al. (2021). For downstream multimodal tasks, we further evaluate six widely used vision-language models, including LLaVA-1.5, LLaVA-1.6 Liu et al. (2023), OpenFlamingo Awadalla et al. (2023), BLIP-2 (FlanT5XL ViT-L), BLIP-2 (FlanT5XL) Li et al. (2023), and InstructBLIP (FlanT5XL) Dai et al. (2023). This diverse model suite allows us to assess the transferability of XSPA across different encoder families, multimodal architectures, and downstream task settings.

Baselines. We compare XSPA with NaturalLightAttack, ShadowAttack, and ITA Hsiao et al. (2024); Zhong et al. (2022); Liu et al. (2025a), covering natural-light, shadow-based, and illumination-transformation attacks. Because their threat models do not perfectly match the fixed sparse geometry imposed by XSPA, these comparisons should be interpreted as robustness-oriented references rather than strict like-for-like rankings. Our goal is to determine whether VLMs remain vulnerable when perturbation support is restricted to a substantially simpler, sparser, and more constrained geometric structure.

Evaluation Metrics. For zero-shot classification, we report Top-1 accuracy and its absolute drop relative to clean samples. T-ASR denotes the proportion of predictions driven toward a predefined target, whereas U-ASR measures untargeted misclassification; higher values indicate stronger attacks. GPT-4o evaluates image naturalness using a 4-point rubric covering visual naturalness, physical consistency, and adversarial plausibility. For captioning and VQA, GPT-4 evaluates response consistency and correctness under a fixed LLM-as-a-Judge prompt and rubric shared across all methods and models Zheng et al. (2023); OpenAI (2023; 2024), ensuring consistent evaluation conditions.

Implementation Details. Main experiments use 200 optimization iterations, while ablation studies follow the frozen protocol described below. XSPA is optimized on a CLIP-style surrogate model and then transferred directly to downstream VLMs without additional adaptation. All experiments are conducted on a single NVIDIA RTX 4090 24GB GPU; the remaining optimization settings and loss scheduling follow Section 3 to ensure reproducibility and consistency across experiments.

Table 2: Image captioning results on six vision-language models under different adversarial attacks. Lower GPT-4 consistency indicates more effective attacks. Numbers in parentheses denote the absolute drop relative to clean samples. Bold indicates the best result.

Image Encoder	Model	Params	Clean	NaturalLightAttack	ShadowAttack	ITA	Ours
OpenAI CLIP ViT-L/14	LLaVA-1.5	7B	78.60	77.00(↓1.60)	74.60(↓4.00)	63.73(↓14.87)	**46.40(↓32.20)**
	LLaVA-1.6	7B	72.10	71.70(↓0.40)	71.17(↓0.93)	61.60(↓10.50)	**44.20(↓27.90)**
EVA-CLIP ViT-G/14	OpenFlamingo	3B	70.20	69.53(↓0.67)	67.80(↓2.40)	53.93(↓16.27)	**11.60(↓58.60)**
	BLIP-2 (FlanT5XL ViT-L)	3.4B	75.10	70.77(↓4.33)	68.57(↓6.53)	60.93(↓14.17)	**21.90(↓53.20)**
	BLIP-2 (FlanT5XL)	4.1B	74.96	71.27(↓3.69)	68.80(↓6.16)	62.01(↓12.95)	**59.00(↓15.96)**
	InstructBLIP (FlanT5XL)	4.1B	76.50	72.07(↓4.43)	69.77(↓6.73)	63.20(↓13.30)	**40.90(↓35.60)**

Table 3: VQA results on six vision-language models under different adversarial attacks. Lower GPT-4 correctness indicates more effective attacks. Numbers in parentheses denote the absolute drop relative to clean samples. Bold indicates the best result.

Image Encoder	Model	Params	Clean	NaturalLightAttack	ShadowAttack	ITA	Ours
OpenAI CLIP ViT-L/14	LLaVA-1.5	7B	68.00	68.00(-)	67.00(↓1.00)	48.00(↓20.00)	**47.45(↓20.55)**
	LLaVA-1.6	7B	64.00	63.00(↓1.00)	64.00(-)	**43.00(↓21.00)**	48.87(↓15.13)
EVA-CLIP ViT-G/14	OpenFlamingo	3B	45.00	39.00(↓6.00)	44.00(↓1.00)	19.00(↓26.00)	**14.29(↓30.71)**
	BLIP-2 (FlanT5XL ViT-L)	3.4B	63.00	58.00(↓5.00)	50.00(↓13.00)	38.00(↓25.00)	**18.75(↓44.25)**
	BLIP-2 (FlanT5XL)	4.1B	54.00	53.00(↓1.00)	54.00(-)	33.00(↓21.00)	**19.78(↓34.22)**
	InstructBLIP (FlanT5XL)	4.1B	68.00	64.00(↓4.00)	62.00(↓6.00)	**44.00(↓24.00)**	44.67(↓23.33)

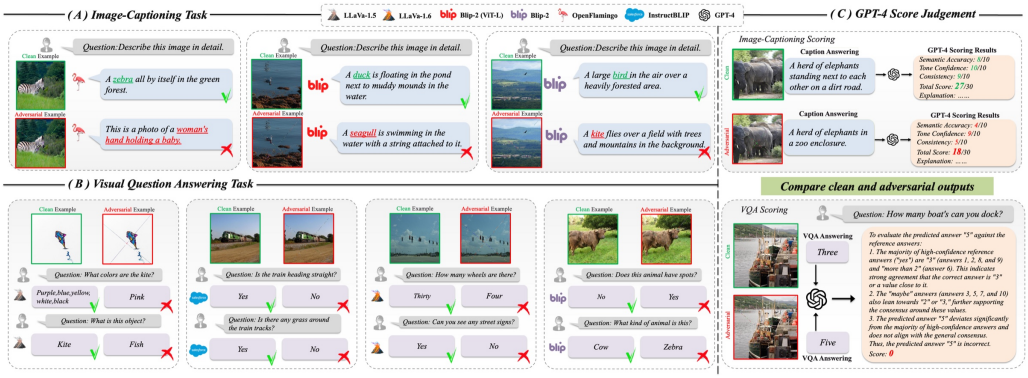

Figure 3: Qualitative examples on image captioning and visual question answering. Compared with clean outputs, XSPA adversarial outputs exhibit clear semantic drift in captions and more error-prone answers in VQA, leading to lower GPT-4-based scores.

4.2 Main Results

Zero-Shot Classification. Table 1 summarizes the zero-shot classification results. Under this protocol, XSPA causes the largest accuracy drop on all four CLIP-style visual encoders while also achieving the highest GPT-4o naturalness scores. Specifically, it reduces the Top-1 accuracy to 30.00%, 60.67%, 63.71%, and 40.67% on OpenCLIP ViT-B/16, Meta-CLIP ViT-L/14, EVA-CLIP ViT-G/14, and OpenAI CLIP ViT-L/14, respectively. Relative to ITA, XSPA further decreases accuracy by 16.00, 3.33, 20.29, and 10.33 percentage points. These results show that even under a highly constrained X-shaped sparse support, CLIP image-text alignment can still be substantially disrupted without introducing the most visible artifacts. Figure 2 further shows qualitative zero-shot classification examples, where clean images are classified correctly, whereas XSPA adversarial images induce clear misclassification under a thin X-shaped perturbation while preserving a largely natural visual appearance, consistent with Table 1.

Image Captioning. Table 2 presents captioning results. Under the judge-based metric, XSPA yields the lowest consistency across six models, demonstrating disruption beyond classification in open-ended generation. Consistency falls to 46.40% and 44.20% on LLaVA-1.5 and LLaVA-1.6, and to 11.60%, 21.90%, 59.00%, and 40.90% on OpenFlamingo, BLIP-2 (FlanT5XL ViT-L), BLIP-2 (FlanT5XL), and InstructBLIP (FlanT5XL), respectively. OpenFlamingo shows the largest decrease, dropping 58.60 points from clean samples. The attack transfers across instruction-tuned

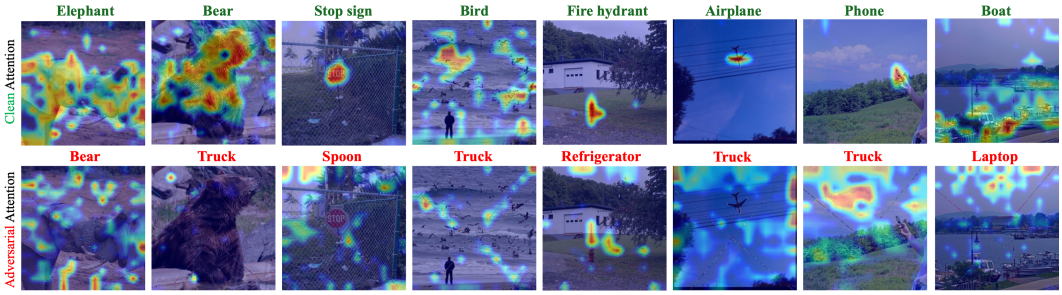

Figure 4: Grad-CAM heatmaps for clean and adversarial samples. Adversarial responses become more diffuse or shift away from the main object.

and non-instruction-tuned systems: the strongest degradation appears on OpenFlamingo and BLIP-2 (FlanT5XL ViT-L), while LLaVA-1.5, LLaVA-1.6, and InstructBLIP still exhibit substantial drops. In many cases, generated captions no longer preserve object category, local attributes, or scene relations, even though the perturbation remains visually sparse and structurally constrained; this tendency is consistent across both short and descriptive captions. Overall, the results suggest that sparse geometric perturbations can propagate from visual features to sentence-level semantics and induce transferable caption drift across diverse model families.

Visual Question Answering. Table 3 reports the VQA results. XSPA achieves the lowest correctness on four of the six models and remains competitive on the other two, indicating that its semantic objectives extend to question-conditioned generation. Specifically, correctness falls to 47.45% and 48.87% on LLaVA-1.5 and LLaVA-1.6, and to 14.29%, 18.75%, 19.78%, and 44.67% on OpenFlamingo, BLIP-2 (FlanT5XL ViT-L), BLIP-2 (FlanT5XL), and InstructBLIP (FlanT5XL), respectively. The largest drop, 44.25 percentage points, occurs on BLIP-2 (FlanT5XL ViT-L). Compared with NaturalLightAttack, ShadowAttack, and ITA, XSPA performs best on three of the four EVA-CLIP-based models, while ITA is slightly stronger on LLaVA-1.6 and InstructBLIP. The trend holds across dialogue-style and encoder-decoder VLMs, suggesting that XSPA disrupts visual evidence used for answer selection rather than merely degrading text fluency. Figure 3 provides qualitative evidence for both tasks. Clean captions generally preserve the main object, scene, and attributes, whereas adversarial captions drift in object identity, material, or fine-grained attributes. Adversarial VQA outputs produce more category, color, count, and local-attribute errors across decoder families, including on simple questions. This indicates that semantic disruption is not confined to difficult or ambiguous prompts. The GPT-4-based judgments support transfer from the surrogate CLIP space to downstream generation and reasoning.

4.3 Grad-CAM-Based Attention Heatmap Analysis

We use Grad-CAM to compare attention in clean and adversarial samples Selvaraju et al. (2017). As shown in Figure 4, clean images concentrate responses on semantically decisive object regions, whereas XSPA produces more diffuse or displaced activation across both foreground-dominant and small-target scenes. This shift is most evident when clean maps form compact peaks, whereas adversarial maps fragment into several weaker responses distributed across less relevant spatial regions. The pattern spans large objects such as elephants and bears and localized targets such as stop signs, airplanes, phones, and boats, indicating that the attention shift is not confined to a particular object scale or scene complexity. Despite being confined to a fixed sparse X-shaped support, the perturbation redirects attention toward background or irrelevant structures, suggesting that its influence extends beyond the modified pixels to internal discriminative features. This redistribution is consistent with the performance degradation in our main experiments: XSPA disrupts the visual evidence used for recognition across varied object scales and scene complexity while largely preserving natural appearance. Together, these shifts consistently weaken object-centered spatial evidence across examples. The effect spans diverse scenes.

4.4 Ablation Studies

We study optimization and perturbation budgets, geometry and shape, loss design, and cross-dataset behavior. Unless varied, ablations use OpenCLIP ViT-B/16 on 300 images with 100 steps, a 0.03 step size, $\epsilon = 0.25$, and 1,532 pixels.

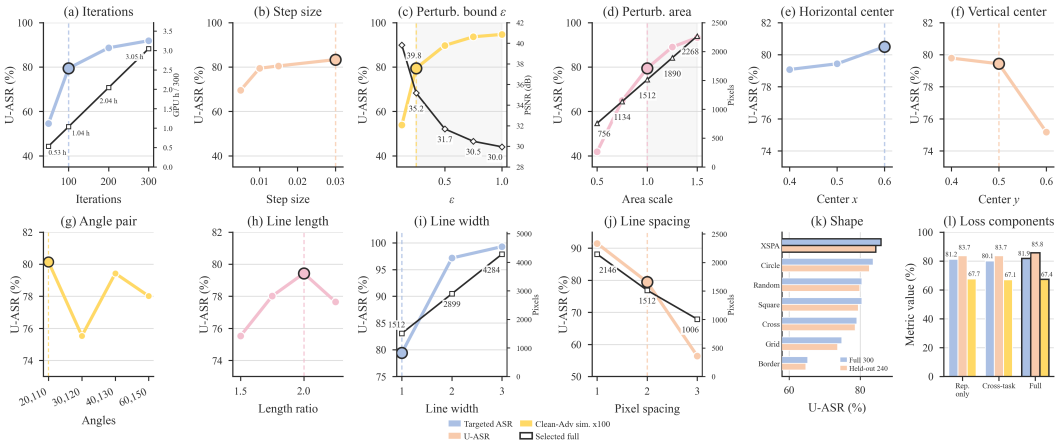

Figure 5: Unified 2-row by 6-column ablations: (a)–(d) optimization and perturbation, (e)–(j) X geometry, (k) shape, and (l) loss components.

Optimization, Geometry, Shape, and Loss Design. Figure 5 consolidates twelve ablations into a grid with two rows and six columns. Panels (a)–(d) show that more iterations, a larger perturbation bound, or a larger area strengthen the attack, but increase GPU time, reduce PSNR, or modify more pixels. The saturation at larger iteration counts also indicates diminishing returns, supporting the selected operating point as a practical compromise between optimization cost, perturbation sparsity, and attack effectiveness for downstream evaluation. Panels (e)–(j) show moderate sensitivity to center, angle, and length, while larger line width or smaller pixel spacing sharply increases U-ASR together with support density. Center and angle sweeps remain flat, whereas width and spacing change sharply, identifying support density as the dominant geometric factor and exposing a trade-off between attack success and feature preservation. This separation helps distinguish robust design choices from budget-driven gains. Panel (k) compares seven shapes under the same 1,532-pixel budget at the final step size of 0.03; XSPA achieves the highest U-ASR on both Full 300 and Held-out 240. Panel (l) shows that no loss alternative dominates targeted ASR, U-ASR, and feature separation simultaneously, so the selected configuration balances attack strength, cost, and sparsity.

Cross-Dataset and Multi-Encoder Behavior. Table **??** reports U-ASR on our original COCO-300 evaluation set and three larger evaluation sets across four CLIP-style encoders. Performance is highest on ImageNet-1K but lower on COCO and ImageNet-R, suggesting that dataset composition and encoder pretraining shape transferability rather than image count alone under the same attack protocol. XSPA remains effective beyond the original 300-image setting, while the variation across datasets and encoders makes the transfer boundary explicit rather than implying uniform vulnerability.

5 Conclusion

This paper presents XSPA, a sparse X-shaped adversarial attack that integrates CLIP supervision, cross-task guidance, source-semantic suppression, and smoothness regularization. Across four CLIP encoders and multiple VLMs, XSPA consistently degrades zero-shot classification, image captioning, and visual question answering on COCO, demonstrating strong cross-model and cross-task transferability. Extensive ablations further examine optimization budgets, perturbation limits, geometric design, and loss components. The results expose the vulnerability of VLMs to geometrically constrained semantic perturbations and highlight the need for structure-aware defenses, stronger robustness evaluation, and more reliable multimodal representation learning.

AI Assistance Statement. AI tools were used only for language polishing, including grammar, clarity, and stylistic refinement. They were not used to generate scientific claims, data, experimental results, figures, tables, or references. The authors remain fully responsible for the content and accuracy of the manuscript.

REFERENCES

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: A query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision*, Cham, Switzerland, 2020. Springer. doi: 10.1007/978-3-030-58592-1_29. URL https://arxiv.org/pdf/1912.00049.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint, 2023.

Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. In *NeurIPS Workshop*, Montreal, Canada, 2018. NeurIPS Foundation. doi: 10.48550/arxiv.1712.09665.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, Los Alamitos, CA, USA, 2023. IEEE Computer Society. doi: 10.1109/CVPR52729.2023.00276. URL https://openaccess.thecvf.com/content/CVPR2023/html/Cherti_Reproducible_Scaling_Laws_for_Contrastive_Language-Image_Learning_CVPR_2023_paper.html.

Xuanming Cui, Alejandro Aparcedo, Young Kyun Jang, and Ser-Nam Lim. On the robustness of large multimodal models against image adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24625–24634, Los Alamitos, CA, USA, 2024. IEEE Computer Society. doi: 10.1109/cvpr52733.2024.02325.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. arXiv preprint, 2023.

Junhao Dong, Liaoyuan Feng, Piotr Koniusz, Qinghua Yao, Yonglong Tian, Hao Zhu, Weiming Liu, and Yew-Soon Ong. Improving zero-shot adversarial robustness in vision-language models by closed-form alignment of adversarial path simplices. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10456–10476, Online, 2025. PMLR. URL https://proceedings.mlr.press/v267/dong25f.html.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9185–9193, Los Alamitos, CA, USA, 2018. IEEE Computer Society. doi: 10.1109/CVPR.2018.00957. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Dong_Boosting_Adversarial_Attacks_CVPR_2018_paper.html.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, Online, 2015. OpenReview.net. doi: 10.48550/arxiv.1412.6572.

Yujun Guo, Jiaming Zhang, YuGang Jiang, and Jitao Sang. 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, Los Alamitos, CA, USA, 2023. IEEE Computer Society. URL https://arxiv.org/pdf/2307.14061.

Teng-Fang Hsiao, Bo-Lun Huang, Zi-Xiang Ni, Yan-Ting Lin, Hong-Han Shuai, YungHui Li, and HaoWen Cheng. Natural light can also be dangerous: Traffic sign misinterpretation under adversarial natural light attacks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, Los Alamitos, CA, USA, 2024. IEEE Computer Society. doi: 10.1109/wacv57701.2024.00387.

Bingyao Huang and Haibin Ling. Spaa: Stealthy projector-based adversarial attacks on deep image classifiers. In *IEEE Conference on Virtual Reality and 3D User Interfaces*, Los Alamitos, CA, USA, 2022. IEEE. doi: 10.1109/vr51125.2022.00073.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, Online, 2023. PMLR. doi: 10.48550/arxiv.2301.12597.

Yang Li, Jia-Li Yin, Luojun Lin, and Wei Lin. Transform to transfer: Boosting adversarial attack transferability on vision-language pre-training models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30141–30150, June 2026a.

Yuanbo Li, Tianyang Xu, Cong Hu, Tao Zhou, Xiao-Jun Wu, and Josef Kittler. Multi-paradigm collaborative adversarial attack against multi-modal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30065–30075, June 2026b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, Cham, Switzerland, 2014. Springer. doi: 10.1007/978-3-319-10602-1_48.

Hanqing Liu, Shouwei Ruan, Yao Huang, Shiji Zhao, and Xingxing Wei. When lighting deceives: Exposing vision-language models' illumination vulnerability through illumination transformation attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10485–10495, Los Alamitos, CA, USA, 2025a. IEEE Computer Society. URL https://openaccess.thecvf.com/content/ICCV2025/html/Liu_When_Lighting_Deceives_Exposing_Vision-Language_Models_Illumination_Vulnerability_Through_Illumination_ICCV_2025_paper.html.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, Red Hook, NY, USA, 2023. Curran Associates, Inc. doi: 10.48550/arxiv.2304.08485.

Yunqi Liu, Xue Ouyang, and Xiaohui Cui. Gleam: Enhanced transferable adversarial attacks for vision-language pre-training models via global-local transformations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1665–1674, Los Alamitos, CA, USA, 2025b. IEEE Computer Society. URL https://openaccess.thecvf.com/content/ICCV2025/html/Liu_GLEAM_Enhanced_Transferable_Adversarial_Attacks_for_Vision-Language_Pre-training_Models_via_ICCV_2025_paper.html.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, Online, 2018. OpenReview.net. doi: 10.48550/arxiv.1706.06083.

Mujtaba Hussain Mirza, Antonio D'Orazio, Odelia Melamed, and Iacopo Masi. A provable energy-guided test-time defense boosting adversarial robustness of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8598–8609, June 2026.

Apostolos Modas, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Sparsefool: A few pixels make a big difference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9087–9096, Los Alamitos, CA, USA, 2019. IEEE Computer Society. doi: 10.1109/cvpr.2019.00930. URL https://arxiv.org/pdf/1811.02248.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2574–2582, Los Alamitos, CA, USA, 2016. IEEE Computer Society. URL https://arxiv.org/pdf/1511.04599.

OpenAI. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023. URL https://arxiv.org/abs/2303.08774.

OpenAI. Gpt-4o system card. OpenAI system card, 2024. URL https://openai.com/index/gpt-4o-system-card/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, Online, 2021. PMLR. doi: 10.48550/arxiv.2103.00020.

Yu Ran, Weijia Wang, Mingjie Li, Lin-Cheng Li, YuanGen Wang, and Jin Li. Cross-shaped adversarial patch attack. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4): 2289–2303, 2023. doi: 10.1109/TCSVT.2023.3307150. URL https://doi.org/10.1109/TCSVT.2023.3307150.

Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 43685–43704, Online, 2024. PMLR. URL https://proceedings.mlr.press/v235/schlarmann24a.html.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, Los Alamitos, CA, USA, 2017. IEEE Computer Society. doi: 10.1109/ICCV.2017.74. URL https://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html.

Lijun Sheng, Jian Liang, Zilei Wang, and Ran He. R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29958–29967, Los Alamitos, CA, USA, 2025. IEEE Computer Society. doi: 10.1109/CVPR52734.2025.02788. URL https://openaccess.thecvf.com/content/CVPR2025/html/Sheng_R-TPT_Improving_Adversarial_Robustness_of_Vision-Language_Models_through_Test-Time_Prompt_CVPR_2025_paper.html.

Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019. doi: 10.1109/tevc.2019.2890858. URL https://arxiv.org/pdf/1710.08864.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. arXiv preprint, 2023.

Baoshun Tong, Hanjiang Lai, Yan Pan, and Jian Yin. On the zero-shot adversarial robustness of vision-language models: A truly zero-shot and training-free approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19921–19930, Los Alamitos, CA, USA, 2025. IEEE Computer Society. doi: 10.1109/CVPR52734.2025.01854. URL https://openaccess.thecvf.com/content/CVPR2025/html/Tong_On_the_Zero-shot_Adversarial_Robustness_of_Vision-Language_Models_A_Truly_CVPR_2025_paper.html.

Yanting Wang, Hongye Fu, Wei Zou, and Jinyuan Jia. Mmcert: Provable defense against adversarial attacks to multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24655–24664, Los Alamitos, CA, USA, 2024. IEEE Computer Society. doi: 10.1109/CVPR52733.2024.02328. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wang_MMCert_Provable_Defense_against_Adversarial_Attacks_to_Multi-modal_Models_CVPR_2024_paper.html.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jihong Wang, Zhou Ren, and Alan L. Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2730–2739, Los Alamitos, CA, USA, 2019. IEEE Computer Society. doi: 10.1109/CVPR.2019.00284. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Xie_Improving_Transferability_of_Adversarial_Examples_With_Input_Diversity_CVPR_2019_paper.html.

Peng Xie, Yequan Bie, Jiazhen 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11970–11979, Los Alamitos, CA, USA, 2025. IEEE Computer Society. doi: 10.1109/cvpr52734.2025.01368. URL https://arxiv.org/pdf/2411.15720.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. In *International Conference on Learning Representations*, Vienna, Austria, 2024. OpenReview.net. doi: 10.48550/arXiv.2309.16671. URL https://openreview.net/forum?id=htWIIXnT7K.

Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models. In *Proceedings of the ACM International Conference on Multimedia*, pp. 5005–5013, New York, NY, USA, 2022. Association for Computing Machinery. doi: 10.1145/3503161.3547801. URL https://arxiv.org/pdf/2206.09391.

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, Los Alamitos, CA, USA, 2025. IEEE Computer Society. doi: 10.1109/CVPR52734.2025.01853. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_Anyattack_Towards_Large-scale_Self-supervised_Adversarial_Attacks_on_Vision-language_Models_CVPR_2025_paper.html.

Jiale Zhao, Xinyang Jiang, Junyao Gao, Yuting Xue, and Cairong Zhao. One object, multiple lies: A benchmark for cross-task adversarial attack on unified vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18140–18152, Los Alamitos, CA, USA, 2025. IEEE Computer Society. doi: 10.48550/arxiv.2507.07709. URL https://arxiv.org/pdf/2507.07709.

Yunqing Zhao, Tianyu Pang, ChaoHai Du, Xiao Yang, Chongxuan Li, NgaiMan Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. In *Advances in Neural Information Processing Systems*, Red Hook, NY, USA, 2023. Curran Associates, Inc. doi: 10.48550/arxiv.2305.16934. URL https://arxiv.org/pdf/2305.16934.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv preprint arXiv:2306.05685, 2023. URL https://arxiv.org/abs/2306.05685.

Xuyang Zhong and Chen Liu. Sparse-pgd: A unified framework for sparse adversarial perturbations generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99), 2025. doi: 10.1109/TPAMI.2025.3630185. URL https://doi.org/10.1109/TPAMI.2025.3630185.

Yiqi Zhong, Xianming Liu, Deming Zhai, Junjun Jiang, and Xiangyang Ji. Shadows can be dangerous: Stealthy and effective physical-world adversarial attack by natural phenomenon. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Los Alamitos, CA, USA, 2022. IEEE Computer Society. doi: 10.1109/cvpr52688.2022.01491.

A Additional Methodological Details

A.1 Mask Construction

For the support Ω_M defined in the main text, the binary mask is

$$M_{h,w} = \mathbf{1}\big[(h, w) \in \Omega_M\big]. \tag{12}$$

We construct Ω_M by sampling the two diagonal centerline paths, rasterizing their coordinates, and applying a width-b neighborhood expansion around every sampled point. Clipping the expanded coordinates to the image boundary produces the final mask M and the ordered centerline path set $\mathcal{P}$ used by the continuity regularizer.

A.2 Complete Mask-constrained Optimization

The current adversarial image and its diversified input are

$$\begin{aligned} x^{adv,(t)} &= \mathrm{clip}(x + M \odot \delta^{(t)}, 0, 1), \\ \tilde{x}^{adv,(t)} &= D(x^{adv,(t)}). \end{aligned} \tag{13}$$

Following DI-FGSM Xie et al. (2019) and MI-FGSM Dong et al. (2018), the gradient and momentum are

$$\begin{aligned} g^{(t)} &= \nabla_\delta \mathcal{L}\big(\tilde{x}^{adv,(t)}\big), \\ m^{(t+1)} &= \mu m^{(t)} + \frac{g^{(t)}}{\mathrm{mean}(|g^{(t)}|) + 10^{-12}}. \end{aligned} \tag{14}$$

We then take a sign step and project it back to the feasible set:

$$\begin{aligned} \tilde{\delta}^{(t+1)} &= \delta^{(t)} - \alpha \ \mathrm{sign}\big(m^{(t+1)}\big), \\ \delta^{(t+1)} &= M \odot \mathrm{clip}\big(\tilde{\delta}^{(t+1)}, -\epsilon, \epsilon\big). \end{aligned} \tag{15}$$

Algorithm 1 Complete Pseudocode of XSPA

Require: Image x, mask M, surrogate f, text pool $\mathcal{T}$, iterations N, step size α, momentum μ, bound ϵ
Ensure: Best adversarial sample x^{best}

```
Initialize δ^(0) ← 0, m^(0) ← 0, and x^best ← x
Build source and target text pools from T
for t = 0 to N − 1 do
    x^adv,(t) ← clip(x + M ⊙ δ^(t), 0, 1)
    Apply input diversity to x^adv,(t)
    Encode image and text features with f
    Compute L_clip, L_tar, L_src, and L_smooth
    Form the joint objective L
    g^(t) ← ∇_δ L
    m^(t+1) ← μ m^(t) + g^(t) / (mean(|g^(t)|) + 10^−12)
    δ̃^(t+1) ← δ^(t) − α sign(m^(t+1))
    Project δ̃^(t+1) to obtain δ^(t+1)
    Update x^best
end for
return x^best
```

B Evaluation and Reproducibility Details

The evaluation universe contains 300 COCO validation images spanning the COCO-80 label space. XSPA is optimized with white-box access to the surrogate encoder, while downstream captioning and VQA models are evaluated without adapting the perturbation to those models. Classification metrics use all 300 image IDs. The strict targeted VQA plan contains 380 questions on 224 of the 300 images; we state the smaller denominator whenever this plan is used.

The reviewer-oriented comparison freezes OpenCLIP ViT-B/16, the image order, optimization loss, and all attack hyperparameters. The only change between XSPA and the saliency Top-k baseline is support selection: XSPA uses its fixed geometric support, whereas Top-k selects the 1,532 clean-image pixels with the largest attack-loss gradient magnitude. Unless a quantity is itself varied, the configuration in Table 4 is held fixed. This design prevents attack strength from being attributed to unequal pixel budgets or optimization effort.

Table 4: Default configuration for the frozen ablation protocol.

Setting	**Default value**
Evaluation set	COCO-300
Surrogate encoder	OpenCLIP ViT-B/16
Input resolution	384×384
Perturbation support	1,532 pixels (approximately 1.04%)
Optimization steps	100
Step size	0.03
L_∞ bound	0.25

For zero-shot classification, primary attack success rate (ASR) is targeted success when a valid predefined target exists and untargeted success otherwise. All 300 images in the matched-budget study have valid targets, so primary and targeted ASR coincide. Untargeted ASR is reported separately. Statistical uncertainty is computed by resampling image IDs while retaining all optimization seeds within an image cluster; the 900 seed-image observations are not treated as independent samples.

C Matched-budget Strong Baseline and Seed Stability

Table 5 reports three independent optimization seeds under the frozen protocol. Saliency Top-k achieves a primary ASR of $94.44 \pm 0.38\%$, compared with $83.44 \pm 0.77\%$ for XSPA. The paired Top-k-minus-XSPA difference is 11.00 percentage points, with a 20,000-replicate image-cluster bootstrap 95% confidence interval of $[7.78, 14.44]$ points. For untargeted ASR, the corresponding difference is 8.00 points with a 95% confidence interval of $[5.00, 11.11]$. Exact paired McNemar tests are significant for every seed after Holm correction; the largest adjusted p-values are 8.70×10^{-6} for primary ASR and 7.17×10^{-4} for untargeted ASR.

Table 5: Matched-budget ASR for saliency Top-k and XSPA on the same 300 image IDs. Every run uses a 1,532-pixel support, $L_\infty = 0.25$, 100 steps, and step size 0.03. Values are percentages.

Method	**ASR**	**3407**	**3408**	**3409**	**Mean $\pm$ SD**
Top-k	Targeted	94.67	94.00	94.67	94.44 ± 0.38
XSPA	Targeted	83.00	83.00	84.33	83.44 ± 0.77
Top-k	Untargeted	95.00	94.67	94.67	94.78 ± 0.19
XSPA	Untargeted	86.67	86.67	87.00	86.78 ± 0.19

These results establish a narrow conclusion. The X-shaped support is effective under a severe sparsity constraint, but it is not the strongest support selector at the same pixel budget. Accordingly, the geometric contribution should be interpreted as a controlled structured-support stress test rather than an attack-strength advantage over adaptive saliency selection.

D Perceptual Distortion and Runtime

We compute full-resolution PSNR, SSIM, AlexNet-LPIPS, mean absolute error (MAE), changed-pixel ratio, and realized L_∞ on all 300 paired image IDs for all three seeds. Table 6 reports method means obtained by first aggregating within each seed; uncertainty for the XSPA-minus-Top-k difference is again obtained by image-cluster bootstrap. Runtime uses seeds 3408 and 3409 only because

the frozen XSPA seed-3407 artifact predates per-image timing. The visual metrics retain all three seeds.

Table 6: Perceptual distortion and runtime under the matched-budget protocol. Differences are XSPA minus saliency Top-k. Higher is better for PSNR and SSIM; lower is better for all other rows.

Metric	**Top-k mean $\pm$ SD**	**XSPA mean $\pm$ SD**	**Difference**	**95% cluster CI**
PSNR (dB)	34.5981 ± 0.0093	34.1714 ± 0.0113	-0.4267	$[-0.4918, -0.3641]$
SSIM	0.96060 ± 0.00007	0.97285 ± 0.00003	$+0.01225$	$[+0.01132, +0.01321]$
LPIPS-Alex	0.08575 ± 0.00016	0.06516 ± 0.00008	-0.02059	$[-0.02359, -0.01767]$
MAE	0.001680 ± 0.000003	0.001798 ± 0.000003	$+0.000118$	$[+0.000100, +0.000137]$
Changed pixels (%)	1.0353 ± 0.0001	1.0383 ± 0.0000	$+0.0030$	$[+0.0017, +0.0046]$
Realized L_∞	0.25098 ± 0.00000	0.25098 ± 0.00000	0.00000	$[0.00000, 0.00000]$
Attack time (s/image)	12.455 ± 0.191	12.354 ± 0.174	-0.102	$[-0.116, -0.087]$

The perceptual evidence is mixed rather than uniformly favorable. XSPA improves SSIM and LPIPS, indicating better structural and learned-feature similarity, whereas Top-k has higher PSNR and lower MAE. Both methods realize the same L_∞ bound and nearly the same support fraction, and their runtime difference is small relative to the approximately 12-second per-image cost. We therefore use "sparse" and "low-coverage" as directly supported descriptors and do not treat attack success as evidence of imperceptibility.

E Same-image Cross-task Joint Success

We additionally test whether classification, captioning, and VQA fail toward their predefined targets on the same adversarial image. This analysis uses seed 3407. Caption outputs are generated by BLIP-2 FlanT5-XL and VQA outputs by LLaVA-1.6 Mistral-7B. An image is jointly eligible when its clean caption does not already mention the target category and at least one planned VQA target answer is absent on the clean input. Of the 300 images, 224 have at least one strictly filtered targeted VQA question and 214 satisfy the joint clean-input eligibility rule.

Table 7: Factorization of same-image targeted success on the 214 jointly eligible images. "VQA any" requires at least one eligible target answer; "VQA all" requires every eligible target answer for that image.

Component	**XSPA**	**Saliency Top-k**
CLIP targeted	177/214 (82.71%)	203/214 (94.86%)
Caption targeted	0/214 (0.00%)	0/214 (0.00%)
VQA any	9/214 (4.21%)	8/214 (3.74%)
VQA all	3/214 (1.40%)	3/214 (1.40%)
Joint with VQA any	0/214 (0.00%)	0/214 (0.00%)
Joint with VQA all	0/214 (0.00%)	0/214 (0.00%)

Neither method achieves a three-task targeted success on this common denominator because the downstream caption model never introduces the predefined target category. The result does not support a claim of correlated targeted failure across all three downstream tasks. It instead separates strong surrogate classification effects from substantially weaker output-space transfer.

F Judge Calibration and Claim Boundary

The GPT-based evaluation is calibrated on an overlapping LLaVA-1.5 subset containing 30 clean and 30 XSPA outputs per task. Each item is independently scored by GPT-4 and GPT-4o-mini under the same task rubric. The bootstrap resamples item IDs and retains the clean/XSPA condition pair. This is an inter-GPT calibration study; it is not human-GPT agreement.

Table 8: Inter-GPT judge calibration on 60 condition-items per task.

Task	Reliability measure	Estimate [95% CI]
Caption	Pearson r	0.580 $[0.255, 0.801]$
Caption	Spearman ρ	0.579 $[0.268, 0.796]$
Caption	ICC(A,1)	0.301 $[0.120, 0.443]$
Caption	MAE (0–30 scale)	3.683 $[3.050, 4.367]$
VQA	Exact agreement	0.933 $[0.850, 1.000]$
VQA	Cohen's κ	0.815 $[0.526, 1.000]$

VQA scoring shows high observed agreement, whereas caption scoring has only moderate rank correlation and low absolute-agreement ICC. GPT-4 assigns caption scores that are, on average, 3.283 points higher on the 0–30 scale than GPT-4o-mini. Because independent human ratings are not yet available, these data do not constitute human validation. We consequently treat judge-based caption differences as secondary evidence and avoid using them alone to support cross-task transfer claims.

G Statistical and Artifact Audit

All matched-budget summaries are reconstructed from per-image JSON records rather than terminal logs. The audit checks 300 unique IDs in every method-seed cell, verifies a 1,532-pixel mask on every image, and requires the method-specific mask variant to agree with the recorded support selector. Paired confidence intervals use 20,000 image-cluster bootstrap replicates with seed 20260731. Binary method comparisons use exact McNemar tests within each optimization seed and Holm correction across the three seeds for a metric. These choices preserve the paired image structure and prevent optimization replicates from inflating the effective sample size.

Overall, the evidence supports sparse surrogate attacks and a perceptual tradeoff, but not universal superiority, uniform imperceptibility, or reliable three-task targeted transfer. Human perceptual and human-GPT studies are incomplete and are not reported as findings.