Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

FASDP: Frequency-Aware Scale-Diverse Purification against Pixel-Space Visual Jailbreaks

Xiaolong Qin
Chengdu University of Information Technology
Chunjiang Wu
westbuddha@cuit.edu.cn

Chengdu University of Information Technology
YiLe Li
Chengdu University of Information Technology
TingHui Chen
Chengdu University of Information Technology
Zhibin Li
Chengdu University of Information Technology
ShiLin Qiu
Chengdu University of Technology

Research Article

Keywords: Vision-language models, visual jailbreak attacks, input purification, frequency analysis, training-free defense

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11336583/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

FASDP: Frequency-Aware Scale-Diverse Purification against Pixel-Space Visual Jailbreaks

Abstract

Vision-language models (VLMs) have achieved remarkable success across a wide range of applications. However, as their deployment becomes increasingly widespread, recent studies have revealed that VLMs are vulnerable to visual jailbreak attacks, where adversarially manipulated images induce safety-aligned models to produce harmful responses, posing serious threats to real-world security. Input purification defenses must balance perturbation suppression against content preservation; a fixed transformation strength may be unsuitable for inputs and attacks with different characteristics. We propose FASDP, a Frequency-Aware Scale-Diverse Purification framework that adaptively allocates purification strength per image. It fuses multi-scale candidates from a single frozen denoiser using low-frequency consistency modulated by a self-referenced frequency cue. The framework requires no task-specific training or access to the target model. We evaluate FASDP on LLaVA-1.5-7B, InstructBLIP-7B, and Qwen2.5-VL-7B under three representative pixel-space attacks. FASDP attains strong protection under these attacks while keeping clean-image performance on downstream tasks close to that of the undefended models.

Keywords: Vision-language models, visual jailbreak attacks, input purification, frequency analysis, training-free defense

1 Introduction

VLMs couple visual encoders with instruction-tuned large language models to support image captioning, visual question answering (VQA), and open-ended multimodal dialogue. Representative systems such as BLIP-2 [1], InstructBLIP [2], LLaVA [3], and the Qwen-VL family [4] are increasingly deployed in real-world applications, where failures of safety alignment become increasingly consequential.

Visual jailbreak attacks exploit the image channel to elicit responses that an aligned model would otherwise refuse. Pixel-optimized adversarial images steer a model toward harmful outputs [5], whereas compositional and typographic attacks encode harmful intent through visible image content [6, 7]. This paper targets the first family, whose perturbations may be attenuated by image transformations, and treats attacks driven by visible semantic content as outside the scope of the proposed purifier.

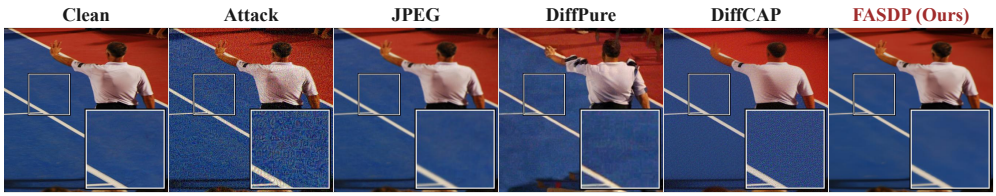


Fig. 1: Purified outputs of representative defenses on a pixel-space jailbreak input.

Classical input transformations, such as random resizing, padding, and feature squeezing [8–10], and diffusion-based purification [11] defend adversarial images without modifying the target model. Recent VLM-specific purifiers, such as F3 and DiffCAP [12, 13], have shown promising results under visual adversarial attacks. Visual jailbreaks differ from conventional adversarial examples: they perturb a larger fraction of pixels at higher ℓ_∞ magnitudes. Mild transformations therefore leave much of the perturbation, whereas stronger purification introduces artifacts or loses fine detail. This trade-off also affects clean images, where preserving the undefended model's performance is itself a requirement. Adaptive approaches reduce it but often rely on model-derived signals, and effectiveness on captioning, VQA, or classification does not by itself establish protection against visual jailbreaks. These observations reduce to one question for an image-only purifier: how to suppress high-budget jailbreak perturbations while minimizing the damage to clean-image performance without target-model signals.

We propose FASDP for pixel-space visual jailbreak defense. Its central design is an adaptive aggregation policy over denoised candidates rather than a new denoiser. For each input, FASDP measures the normalized change in the high-frequency energy ratio under a mild image transform. This self-referenced statistic is a transformation-sensitivity cue; it does not directly measure perturbation magnitude or distinguish adversarial perturbations from natural texture. This cue then modulates the fusion weights of the scale-diverse denoised candidates. Figure 1 compares purified outputs of representative defenses under a pixel-space jailbreak input. The reference configuration requires neither additional training nor access to target-model parameters, gradients, or embeddings.

Our contributions are threefold.

- **Training-free per-image strength allocation.** FASDP adapts purification strength per image by reweighting scale-diverse candidates from a single frozen denoiser, requiring neither task-specific training nor target-model access.
- **A self-referenced frequency statistic.** The statistic Δ measures the normalized change in the DCT high-frequency energy ratio under a mild image transform, serving as a target-model-free cue that calibrates the denoising strength η and modulates candidate fusion.
- **A low-frequency-consistency fusion rule.** Input- and pairwise-consistency distances of low-frequency components yield closed-form softmax weights that aggregate the candidates under this cue.

2 Related Work

2.1 Vision-Language Models

VLMs have emerged as a fundamental paradigm bridging visual and textual understanding. Early representative models, such as CLIP [14], align images and text through contrastive pretraining, and subsequent systems, including Flamingo [15] and BLIP-2 [1], couple frozen visual encoders with large language models through lightweight connectors. More recent instruction-tuned and open models, such as LLaVA [3], MiniGPT-4 [16], InstructBLIP [2], Qwen-VL [4], and InternVL [17], together with proprietary systems, such as GPT-4V [18] and Gemini [19], are now widely used in real-world applications. With their rapidly expanding use, the safety issues of these models have become increasingly severe.

2.2 Security and Safety Risks of VLMs

Recent studies have exposed these safety issues across the entire model pipeline. Backdoor attacks, such as BadNets [20] and CBV [21], implant hidden triggers through poisoned training corpora, and such poisoning is practical at web scale [22]. At inference time, jailbreak prompts based on competing objectives [23] and learned adversarial suffixes [24] circumvent safety alignment. Visual attacks, such as HADES [25] and FigStep [7], steer aligned models toward attacker-chosen behavior through optimized perturbations or visible harmful content, and remain effective even under black-box access [26]. Systematic benchmarks, such as MM-SafetyBench [27] and OmniSafeBench [28], confirm that current VLMs remain vulnerable to these attacks. Among them, visual jailbreak attacks are particularly practical, as they require neither access to training data nor modification of model parameters, and thus directly threaten deployed systems; they are the focus of this paper.

2.3 Visual Jailbreak Attacks

Visual jailbreak attacks craft images that induce aligned VLMs to produce harmful outputs. One family of such attacks optimizes pixel perturbations, extending the classical adversarial-example literature [29–32]: Qi et al. [5] show that a single universal image can jailbreak aligned VLMs, Image Hijacks [33] control model outputs at runtime, and HADES [25] optimizes adversarial perturbations to defeat both open-source and proprietary models, while transfer- or score-based variants remain effective under black-box access [26]. Other attacks embed harmful intent in visible content rather than perturbations: FigStep [7] typesets prohibited instructions as in-image text, and Jailbreak in Pieces [6] spreads a harmful concept across the visual and textual components of a query. These failures have been systematized by recent benchmarks and surveys [27, 28, 34], and effective defenses remain an urgent open problem.

2.4 Defense Methods

Existing defenses intervene at different positions of the inference pipeline. Model-involved methods operate within or around the target VLM: UniGuard learns

multimodal input guardrails [35], BaThe inserts a virtual rejection prompt [36], Tit-for-Tat combines visual and textual safety instructions [37], SafePTR prunes harmful multimodal tokens [38], and E^2AT performs multimodal adversarial training [39]. Input-level methods act before inference; ECSO, for example, replaces the image with its textual description [40].

Input purifiers instead retain the image and seek to suppress adversarial perturbations through classical transforms [8–10], learned compression [41], generative projection [11, 42, 43], or randomized smoothing [44–46]. A motivation for these approaches is that adversarial perturbations can exploit non-robust features, including high-frequency components [47–49], and may be attenuated by scale transformations or denoising priors [50–54]. For VLMs, F3 repairs adversarial images via target-model cross-modal attention with added noise [12], and DiffCAP selects a per-input diffusion duration by embedding convergence [13]. Strong or unconstrained jailbreak perturbations [5], however, may be difficult to suppress with mild transforms such as resizing alone. Increasing the noise level in diffusion purification can strengthen perturbation suppression at the cost of image fidelity [11, 13]. Content-faithful purification under strong jailbreak perturbations therefore remains challenging.

3 Threat Model

Adversary goal. The adversary aims to induce an aligned VLM to generate harmful responses to image-containing queries, circumventing the safety alignment of the target model.

Adversary capability and knowledge. The adversary perturbs only the input image in pixel space, possibly in composition with harmful imagery [5, 25, 33]. We consider both black-box transfer [26] and white-box adaptive attackers, including gradient-based attacks against non-differentiable or randomized purification pipelines [55–57]. Attacks encoding harmful intent in visible content or in the text prompt are out of scope.

Defense assumptions and scope. The purifier observes only the input image and requires neither task-specific training nor access to target-model parameters, gradients, or embeddings, leaving the target model and its safety alignment unchanged. It targets perturbation-based pixel-space attacks, whereas attacks conveying harmful semantics through visible content are out of scope.

4 Method

4.1 Overview

FASDP is an input-purification framework for pixel-space visual jailbreak defense. An image is an element of $[0,1]^{C\times H\times W}$ with $C = 3$ color channels; I_{clean}, I_{adv}, and I_{pur} denote the clean, attacked, and purified images, with $\delta = I_{\text{adv}} - I_{\text{clean}}$, and the defender observes only the attacked image. As shown in Figure 2, a frequency analyzer Φ_θ measures the normalized change Δ in the input's high-frequency energy ratio under a fixed reference transformation and calibrates denoising strength η using multiple reference scales. A fixed set of scale transformations $\{\mathcal{T}_s\}_{s\in\mathcal{S}}$, followed by one shared

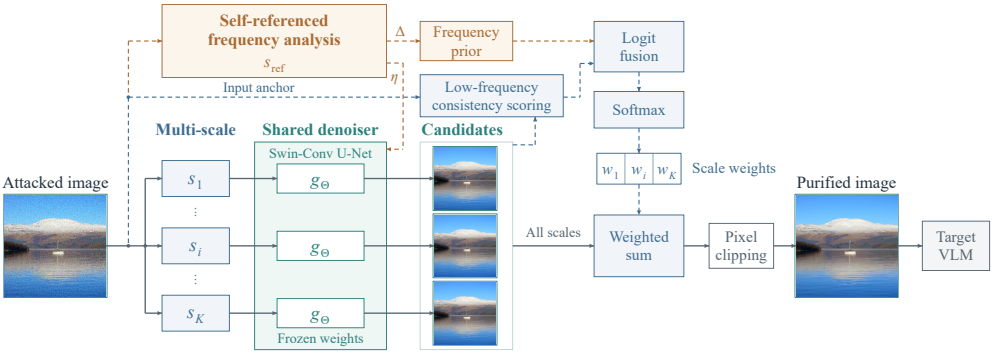


Fig. 2: The FASDP framework: self-referenced frequency analysis calibrates denoising strength and modulates the fusion of scale-diverse denoised candidates, with no training or target-model signal.

denoiser $\mathcal{D}$ with strength η, produces the candidates. A fusion rule Ψ combines these candidates using low-frequency consistency weights modulated by Δ, before $\mathcal{R}$ clips the result to the valid range. Algorithm 1 summarizes the procedure, which requires neither task-specific training nor target-model signals.

4.2 Self-Referenced Frequency Statistic and Scale Transformations

Fixed-strength purification uses the same transformation setting for images with substantially different levels of high-frequency activity. FASDP therefore allocates purification strength per image with self-referenced frequency analysis, without any target-model signal; multiple reference scales calibrate denoising strength, while the statistic Δ enters the fusion of Section 4.4 as a modulating prior, not as a binary attack detector.

The frequency statistic is based on the channel-averaged high-frequency energy ratio ρ_{HF} under the two-dimensional DCT-II transform. For one channel $X^c \in \mathbb{R}^{H\times W}$ of an image X, the coefficient at frequency (u, v) is

$$\mathcal{C}(X^c)[u,v] = \alpha_u \alpha_v \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} X^c[x,y] \cos\frac{\pi u(2x+1)}{2H} \cos\frac{\pi v(2y+1)}{2W}, \tag{1}$$

with normalization factors $\alpha_0 = \sqrt{1/H}$, $\alpha_u = \sqrt{2/H}$ for $u > 0$, and α_v defined analogously. With the normalized radial coordinate $r(u, v)$ and the high-frequency mask

$$r(u,v) = \frac{1}{\sqrt{2}} \sqrt{\left(\frac{u}{H-1}\right)^2 + \left(\frac{v}{W-1}\right)^2}, \qquad \Omega_{\text{HF}}(\gamma) = \{(u,v) : r(u,v) > 1-\gamma\}, \tag{2}$$

Algorithm 1 FASDP: training-free purification and consistency-weighted fusion

Require: Input image I_{adv}; scale set $\mathcal{S}$; reference scales $\mathcal{S}_{\text{ref}} \subset (0,1)$ with $s_{\text{ref}} \in \mathcal{S}_{\text{ref}}$; frequency parameters θ; shared denoiser $\mathcal{D}$; fusion parameters $\lambda_{\text{in}}, \beta, T, \kappa$

Ensure: Purified image I_{pur}

for $s_i \in \mathcal{S}$ **do**
  $X_i \leftarrow \mathcal{T}_{s_i}(I_{\text{adv}})$
end for
$\rho_0 \leftarrow \rho_{\text{HF}}(I_{\text{adv}})$
for $s \in \mathcal{S}_{\text{ref}}$ **do**
  Obtain $\mathcal{T}_s(I_{\text{adv}})$, reusing X_i when $s = s_i$
  $\rho_s \leftarrow \rho_{\text{HF}}(\mathcal{T}_s(I_{\text{adv}}))$
end for
$\rho_{\text{ref}} \leftarrow \rho_{s_{\text{ref}}}$; $\Delta \leftarrow (\rho_0 - \rho_{\text{ref}})/(\rho_0 + \varepsilon)$
$\eta \leftarrow \text{clip}\Big(\frac{1}{|\mathcal{S}_{\text{ref}}|} \sum_{s \in \mathcal{S}_{\text{ref}}} \frac{\rho_0 - \rho_s}{\rho_0 + \varepsilon}, 0, 1\Big)$
for $s_i \in \mathcal{S}$ **do**
  $\widehat{I}_{s_i} \leftarrow \mathcal{D}(X_i; \eta)$ using Eq. (6)
end for
$L_0 \leftarrow \text{LPF}(I_{\text{adv}})$; $L_i \leftarrow \text{LPF}(\widehat{I}_{s_i})$ for all i
$d_i \leftarrow \lambda_{\text{in}} d_i^{\text{in}} + (1 - \lambda_{\text{in}}) d_i^{\text{pair}}$ using Eqs. (10)–(12)
$q_i \leftarrow (s_i - s_{\min})/(s_{\max} - s_{\min} + \varepsilon)$
$c_i \leftarrow 1/(d_i + \varepsilon)$; $c_i' \leftarrow c_i(1 + \beta q_i)$
$w_i \leftarrow \text{softmax}_i(\log(c_i')/T - \kappa \Delta q_i)$ **return** $I_{\text{pur}} = \text{clip}\left(\sum_i w_i \widehat{I}_{s_i}, 0, 1\right)$

where $\gamma \in (0,1)$ controls the radial cutoff rather than the fraction of selected coefficients, the ratio is

$$\rho_{\text{HF}}(X) = \frac{1}{C} \sum_{c=1}^{C} \frac{\sum_{(u,v) \in \Omega_{\text{HF}}} |\mathcal{C}(X^c)[u,v]|^2}{\sum_{u,v} |\mathcal{C}(X^c)[u,v]|^2 + \varepsilon}, \tag{3}$$

with $\varepsilon > 0$ preventing division by zero. The analyzer parameters θ comprise the DCT configuration, the cutoff parameter γ, and the reference scales s_{ref} and $\mathcal{S}_{\text{ref}}$ defined below. Since natural images concentrate spectral energy at low frequencies, ρ_{HF} is an image-level proxy whose absolute value depends on image content and the mask.

Each scale $s \in (0,1]$ in the fixed set $\mathcal{S} = \{s_1 > s_2 > \cdots > s_K\}$ downsamples the input and restores the target size,

$$\mathcal{T}_s(X) = U_{\text{bilinear}}\left(D_{\text{area}}(X, sH, sW), H, W\right), \tag{4}$$

where area interpolation averages local content during downsampling and bilinear interpolation restores the target size. The reference configuration uses $\mathcal{S} = \{1.0, 0.75, 0.5\}$ to span mild to more aggressive transformations. In Eq. (4), $\mathcal{T}_1$ is the identity; this candidate retains the original spatial resolution before denoising. The resize-only operator is used without an additional Gaussian prefilter.

The absolute value of ρ_{HF} depends on content: natural texture, edges, and embedded text can also increase it. FASDP therefore compares the input with its own transformed version, using a reference scale $s_{\mathrm{ref}} \in (0, 1)$ defined separately from $\mathcal{S}$. The reference scale is fixed at $s_{\mathrm{ref}} = 0.75$, with the candidate scales retained as $\mathcal{S} = \{1.0, 0.75, 0.5\}$. The self-referenced frequency statistic is

$$\Delta(X) = \frac{\rho_{\mathrm{HF}}(X) - \rho_{\mathrm{HF}}(\mathcal{T}_{s_{\mathrm{ref}}}(X))}{\rho_{\mathrm{HF}}(X) + \varepsilon}. \tag{5}$$

It measures the normalized change in the high-frequency energy ratio, rather than the fraction of absolute high-frequency energy removed. A positive value indicates that the reference transformation reduces this ratio; it serves as a heuristic cue for modulating purification strength. Natural fine texture can also be attenuated, and low-frequency perturbations may persist, so Δ does not identify adversarial content or estimate its true magnitude. The statistic is bounded above by 1 but may be negative when the transformed ratio increases. Negative values bias the prior toward milder candidates; this describes the weighting direction and does not imply a safety guarantee. To calibrate denoising strength from multiple scales, we additionally fix $\mathcal{S}_{\mathrm{ref}} = \{0.75, 0.5\}$ and compute

$$\eta(X) = \mathrm{clip}\left(\frac{1}{|\mathcal{S}_{\mathrm{ref}}|} \sum_{s \in \mathcal{S}_{\mathrm{ref}}} \frac{\rho_{\mathrm{HF}}(X) - \rho_{\mathrm{HF}}(\mathcal{T}_s(X))}{\rho_{\mathrm{HF}}(X) + \varepsilon},\, 0,\, 1 \right).$$

The analyzer outputs $(\Delta, \eta) = \Phi_\theta(I_{\mathrm{adv}})$, with $\eta = \eta(I_{\mathrm{adv}})$ shared across all candidate scales. This fixed mapping bounds the residual denoising strength in $[0, 1]$ without introducing an additional tunable coefficient; it is a heuristic control signal, not an estimate of the true noise level. When reference scales occur in $\mathcal{S}$, their images are reused from the corresponding candidate inputs. The reference scales remain fixed when the candidate set changes; if absent from that set, their images are generated separately without extra denoiser passes.

4.3 Shared Denoiser and Candidate Construction

Each scale $s \in \mathcal{S}$ yields one candidate through a shared denoiser $\mathcal{D}(\cdot; \eta) : [0, 1]^{C \times H \times W} \to [0, 1]^{C \times H \times W}$,

$$\begin{aligned} \widehat{I}_s &= \mathcal{D}(\mathcal{T}_s(I_{\mathrm{adv}}); \eta), \\ \mathcal{D}(Y; \eta) &= (1 - \eta)Y + \eta g_\Theta(Y). \end{aligned} \tag{6}$$

where scale diversity enters through $\mathcal{T}_s$, not through different denoisers or strength settings across scales. The coefficient η controls the fraction of the residual $g_\Theta(Y) - Y$ applied to each input; $\eta = 0$ returns the scale-transformed image, and $\eta = 1$ recovers the full blind-denoiser output.

The reference setting uses g_Θ, a publicly released blind denoiser of the U-shaped window-attention–convolution family [54], inside $\mathcal{D}$. A head convolution maps the

input Y to F feature channels, and three encoder stages apply a stride-2 downsampling convolution followed by hybrid blocks,

$$Z^0 = \mathrm{Conv}_{3\times3}^{C\to F}(Y), \qquad Z^\ell = \mathrm{SC}_\ell\big(\mathrm{Down}_\ell(Z^{\ell-1})\big), \quad \ell = 1,2,3, \tag{7}$$

where Down_ℓ doubles the channels, $F_\ell = 2^\ell F$. A bottleneck block yields Z^4; the decoder mirrors the encoder with transposed convolutions and additive skip connections, for $\ell = 2,3,4$ with $\widehat{Z}^4 = Z^4$,

$$\widehat{Z}^{\ell-1} = \mathrm{SC}'_{\ell-1}\Big(\mathrm{Up}_\ell(\widehat{Z}^\ell)\Big) + Z^{\ell-2}, \qquad g_\Theta(Y) = \mathrm{Conv}_{3\times3}^{F\to C}\Big(\widehat{Z}^1\Big). \tag{8}$$

Each hybrid block splits its input along channels, $[Z_a; Z_b] = Z$, and couples window self-attention with local convolution,

$$\mathrm{SC}(Z) = Z + \mathrm{Conv}_{1\times1}[\mathrm{STL}(Z_a);\ \phi(\mathrm{Conv}_{3\times3}(Z_b))], \tag{9}$$

where ϕ denotes ReLU and $\mathrm{STL}(Z) = \mathrm{MLP}(\mathrm{LN}(\widetilde{Z})) + \widetilde{Z}$ with $\widetilde{Z} = \text{W-MSA}(\mathrm{LN}(Z)) + Z$; successive blocks alternate regular and shifted windows.

The weights Θ are public and frozen, so purification involves no task-specific training. As a blind denoiser, g_Θ takes no noise-level input; η controls the residual interpolation outside this network, with one Θ shared across scales. Alternative denoisers implementing the same image-to-image interface replace g_Θ only in the denoiser ablation, with the strength mapping, scale set and fusion rule fixed.

4.4 Low-Frequency-Consistency Fusion

The candidates provide different trade-offs: larger scales retain more detail, whereas smaller scales apply stronger smoothing before denoising. Uniform averaging assigns equal importance to all candidates, while single-candidate selection discards information from the remaining scales. FASDP instead aggregates candidates using low-frequency consistency scores modulated by the statistic in Eq. (5). The consistency scores are already input-dependent; the frequency statistic provides an additional cue for adjusting their relative weights.

Let $L_0 = \mathrm{LPF}(I_{\mathrm{adv}})$ and $L_i = \mathrm{LPF}(\widehat{I}_{s_i})$ denote the low-frequency representations of the input and of each candidate of Eq. (6) under a fixed Gaussian operator. The input-consistency distance

$$d_i^{\mathrm{in}} = \frac{1}{CHW}\|L_i - L_0\|_2^2, \tag{10}$$

measures how much a candidate alters the low-frequency content of the received image, and the pairwise-consistency distance

$$d_i^{\mathrm{pair}} = \begin{cases} \dfrac{1}{(K-1)CHW}\sum_{j\neq i}\|L_i - L_j\|_2^2, & K > 1, \\ 0, & K = 1. \end{cases} \tag{11}$$

measures its disagreement with the other scales. Their combination

$$d_i = \lambda_{\text{in}} d_i^{\text{in}} + (1 - \lambda_{\text{in}}) d_i^{\text{pair}}, \tag{12}$$

with $\lambda_{\text{in}} \in [0, 1]$, anchors each candidate to the received image. For $K = 2$, a positive input-anchor weight can distinguish candidates when their input distances differ, whereas their pairwise distances are equal.

Smaller distances receive larger raw scores,

$$c_i = \frac{1}{d_i + \varepsilon}. \tag{13}$$

A fidelity bias $c_i' = c_i(1 + \beta q_i)$ favors larger scales, where $q_i = (s_i - s_{\min})/(s_{\max} - s_{\min} + \varepsilon)$ and $\beta \geq 0$. A statistic-conditioned prior modulates these scores:

$$p_i = \frac{\exp(-\kappa\, \Delta\, q_i)}{\sum_{j=1}^{K} \exp(-\kappa\, \Delta\, q_j)}, \qquad \kappa \geq 0. \tag{14}$$

As $\Delta \to 0$, the prior approaches a uniform distribution; for positive Δ, it downweights milder candidates with larger q_i. Setting $\kappa = 0$ removes the fusion frequency prior while retaining the consistency scores, fidelity bias and denoising-strength calibration. For positive Δ, the fidelity bias and frequency prior favor opposite scale directions; their balance depends on β, κ, T, and the consistency distances. The weights combine temperature-normalized log scores with the prior, with $T > 0$. Because the prior's normalizing constant is shared by all candidates, it cancels inside the softmax:

$$w_i = \text{softmax}_i \left(\frac{\log(c_i')}{T} - \kappa \Delta q_i \right), \qquad \sum_i w_i = 1, \tag{15}$$

This equivalent form avoids explicitly exponentiating and then taking the logarithm of the prior. The fused image is

$$I_{\text{fused}} = \sum_{i=1}^{K} w_i \widehat{I}_{s_i}. \tag{16}$$

Per-image strength allocation is therefore realized through residual denoising strength and continuous fusion weights rather than discrete scale selection. Equivalently, $w_i \propto [c_i(1+\beta q_i)]^{1/T} \exp(-\kappa \Delta q_i)$. Temperature controls the concentration of the consistency contribution, but neither the logarithm nor finite temperature guarantees that a single candidate cannot dominate.

4.5 Reconstruction and Complexity

Since every candidate is already restored to the target size before fusion, reconstruction consists only of clipping, $\mathcal{R}(I_{\text{fused}}) = \text{clip}(I_{\text{fused}}, 0, 1)$.

Let $K = |\mathcal{S}|$ be the fixed number of scales. The frequency analysis costs $1 + |\mathcal{S}_{\text{ref}}|$ DCT evaluations, $O(HW \log(HW))$ per channel each, and the scaling, residual interpolation and low-pass filtering are linear in the pixel count; with $C_{\mathcal{D}}(H, W)$ denoting one denoiser pass at the target resolution and the reference set fixed, the total cost is

$$O\left(CHW \log(HW) + K^2 CHW + K C_{\mathcal{D}}(H, W)\right). \tag{17}$$

The K^2CHW term accounts for directly evaluating the pairwise distances in Eq. (11). For the denoiser of Eq. (8), $C_{\mathcal{D}}(H, W) = O\Big(HW \sum_{\ell=0}^{3} F_\ell^2 4^{-\ell}\Big)$, as convolutions thin out by $4^{-\ell}$ per scale and window attention is linear in the pixel count for a fixed window size. With $K = 3$ and $|\mathcal{S}_{\text{ref}}| = 2$, FASDP uses three shared-denoiser forward passes and three DCT evaluations per input, without iterative diffusion sampling. This operation count does not establish a measured latency advantage; runtime depends on resolution, hardware, and implementation.

5 Evaluation

5.1 Experimental Setup

Datasets. Jailbreak evaluation uses WildJailbreak [58], and clean-task utility is measured on COCO and VQAv2.

Target Models. We evaluate three classic open-source VLMs: LLaVA-1.5-7B [3], InstructBLIP-7B [2], and Qwen2.5-VL-7B [59], spanning different vision encoders, connector designs, and model families.

Attack Methods. We construct the jailbreak attack set with three representative pixel-space attacks: visual adversarial examples (VAE) [5], HADES [25], and Image Hijacks [33]. Unless stated otherwise, perturbation-based attacks use an ℓ_∞ budget of 32/255, following the strong-budget setting of visual jailbreak attacks [5].

Baseline Defenses. We compare FASDP with five inference-time defenses: JPEG compression [8], random resizing and padding (R&P) [9], DiffPure [11], F3 [12], and DiffCAP [13], with No Defense as the unprotected reference.

FASDP Configuration. The frozen reference configuration uses the candidate scale set $\mathcal{S} = \{1.0, 0.75, 0.5\}$, prior strength $\kappa = 10$, consistency mixing coefficient $\lambda_{\text{in}} = 0.5$, frequency cutoff $\gamma = 0.25$, and fidelity bias $\beta = 0.05$, fixed once and reused across all target models and attacks without per-setting retuning.

Evaluation Metrics. We measure safety by the attack success rate (ASR), the fraction of attacked inputs that elicit a policy-violating response as judged by an online GPT-5.4 API. Clean-task utility is measured by VQA accuracy on VQAv2 and CIDEr on COCO captioning; purified-image fidelity is measured by SSIM, LPIPS, and CLIP similarity.

5.2 Main Experiments

The main experiments ask three questions: how much protection FASDP provides against pixel-space jailbreak attacks, how well the purified images preserve visual quality, and what this protection costs on benign inputs.

Table 1: ASR and post-purification model performance of five inference-time defenses and FASDP under three pixel-space jailbreak attacks on three classic open-source VLMs.

Attack	Defense	LLaVA-1.5-7B			InstructBLIP-7B			Qwen2.5-VL-7B		
		ASR↓	VQA↑	CIDEr↑	ASR↓	VQA↑	CIDEr↑	ASR↓	VQA↑	CIDEr↑
No Attack	–	34.78	67.13	122.48	27.12	72.03	136.53	18.67	81.60	144.31
VAE [5]	No Defense	84.57	58.96	99.29	76.34	59.45	110.57	79.41	63.54	113.88
	JPEG [8]	38.42	59.26	111.93	36.50	60.13	119.59	25.03	75.90	131.07
	R&P [9]	41.15	57.36	112.12	35.49	62.73	119.71	28.25	76.07	130.19
	DiffPure [11]	41.91	52.35	77.12	33.71	53.02	72.24	23.09	55.81	85.98
	F3 [12]	35.65	58.15	**119.93**	37.14	67.92	131.41	22.77	72.39	136.40
	DiffCAP [13]	37.79	59.01	107.75	37.51	**69.90**	115.83	21.69	72.21	122.53
	FASDP (Ours)	**35.15**	**64.29**	116.19	**28.20**	67.17	**133.70**	**21.26**	**77.93**	**143.79**
HADES [25]	No Defense	86.97	56.94	104.65	76.85	59.50	111.89	83.63	61.63	112.62
	JPEG [8]	41.53	59.87	112.78	36.77	62.80	121.05	27.46	75.30	138.70
	R&P [9]	38.37	56.18	111.57	35.05	61.04	122.63	29.72	76.58	131.50
	DiffPure [11]	42.94	48.23	78.47	34.97	50.02	71.78	22.07	56.01	90.14
	F3 [12]	38.63	60.15	115.04	39.06	**68.20**	131.83	21.63	71.21	138.22
	DiffCAP [13]	36.99	62.80	107.25	36.27	65.66	115.11	**20.54**	71.87	122.21
	FASDP (Ours)	**35.04**	**64.59**	**116.43**	**28.46**	67.44	**134.94**	21.22	**76.81**	**142.88**
Image Hijacks [33]	No Defense	89.26	57.07	101.23	81.83	61.65	108.42	84.25	64.24	119.38
	JPEG [8]	41.57	59.51	112.16	37.23	61.07	122.86	28.00	75.37	135.55
	R&P [9]	38.19	55.57	115.27	35.91	62.72	121.11	25.81	75.25	138.82
	DiffPure [11]	40.97	49.44	75.56	34.74	52.20	70.97	24.91	58.82	82.08
	F3 [12]	**36.40**	58.85	**116.58**	36.09	**69.45**	131.56	22.57	70.42	139.10
	DiffCAP [13]	37.94	60.73	107.56	39.47	69.19	115.42	22.20	69.64	122.09
	FASDP (Ours)	37.80	**62.53**	115.90	**30.21**	67.39	**133.85**	**20.47**	**77.11**	**143.85**

Defense effectiveness.

Table 1 reports the ASR and post-purification task performance of the compared defenses under the three attacks, together with each model's harmful-response rate and clean-task performance in the absence of an image attack. Relative to the no-attack reference, all three attacks increase the harmful-response rate and degrade task performance. Across the evaluated settings, FASDP attains the lowest ASR in a clear majority, and in the remaining cases its ASR stays close to the best value attained by any defense. Among the compared defenses, DiffPure incurs the largest utility cost, with post-purification scores below those of the undefended models under the same attack, whereas FASDP attains task performance closest to the no-attack reference in most settings.

Purified image quality. Table 2 reports the SSIM, LPIPS, and CLIP similarity between the purified images and the corresponding clean images on adversarial and clean inputs. On adversarial inputs, FASDP attains the highest SSIM and CLIP similarity and the lowest LPIPS among the compared defenses. On clean inputs, FASDP attains the highest SSIM and CLIP similarity, indicating that purification leaves the image content largely intact.

Table 3 reports the benign-input performance of each defense, measured by VQA accuracy on VQAv2 and CIDEr on COCO captioning across the three target models. JPEG compression discards fine image detail and lowers benign scores only slightly, whereas DiffPure degrades them substantially. After processing clean images, FASDP

Table 2: Purified image quality of the compared defenses on adversarial and clean inputs.

Defense	Adversarial inputs			Clean inputs		
	SSIM ↑	LPIPS ↓	CLIP ↑	SSIM ↑	LPIPS ↓	CLIP ↑
No Defense	65.33	32.05	82.48	–	–	–
JPEG [8]	78.18	35.59	87.84	92.93	**4.89**	90.71
DiffPure [11]	64.98	27.30	84.61	68.08	26.56	85.52
DiffCAP [13]	79.71	33.07	88.35	79.08	21.26	88.26
FASDP (Ours)	**87.00**	**17.71**	**95.45**	**95.26**	7.97	**94.94**

Table 3: Benign-input performance of the compared defenses on three classic open-source VLMs.

Defense	LLaVA-1.5-7B		InstructBLIP-7B		Qwen2.5-VL-7B	
	VQA Acc. ↑	CIDEr ↑	VQA Acc. ↑	CIDEr ↑	VQA Acc. ↑	CIDEr ↑
No Defense	67.13	122.48	72.03	136.53	81.60	144.31
JPEG [8]	64.50	112.92	68.80	134.39	**78.07**	138.49
DiffPure [11]	50.33	79.43	58.17	77.92	59.57	90.17
DiffCAP [13]	62.07	107.96	**69.47**	115.90	72.13	122.91
FASDP (Ours)	**67.11**	**116.62**	67.50	**135.07**	**78.07**	**143.88**

keeps task performance close to that of the undefended models on all three VLMs, and its benign utility is comparable to that of the compared defenses.

5.3 Visualization Analysis

This section visualizes purified outputs to illustrate differences in visible noise and content preservation.

Purified outputs. Figure 3 compares the purified outputs of representative defenses on VAE-attacked inputs. In these examples, the VAE perturbation appears as dense texture noise. JPEG compression smooths visible noise but blurs fine structures; DiffPure reduces visible noise yet introduces pronounced artifacts; DiffCAP preserves the global layout but over-smooths textured regions. FASDP reduces visible noise while retaining more detail and color fidelity in the shown examples, consistent with the fidelity results in Table 2.

Residual perturbations. Figure 4 shows two further examples, each with its purified output and the pixel-wise residual against the clean source. A darker residual map indicates closer agreement with the clean image. The attacked input leaves a residual of dense unstructured noise, whereas the compared defenses retain visible semantic structure. JPEG compression over-smooths and leaves pronounced edge traces, DiffPure gives the brightest residual and the strongest content drift, and DiffCAP lowers the residual magnitude yet distorts object contours. FASDP yields the faintest residuals, confined to thin object boundaries and thus closest to the clean



Fig. 3: Qualitative comparison of purified outputs under the VAE attack.

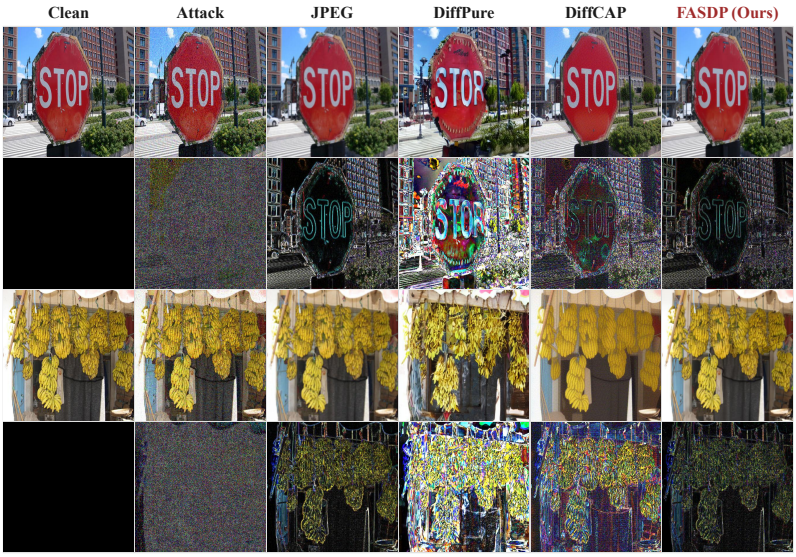


Fig. 4: Each case under the VAE attack shows the purified output together with the pixel-wise residual against the clean source.

images at pixel level. Residual magnitude alone does not determine jailbreak success; it measures only how much the image has changed.

Attention analysis. Figure 5 compares Grad-CAM activation maps of the target model under the VAE attack and representative defenses. The attack spreads dense, high-frequency activation across the maps and displaces the main object, and all defenses suppress this pattern. FASDP removes the speckle while keeping the main object activated.

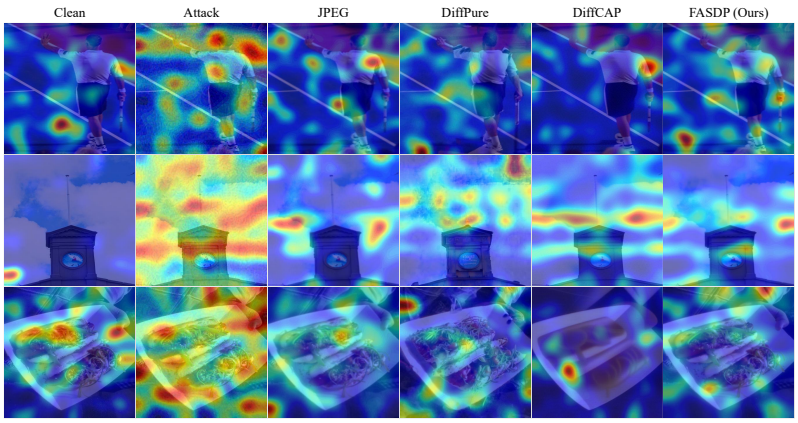


Fig. 5: Grad-CAM visualizations of the target model under the VAE attack and representative defenses.

5.4 Ablation Study

We conduct a leave-one-out ablation study to examine the role of each module of FASDP, disabling one module at a time from the frozen reference configuration. All variants are evaluated on Qwen2.5-VL-7B under the Image Hijacks attack.

Frequency prior. Removing the prior increases ASR, while benign performance remains closer to the full model than in the scale and fusion ablations. This supports the contribution of frequency modulation; input-dependent consistency weights still provide adaptivity without the prior.

Scale diversity. Reduced to a single fixed scale, security degrades only slightly, while fidelity and benign utility take the largest drops of any variant; the scale set enables a per-image choice of strength rather than a single fixed strength.

Consistency fusion. Replaced by uniform averaging, both security and benign-side performance degrade together, since every scale then contributes equally; the consistency weights balance detail preservation against perturbation suppression.

Shared denoiser. Removing the denoiser leaves scale transformations and fusion, and yields the highest ASR among the examined variants. Benign utility is higher than for the full model, showing that denoising contributes to both attack suppression and the utility cost.

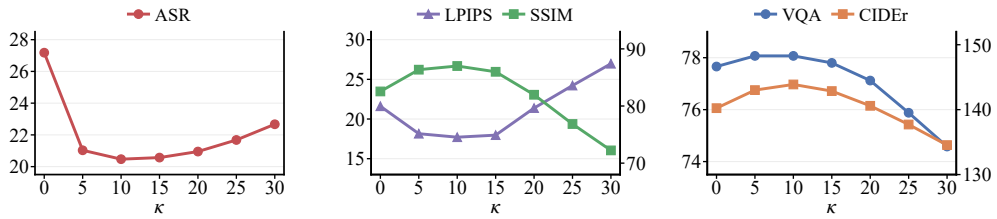
5.5 Hyperparameter Analysis

This section examines four scalar hyperparameters: the prior strength κ, the consistency mixing coefficient λ_{in}, the frequency-cutoff parameter γ, and the fidelity bias β. Each parameter is varied individually around the reference configuration while the others remain fixed.

Prior strength κ. Figure 6 varies the prior strength with the remaining hyperparameters fixed. ASR initially decreases and then increases at larger κ, while fidelity and benign utility also vary across the sweep. These trends show that stronger frequency modulation does not necessarily improve every evaluation metric.

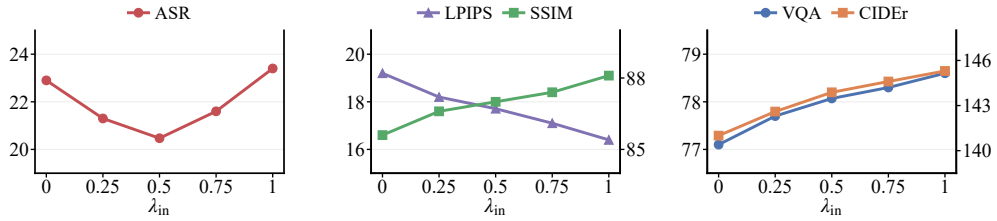
Table 4: Module ablation: ASR and fidelity under attack, and utility on benign images.

	Defense			Clean performance	
Variant	ASR ↓	SSIM ↑	LPIPS ↓	VQA Acc. ↑	CIDEr ↑
w/o prior	27.18	82.55	21.61	77.66	140.23
w/o multi-scale	23.17	70.46	28.14	65.81	122.64
w/o fusion	28.61	77.67	21.18	70.96	131.32
w/o denoiser	31.14	72.92	27.20	80.09	143.95
FASDP (full)	20.47	87.00	17.71	78.07	143.88



(a) Attack success rate (b) Purification fidelity (c) Benign-input utility

Fig. 6: Sensitivity of FASDP to the prior strength κ on attack effectiveness, purification fidelity, and benign-input utility.

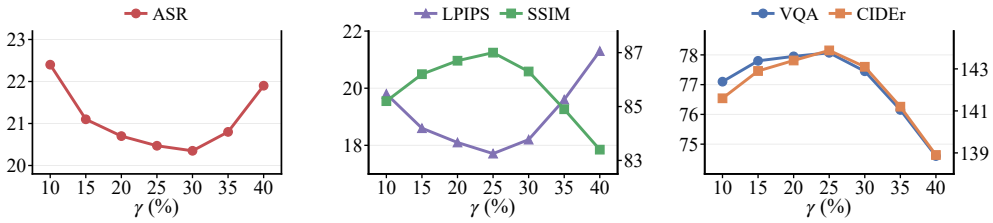


(a) Attack success rate (b) Purification fidelity (c) Benign-input utility

Fig. 7: Sensitivity of FASDP to the consistency mixing coefficient λ_{in} on attack effectiveness, purification fidelity, and benign-input utility.

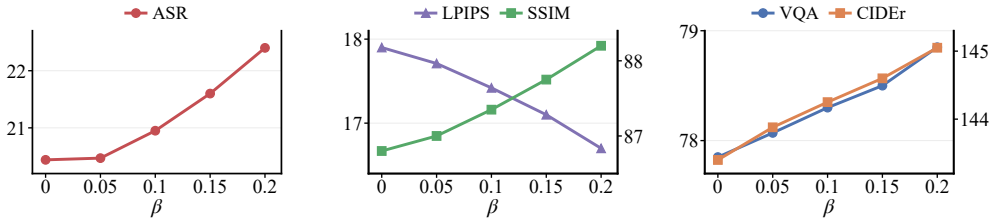
Consistency mixing λ_{in}. Figure 7 varies the input-anchor weight with the remaining hyperparameters fixed. ASR is lowest near the middle of the plotted range, while image fidelity and benign utility generally improve with larger input-anchor weights. This illustrates the trade-off between input fidelity and attack suppression across the tested settings.

Frequency cutoff γ. Figure 8 varies the cutoff parameter while keeping the remaining hyperparameters fixed. Increasing γ expands the selected frequency band



(a) Attack success rate (b) Purification fidelity (c) Benign-input utility

Fig. 8: Sensitivity of FASDP to the frequency-cutoff parameter γ on attack effectiveness, purification fidelity, and benign-input utility.



(a) Attack success rate (b) Purification fidelity (c) Benign-input utility

Fig. 9: Sensitivity of FASDP to the fidelity bias β on attack effectiveness, purification fidelity, and benign-input utility.

according to Eq. (2). The plotted results show a trade-off between the frequency band used for weighting and the measured safety and utility outcomes.

Fidelity bias β. Figure 9 varies the fidelity bias, which favors milder candidates without adding frequency information. Across the plotted settings, larger β improves fidelity and benign utility while increasing ASR, illustrating the corresponding safety–utility trade-off.

Scale configuration. Table 5 varies the candidate set from a single fixed scale to four, with the reference scale fixed. A single scale reproduces the fixed-strength trade-off, whereas coverage of the scale range matters more than the number of candidates: sets that reach the aggressive end approach the reference, and extending beyond it adds little suppression.

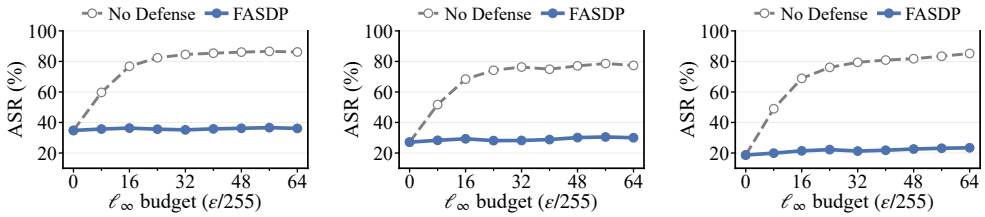
5.6 Robustness Analysis

Perturbation budget. Figure 10 shows the defense effectiveness of FASDP against the VAE attack under different perturbation budgets, with the budget swept up to 64/255, indicating that the method remains effective across all the budgets examined.

Denoiser instantiation. Table 6 swaps the shared denoiser with everything else frozen. Gaussian filtering degrades performance noticeably, DnCNN is intermediate, and the ASRs of Restormer, NAFNet, and SCUNet are close to one another. The

Table 5: Attack success rate and benign-input VQA accuracy of the scale configurations examined.

Scale set $\mathcal{S}$	ASR↓	VQA Acc.↑
$\{1.00\}$	26.88	75.62
$\{0.75\}$	23.63	73.66
$\{0.50\}$	23.17	65.81
$\{1.00, 0.75\}$	24.22	75.39
$\{1.00, 0.50\}$	21.81	75.61
$\{0.75, 0.50\}$	22.52	70.72
$\{1.00, 0.90, 0.80\}$	26.00	77.38
$\{1.00, 0.75, 0.50\}^*$	20.47	78.07
$\{1.00, 0.70, 0.40\}$	22.18	75.79
$\{1.00, 0.90, 0.80, 0.70\}$	24.04	75.62
$\{1.00, 0.80, 0.60, 0.40\}$	21.26	77.71
$\{1.00, 0.75, 0.50, 0.25\}$	20.57	76.50



(a) LLaVA-1.5-7B (b) InstructBLIP-7B (c) Qwen2.5-VL-7B

Fig. 10: Attack success rate of the undefended model and of FASDP under the VAE attack as the perturbation budget grows from 0 to 64/255.

results support compatibility with a range of denoisers and indicate that FASDP is relatively robust to the choice of denoiser.

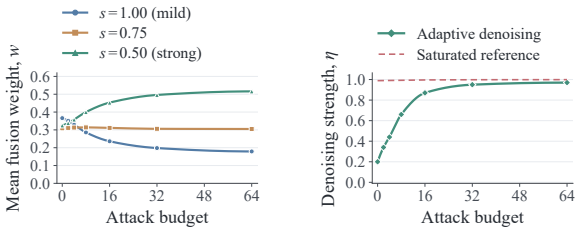
5.7 Analysis of Per-Image Strength Allocation

We analyze the response of the per-image allocation to the perturbation budget, reporting the allocated strength η and the induced fusion weights under the VAE attack on Qwen2.5-VL-7B.

Response to the perturbation budget. Figure 11 shows the allocated strength rising monotonically with the budget and approaching its upper bound at the evaluation budget, whereas the mean weight of the mild candidate declines and that of the aggressive candidate increases. The adjustment therefore shifts from the strength channel to the fusion weights as the strength approaches its bound.

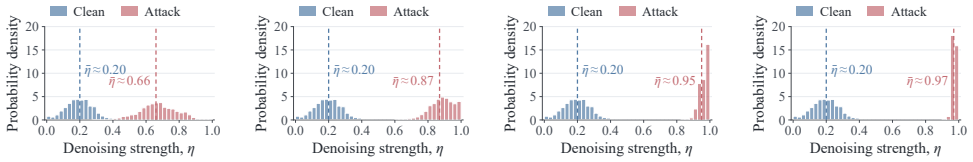
Table 6: Denoiser robustness under Image Hijacks on Qwen2.5-VL-7B.

	Defense			Clean performance	
Denoiser	ASR ↓	SSIM ↑	LPIPS ↓	VQA Acc. ↑	CIDEr ↑
Gaussian	30.32	66.28	32.05	75.14	135.78
DnCNN [51]	26.56	81.06	24.02	76.32	139.61
Restormer [60]	21.35	87.07	17.86	77.72	143.86
NAFNet [53]	21.01	85.39	18.84	78.38	142.44
SCUNet (reference) [54]	20.47	87.00	17.71	78.07	143.88



(a) Mean fusion weights. (b) Denoising strength.

Fig. 11: Response of the per-image allocation to the perturbation budget.



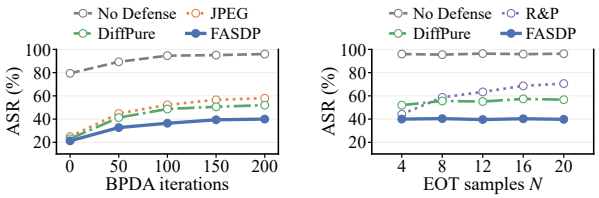
(a) $\varepsilon = 8/255$ (b) $\varepsilon = 16/255$ (c) $\varepsilon = 32/255$ (d) $\varepsilon = 64/255$

Fig. 12: Distribution of the allocated denoising strength on clean and attacked inputs across four perturbation budgets.

Separation between clean and attacked inputs. Figure 12 shows clean inputs concentrating in a narrow low range of allocated strength, whereas attacked inputs shift to higher strength as the budget grows. The limited overlap at the smallest budget implies that no fixed η serves both populations.

5.8 Adaptive Attacks

Under adaptive BPDA and EOT attacks [55–57] on Qwen2.5-VL-7B, the ASR of FASDP grows as the number of BPDA iterations increases and stays approximately at its plateau as the number N of transformations averaged by EOT grows (Figure 13),



(a) BPDA (b) EOT

Fig. 13: Attack success rate on Qwen2.5-VL-7B under adaptive attacks on the VAE perturbation. The BPDA panel compares no defense, JPEG, DiffPure, and FASDP; the EOT panel compares no defense, R&P, DiffPure, and FASDP.

so adaptive optimization recovers part of the effectiveness that the static setting suppresses. This increase is nevertheless the smallest among the compared defenses, whose ASR rises more steeply or settles at a markedly higher level, so FASDP remains the most robust purifier and the advantage of per-image strength allocation carries over to the adaptive case.

6 Limitations

FASDP attains strong protection on the perturbation-based jailbreaks evaluated here, but its scope is bounded by the image channel it acts on. Jailbreaks that exploit the alignment of the target model itself, such as text-only prompts [23, 24], and other visual jailbreak families remain outside what the purifier is designed to suppress, since purification neither rewrites the prompt nor repairs the model. Broadening the defense to these settings is left for future work.

7 Conclusion

We propose FASDP, a training-free purifier that allocates purification strength per image by modulating the low-frequency-consistency fusion of scale-diverse candidates with a self-referenced frequency statistic, using no target-model signals. Across the evaluated attacks on three VLMs, FASDP attains strong protection and improves purified-image fidelity, while keeping clean-image performance close to that of the undefended models.

References

[1] Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning (ICML), pp. 19730–19742 (2023)

[2] Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction

tuning. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 49250–49267 (2023)

[3] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems (NeurIPS) (2023). Oral presentation

[4] Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)

[5] Qi, X., Huang, K., Panda, A., Henderson, P., Wang, M., Mittal, P.: Visual adversarial examples jailbreak aligned large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 21527–21536 (2024)

[6] Shayegani, E., Dong, Y., Abu-Ghazaleh, N.: Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. In: International Conference on Learning Representations (ICLR) (2024)

[7] Gong, Y., Ran, D., Liu, J., Wang, C., Cong, T., Wang, A., Duan, S., Wang, X.: Figstep: Jailbreaking large vision-language models via typographic visual prompts. In: Proceedings of the AAAI Conference on Artificial Intelligence (2025). Oral presentation

[8] Guo, C., Rana, M., Cisse, M., Maaten, L.: Countering adversarial images using input transformations. In: International Conference on Learning Representations (ICLR) (2018)

[9] Xie, C., Wang, J., Zhang, Z., Ren, Z., Yuille, A.: Mitigating adversarial effects through randomization. In: International Conference on Learning Representations (ICLR) (2018)

[10] Xu, W., Evans, D., Qi, Y.: Feature squeezing: Detecting adversarial examples in deep neural networks. In: Network and Distributed Systems Security Symposium (NDSS) (2018)

[11] Nie, W., Guo, B., Huang, Y., Xiao, C., Vahdat, A., Anandkumar, A.: Diffusion models for adversarial purification. In: International Conference on Machine Learning (ICML), pp. 16805–16827 (2022)

[12] Zhang, Y., Xie, R., Huang, Y., Chen, J., Sun, X., Kang, Z., Wang, D., Wang, Y.: Fighting fire with fire (f3): A training-free and efficient visual adversarial example purification method in lvlms. In: Proceedings of the 33rd ACM International Conference on Multimedia (2025). https://doi.org/10.1145/3746027.3758152

[13] Fu, J., Wu, Y., Chen, Y., Peng, K., Zhang, X., Cevher, V., Pashami, S., Holst, A.: Diffusion-based cumulative adversarial purification for vision language models. Transactions on Machine Learning Research (2026)

[14] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML), pp. 8748–8763 (2021)

[15] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)

[16] Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)

[17] Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., Li, B., Luo, P., Lu, T., Qiao, Y., Dai, J.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 24185–24198 (2024)

[18] OpenAI: GPT-4V(ision) System Card. https://cdn.openai.com/papers/GPTV_System_Card.pdf (2023)

[19] Gemini Team: Gemini: A family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)

[20] Gu, T., Dolan-Gavitt, B., Garg, S.: Badnets: Evaluating backdooring attacks on deep neural networks. IEEE Access **7**, 47230–47244 (2019)

[21] Guo, J., Qin, X., Liu, C., Wang, J., Chen, J., Jiang, W.: CBV: Clean-label backdoor attacks on vision language models via diffusion models. In: Proceedings of the 43rd International Conference on Machine Learning (ICML). PMLR, vol. 306 (2026)

[22] Carlini, N., Jagielski, M., Choquette-Choo, C.A., Paleka, D., Pearce, W., Anderson, H., Terzis, A., Thomas, K., Tramer, F.: Poisoning web-scale training datasets is practical. In: IEEE Symposium on Security and Privacy (S&P), pp. 407–425 (2024)

[23] Wei, A., Haghtalab, N., Steinhardt, J.: Jailbroken: How does LLM safety training fail? In: Advances in Neural Information Processing Systems (NeurIPS) (2023)

[24] Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J.Z., Fredrikson, M.: Universal and transferable adversarial attacks on aligned language models. arXiv preprint arXiv:2307.15043 (2023)

[25] Li, Y., Guo, H., Zhou, K., Zhao, W.X., Wen, J.-R.: Images are Achilles' heel of alignment: Exploiting visual vulnerabilities for jailbreaking multimodal large language models. In: European Conference on Computer Vision (ECCV), pp. 174–189 (2024)

[26] Zhao, Y., Pang, T., Du, C., Yang, X., Li, C., Cheung, N.-M., Lin, M.: On evaluating adversarial robustness of large vision-language models. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)

[27] Liu, X., Zhu, Y., Gu, J., Lan, Y., Yang, C., Qiao, Y.: Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In: European Conference on Computer Vision (ECCV) (2024)

[28] Jia, X., Liao, J., Guo, Q., Ma, T., Qin, S., Duan, R., Li, T., Huang, Y., Zeng, Z., Wu, D., Li, Y., Ren, W., Cao, X., Liu, Y.: Omnisafebench-mm: A unified benchmark and toolbox for multimodal jailbreak attack-defense evaluation. arXiv preprint arXiv:2512.06589 (2025)

[29] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: International Conference on Learning Representations (ICLR) (2014)

[30] Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: International Conference on Learning Representations (ICLR) (2015)

[31] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (ICLR) (2018)

[32] Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: Proc. IEEE Symposium on Security and Privacy (S&P), pp. 39–57 (2017)

[33] Bailey, L., Ong, E., Russell, S., Emmons, S.: Image hijacks: Adversarial images can control generative models at runtime. In: International Conference on Machine Learning (ICML), pp. 2443–2455 (2024)

[34] Liu, X., Cui, X., Li, P., Li, Z., Huang, H., Xia, S., Zhang, M., Zou, Y., He, R.: Jailbreak attacks and defenses against multimodal generative models: A survey. arXiv preprint arXiv:2411.09259 (2024)

[35] Oh, S., Jin, Y., Sharma, M., Kim, D., Ma, E., Verma, G., Kumar, S.: Uniguard: Towards universal safety guardrails for jailbreak attacks on multimodal large language models. arXiv preprint arXiv:2411.01703 (2024)

[36] Chen, Y., Li, H., Zhang, Y., Zheng, Z., Song, Y., Hooi, B.: Bathe: Defense against the jailbreak attack in multimodal large language models by treating harmful

instruction as backdoor trigger. arXiv preprint arXiv:2408.09093 (2024)

[37] Hao, S., Wang, Y., Hooi, B., Yang, M.-H., Liu, J., Tang, C., Huang, Z., Cai, Y.: Tit-for-tat: Safeguarding large vision-language models against jailbreak attacks via adversarial defense. arXiv preprint arXiv:2503.11619 (2025)

[38] Chen, B., Lyu, X., Gao, L., Song, J., Shen, H.T.: Safeptr: Token-level jailbreak defense in multimodal llms via prune-then-restore mechanism. In: Advances in Neural Information Processing Systems (2025)

[39] Lu, L., Gu, X., Pang, S., Liang, S., Zhu, H., Zeng, X., Zheng, X., Zhou, Y.: E2at: Multimodal jailbreak defense via dynamic joint optimization for multimodal large language models. arXiv preprint arXiv:2503.04833 (2025)

[40] Gou, Y., Chen, K., Liu, Z., Hong, L., Xu, H., Li, Z., Yeung, D.-Y., Kwok, J.T., Zhang, Y.: Eyes closed, safety on: Protecting multimodal LLMs via image-to-text transformation. In: European Conference on Computer Vision (ECCV), pp. 388–404 (2024)

[41] Jia, X., Wei, X., Cao, X., Foroosh, H.: Comdefend: An efficient image compression model to defend adversarial examples. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6077–6085 (2019)

[42] Song, Y., Kim, T., Nowozin, S., Ermon, S., Kushman, N.: Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In: International Conference on Learning Representations (ICLR) (2018)

[43] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 6840–6851 (2020)

[44] Salman, H., Sun, M., Yang, G., Kapoor, A., Kolter, J.Z.: Denoised smoothing: A provable defense for pretrained classifiers. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)

[45] Cohen, J.M., Rosenfeld, E., Kolter, J.Z.: Certified adversarial robustness via randomized smoothing. In: International Conference on Machine Learning (ICML), pp. 1310–1320 (2019)

[46] Yang, G., Duan, T., Hu, J.E., Salman, H., Razenshteyn, I., Li, J.: Randomized smoothing of all shapes and sizes. In: International Conference on Machine Learning (ICML) (2020)

[47] Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., Madry, A.: Adversarial examples are not bugs, they are features. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)

[48] Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., Madry, A.: Robustness may be at odds with accuracy. In: International Conference on Learning Representations (ICLR) (2019)

[49] Wang, H., Wu, X., Huang, Z., Xing, E.P.: High frequency component helps explain the generalization of convolutional neural networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020). Oral presentation

[50] Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3-d transform-domain collaborative filtering. IEEE Transactions on Image Processing **16**(8), 2080–2095 (2007)

[51] Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. IEEE Transactions on Image Processing **26**(7), 3142–3155 (2017)

[52] Zhang, K., Zuo, W., Zhang, L.: Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. IEEE Transactions on Image Processing **27**(9), 4608–4622 (2018)

[53] Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. In: European Conference on Computer Vision (ECCV), pp. 17–33 (2022)

[54] Zhang, K., Li, Y., Liang, J., Cao, J., Zhang, Y., Tang, H., Fan, D.-P., Timofte, R., Van Gool, L.: Practical blind image denoising via swin-conv-unet and data synthesis. Machine Intelligence Research **20**, 822–836 (2023)

[55] Athalye, A., Carlini, N., Wagner, D.: Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In: International Conference on Machine Learning (ICML), pp. 274–283 (2018)

[56] Tramer, F., Carlini, N., Brendel, W., Madry, A.: On adaptive attacks to adversarial example defenses. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 1633–1645 (2020)

[57] Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: International Conference on Machine Learning (ICML), pp. 2206–2216 (2020)

[58] Jiang, L., Rao, K., Han, S., Ettinger, A., Brahman, F., Kumar, S., Mireshghallah, N., Lu, X., Sap, M., Choi, Y., Dziri, N.: Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 37, pp. 47094–47165 (2024)

[59] Qwen Team: Qwen2.5-VL Technical Report. arXiv preprint arXiv:2502.05423 (2025)

[60] Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.-H.: Restormer: Efficient transformer for high-resolution image restoration. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)

Declarations

Data availability. All datasets analysed in this study are publicly available: WildJailbreak at https://huggingface.co/datasets/allenai/wildjailbreak, COCO at https://cocodataset.org, and VQAv2 at https://visualqa.org. Implementations of the evaluated attacks and defense baselines are obtained from the sources cited in Sect. 5.1. No new datasets were created in this study.