PureProof: Diffusion-Resistant Black-box Targeted Attack on Large Vision-Language Models

Supplementary Material

6. Appendix

6.1. Proof of Theorem 1

We provide a detailed proof for Theorem 1 in Sec. 3.

Notation. During adversarial optimization in PureProof, at each iteration the current adversarial image $\mathbf{x}_{\text{adv}}$ is forward-noised to a randomly sampled timestep $t \sim \text{Unif}\{1, \dots, T_p\}$, producing $\mathbf{x}_t$. The diffusion model predicts a clean estimate $\hat{\mathbf{x}}_0(\mathbf{x}_t, t)$, and we re-noise it as:

$$\tilde{\mathbf{x}}_t = \sqrt{\bar{\alpha}_t}\,\hat{\mathbf{x}}_0(\mathbf{x}_t, t) + \sigma_t \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{12}$$

where $\sigma_t^2 = 1 - \bar{\alpha}_t$. Its mean is:

$$\boldsymbol{\mu}_t = \mathbb{E}[\tilde{\mathbf{x}}_t] = \sqrt{\bar{\alpha}_t}\,\hat{\mathbf{x}}_0(\mathbf{x}_t, t). \tag{13}$$

We define the alignment loss as the negative cosine similarity between the CLIP image embedding of the image $\mathbf{x}$ and that of the target image $\mathbf{x}_{\text{tar}}$:

$$\ell(\mathbf{x}) = -sim(\mathbf{x}, \mathbf{x}_{\text{tar}}). \tag{14}$$

Assumptions. We next introduce two mild assumptions.

(A1) Local smoothness. We assume that ℓ is twice differentiable with bounded third derivatives in a neighborhood of $\boldsymbol{\mu}_t$. This is a mild practical assumption, as ℓ is composed of standard differentiable operations (CLIP embeddings and cosine similarity) within a bounded image region.

(A2) Non-negative average Hessian trace in neighborhood. We assume that the trace of the Hessian of ℓ in a neighborhood of $\boldsymbol{\mu}_t$ is non-negative on average. This assumption simply excludes neighborhoods that are strongly concave on average and does not require global convexity. It is used only to support the interpretation of the second-order term.

Proof. We perform a second-order Taylor expansion of $\ell(\tilde{\mathbf{x}}_t)$ around the mean $\boldsymbol{\mu}_t$:

$$\begin{aligned}\ell(\tilde{\mathbf{x}}_t) = \ell(\boldsymbol{\mu}_t) &+ \nabla\ell(\boldsymbol{\mu}_t)^\top(\tilde{\mathbf{x}}_t - \boldsymbol{\mu}_t) \\ &+ \frac{1}{2}(\tilde{\mathbf{x}}_t - \boldsymbol{\mu}_t)^\top H_\ell(\boldsymbol{\mu}_t)(\tilde{\mathbf{x}}_t - \boldsymbol{\mu}_t) \\ &+ R_t^{\text{raw}},\end{aligned} \tag{15}$$

where the remainder R_t^{raw} satisfies $R_t^{\text{raw}} = \mathcal{O}\big(\|\tilde{\mathbf{x}}_t - \boldsymbol{\mu}_t\|^3\big)$.

Next, we take expectation over the Gaussian noise $\boldsymbol{\epsilon}$ in the Taylor expansion of $\ell(\tilde{\mathbf{x}}_t)$ from Eq. (15).

The first-order term in the expansion is:

$$\nabla\ell(\boldsymbol{\mu}_t)^\top(\tilde{\mathbf{x}}_t - \boldsymbol{\mu}_t) = \sigma_t \nabla\ell(\boldsymbol{\mu}_t)^\top \boldsymbol{\epsilon}. \tag{16}$$

Since $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$, its expectation vanishes:

$$\mathbb{E}\big[\nabla\ell(\boldsymbol{\mu}_t)^\top(\tilde{\mathbf{x}}_t - \boldsymbol{\mu}_t)\big] = \sigma_t \nabla\ell(\boldsymbol{\mu}_t)^\top \mathbb{E}[\boldsymbol{\epsilon}] = 0. \tag{17}$$

For the second-order term, we have:

$$(\tilde{\mathbf{x}}_t - \boldsymbol{\mu}_t)(\tilde{\mathbf{x}}_t - \boldsymbol{\mu}_t)^\top = \sigma_t^2 \boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top. \tag{18}$$

Since $\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top] = \mathbf{I}$, thus:

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\epsilon}}\big[(\tilde{\mathbf{x}}_t - \boldsymbol{\mu}_t)^\top H_\ell(\boldsymbol{\mu}_t)(\tilde{\mathbf{x}}_t - \boldsymbol{\mu}_t)\big] \\ &\quad= \sigma_t^2\, \mathbb{E}_{\boldsymbol{\epsilon}}\big[\boldsymbol{\epsilon}^\top H_\ell(\boldsymbol{\mu}_t)\boldsymbol{\epsilon}\big] \\ &\quad= \sigma_t^2\, \text{tr}(H_\ell(\boldsymbol{\mu}_t))\,.\end{aligned} \tag{19}$$

Finally, for the remainder term, we have:

$$\mathbb{E}_{\boldsymbol{\epsilon}}[|R_t^{\text{raw}}|] = \mathcal{O}(\sigma_t^3), \tag{20}$$

since $\mathbb{E}[\|\boldsymbol{\epsilon}\|^3] < \infty$ for Gaussian $\boldsymbol{\epsilon}$. Let R_t denote the signed expected remainder. Its magnitude is therefore bounded by $\mathcal{O}(\sigma_t^3)$.

Combining all terms:

$$\mathbb{E}_{\boldsymbol{\epsilon}}[\ell(\tilde{\mathbf{x}}_t)] = \ell(\boldsymbol{\mu}_t) + \frac{1}{2}\sigma_t^2 \text{tr}\big(H_\ell(\boldsymbol{\mu}_t)\big) + R_t. \tag{21}$$

Substituting $\boldsymbol{\mu}_t = \mathbb{E}[\tilde{\mathbf{x}}_t]$ yields the expression in Theorem 1, which formally shows that Adaptive Re-noising Augmentation (ARA) acts as a curvature-aware regularizer. The second-order term penalizes regions with high local curvature, effectively smoothing the loss landscape. As σ_t^2 grows with t, ARA applies stronger regularization at noisier timesteps where $\hat{\mathbf{x}}_0(\mathbf{x}_t, t)$ is less reliable, while smaller timesteps with more accurate clean image estimates contribute more to alignment. The first term ensures the mean of the re-noised sample remains aligned with the target and the remainder captures higher-order effects that are typically negligible in practice.

$$\boxed{\mathbb{E}_{\boldsymbol{\epsilon}}[\ell(\tilde{\mathbf{x}}_t)] = \ell(\mathbb{E}[\tilde{\mathbf{x}}_t]) + \frac{1}{2}\sigma_t^2 \text{tr}(H_\ell(\mathbb{E}[\tilde{\mathbf{x}}_t])) + R_t.} \tag{22}$$

6.2. More Implementation Details

Implementation details of DBPs. All purification settings follow the configurations reported in the original papers. We evaluate three diffusion-based purification defenses: DiffPure, GDMP, and LM. For DiffPure, we set the normalized diffusion timestep to $t^* = 0.15$. For GDMP, we set the purification length to 45 under DDPM acceleration, which respaces the original $T = 1000$ steps to 250

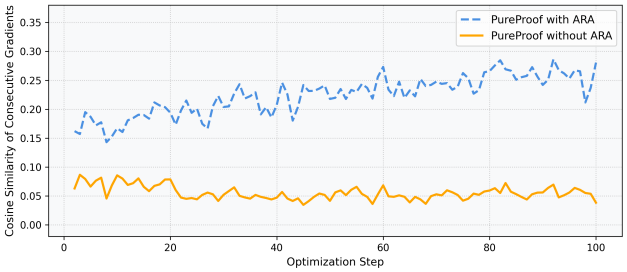


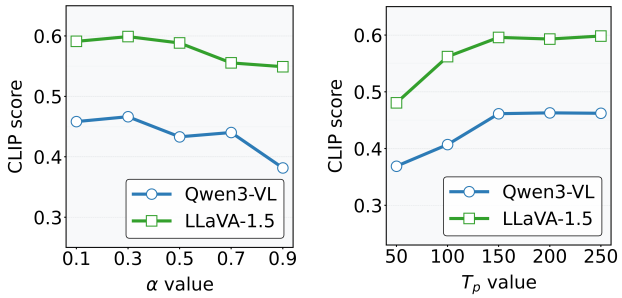
Figure 5. Stability of gradient directions during optimization. We show the average cosine similarity between consecutive gradients for PureProof with and without Adaptive Re-noising Augmentation (ARA) across optimization steps. Higher similarity for PureProof with ARA indicates that gradients maintain more consistent directions, demonstrating that ARA stabilizes the optimization process by effectively smoothing the loss landscape.

steps. SSIM-based guidance is implemented as in the original work. For DiffPure and GDMP, the pretrained diffusion model is Guided Diffusion for the ImageNet experiments. For LM, we adopt the EDM diffusion model released with the original paper and perform likelihood maximization with $N = 5$ optimization steps.

Configurations of VLMs. Our evaluation covers a diverse suite of open-source VLMs with varying architectures and scales, including LLaVA-1.5, MiniGPT-4, LLaVA-1.6, Gemma 3, and Qwen3-VL. For the vision backbones and language model of the VLMs, we use the following configurations. LLaVA-1.5 employs a CLIP ViT-L/14 vision encoder paired with Vicuna-7B-v1.5. MiniGPT-4 uses the EVA-CLIP ViT-g/14 vision encoder with Vicuna-13B. LLaVA-1.6 adopts the CLIP ViT-L/14 encoder together with Mistral-7B. Gemma 3 incorporates a 400M-parameter SigLIP vision encoder and the Gemma-3-27B language model. Qwen3-VL-8B-Instruct adopts the Qwen3-8B language model. These models span parameter sizes from 7B (LLaVA-1.5) to 27B (Gemma 3). For GPT-5 and Gemini-2.5-Flash, reasoning capabilities are enabled in all experiments, with no manual tuning of default black-box API parameters. For GPT-5, we use the default reasoning effort `medium`. For Gemini-2.5-Flash, the default dynamic thinking mode is used, which automatically allocates reasoning based on input complexity.

6.3. More Experimental Results

Analysis of ARA's smoothing effect. To quantify the impact of Adaptive Re-noising Augmentation (ARA) on the optimization landscape of PureProof, we randomly sample 20 images and track the evolution of gradient directions during adversarial optimization. Specifically, we compute the average cosine similarity between consecutive gradients at each optimization step, comparing PureProof with and without ARA, as illustrated in Fig. 5. The results demonstrate that PureProof with ARA consistently exhibits higher gradient similarity across steps. This indicates that the gradient directions remain more stable, with updates pointing in more consistent directions compared to the baseline without ARA, where gradients change directions more. Notably, the gradient similarity for PureProof with ARA shows a gradual increasing trend over the course of optimization, suggesting that the optimization trajectory becomes progressively more stable as training proceeds. In contrast, the gradient similarity without ARA remains low and does not exhibit a comparable trend. These observations provide empirical evidence that ARA effectively smooths the loss landscape. By introducing timestep-dependent re-noising of the clean image estimate, ARA reduces the stochasticity inherent in the diffusion process, aligns gradient directions, and promotes robust updates. As a result, the optimization process is stabilized, which facilitates more reliable convergence and ultimately enhances attack effectiveness.



(a) α (b) T_p

Figure 6. Left: Attack efficacy of PureProof with varying α values. Right: Attack efficacy of PureProof with varying T_p values.

Ablation study on α. We examine the effect of the weighting coefficient α in the final PureProof objective (Eq. (11)) on Qwen3-VL and LLaVA-1.5 against DiffPure. As shown in Fig. 6a, we vary $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. The attack performance consistently achieves its maximum at $\alpha = 0.3$ across both models, suggesting that this value provides the best balance between the two alignment objectives. This indicates that while Stochastic Reverse Alignment (SRA) and Adaptive Re-noising Augmentation (ARA) both improve attack strength, ARA contributes more significantly to producing diffusion-resistant adversarial examples.

Ablation study on T_p. We investigate how the maximum sampling timestep T_p affects the performance of PureProof against DiffPure. As shown in Fig. 6b, we vary $T_p \in \{50, 100, 150, 200, 250\}$ on Qwen3-VL and LLaVA-1.5. Attack performance consistently increases from 50 to 150. This trend suggests that sampling from a sufficiently wide timestep range is crucial because relatively larger timesteps expose the optimization to noisier diffusion states, enabling the surrogate model to produce more challenging

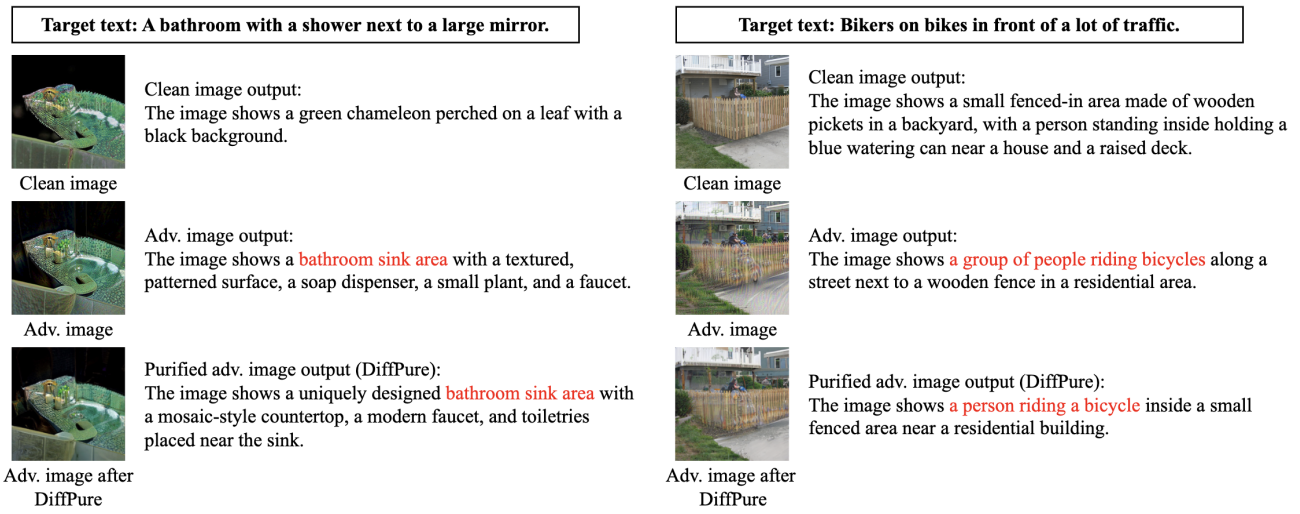


Figure 7. Qualitative results of PureProof against diffusion-based purification (DBP) on OpenAI GPT-5. For each target text (top), we show the clean image, the adversarial image produced by PureProof, and the DBP-purified adversarial image, each paired with the model's generated description. The model's responses highlighted in red reflect the semantic content of the target. PureProof successfully preserves the target semantics even after DBP purification.

Table 5. Attack performance comparison of PureProof and DG-cos on VLMs against different DBP defenses. We report the Ensemble CLIP score (↑). The gray shading highlights our proposed method, while **bold** numbers indicate the best results.

VLM model	Method	DiffPure	GDMP	LM
LLaVA-1.5	DG-cos	0.5041	0.4977	0.4919
	PureProof	**0.5983**	**0.5540**	**0.6231**
MiniGPT-4	DG-cos	0.4675	0.4594	0.4632
	PureProof	**0.5453**	**0.5105**	**0.5689**
Qwen3-VL	DG-cos	0.3626	0.3597	0.3430
	PureProof	**0.4493**	**0.4140**	**0.4613**

clean image estimates. This in turn leads to stronger semantic alignment and more effective adversarial perturbations. The improvement saturates around $T_p = 150$. This also matches DiffPure's diffusion timestep. Once PureProof's sampling range reaches this optimal noise level, expanding it further no longer brings additional significant gains.

Comparison with DiffGrad. DiffGrad [16] was recently proposed as a reliable differentiation module that enables backpropagation through diffusion-based purification (DBP), leading to stronger adaptive attacks than prior baselines. To ensure a fair comparison, we adapt DiffGrad to our black-box VLM setting. Because the original method targets white-box classification, we reformulate it as a transfer-based attack by employing our surrogate diffusion model and replacing its classification loss with a cosine-similarity objective between adversarial and target image embeddings (denoted as DG-cos). The results in Tab. 5 show that PureProof consistently achieves higher Ensemble CLIP scores across DiffPure, GDMP, and LM, indicating stronger targeted attack performance under DBP than the adapted DiffGrad variant. This suggests that, although DiffGrad reduces noise in gradient estimation, it does not address the intrinsic randomness of stochastic diffusion-based purification. In contrast, PureProof explicitly addresses this stochasticity through Adaptive Re-noising Augmentation (ARA), which effectively smooths the loss landscape and significantly improves attack robustness in black-box DBP settings.

More qualitative results. Fig. 7 provides the qualitative examples of PureProof against DBP on OpenAI GPT-5. For each target description, we show the clean image with its model-generated caption, the adversarial image generated by PureProof along with its model-generated caption, and the corresponding purified adversarial image after applying DiffPure together with the model-generated caption. PureProof reliably steers the model toward the target semantics. For instance, although the clean image is originally described as "a green chameleon perched on a leaf", the adversarial image is interpreted as depicting "a bathroom sink area", and this targeted interpretation persists after purification by DiffPure. These observations demonstrate that PureProof generates diffusion-resistant adversarial perturbations that remain effective against strong DBP defenses, consistently causing the model to output the target semantics despite purification.