Journal of the Brazilian Computer Society, 2026, 32:1, doi: 10.5753/jbcs.2026.6584

CertifiedGPT: Evaluating Adversarial Robustness of Vision and Language Models Against Targeted Black-Box Attacks

Leonardo Souza [**Universidade Federal do Estado do Rio de Janeiro (UNIRIO)** | ***leonardo.souza@edu.unirio.br***]
Pedro Nuno de Souza Moura [**Universidade Federal do Estado do Rio de Janeiro (UNIRIO)** | ***pedro.moura@uniriotec.br***]
Maíra Gatti [**Universidade Federal do Estado do Rio de Janeiro (UNIRIO)** | ***gattidebayser@gmail.com***]

Graduate Program in Computer Science, Universidade Federal do Estado do Rio de Janeiro (UNIRIO), Av. Pasteur, 458 - Urca.

Received: 18 September 2025 • **Accepted:** 10 August 2026 • **Published:** 28 August 2026

Abstract. Multimodal Vision and Language Models (VLMs) have attracted significant academic interest due to their strong performance in tasks that require simultaneous inference across image and text modalities. The impact of VLMs is evident both in general society, through powerful chatbots such as ChatGPT and Google Gemini, and in specific industry sectors, including industrial processes, biomedical engineering, systems engineering, and medicine. In contrast to proprietary models, open-source VLMs such as MiniGPT-4 have emerged, offering compact alternatives that demonstrate strong performance. However, such VLMs show certain vulnerabilities to noisy input data (e.g., images), which can compromise the reliability of the output and can be explored by adversarial attacks, even under black-box access. More sophisticated attacks can also force a VLM to generate specific output using targeted attack configurations. In this work, we explored the concept of robustness certification in order to make a VLM robust against targeted black-box attacks through the use of a randomized smoothing technique. First, we fine-tuned the MiniGPT-4 model on a small subset of VQAv2 and applied Gaussian noise to the input images. We also adapted a smoothed method that encapsulates the original decoder and generates the most probable answer, even with the noise in the input images. Finally, we evaluated the model against targeted black-box attacks. Our certified version of MiniGPT-4, when evaluated on a small VQAv2 subset, produced statistically significant results, demonstrating that randomized smoothing is a feasible approach to certifying the robustness of VLMs, especially in scenarios where access to high-performance GPU is limited.

Keywords: Visual Language Models, Fine-tuning, Certified robustness, Randomized smoothing, Adversarial attacks

1 Introduction

Deep Learning (DL) models have demonstrated increasingly impressive results since 2012, when computer scientists Alex Krizhevsky and Ilya Sutskever, under the supervision of Geoffrey Hinton at the University of Toronto, introduced what became known as AlexNet during the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) Krizhevsky *et al*. [2012].

The advances in DL have also boosted other domains of Artificial Intelligence (AI), such as computer vision, speech recognition, and Natural Language Processing (NLP). Within DL research, the Transformer architecture (Islam *et al*. [2023]; Vaswani *et al*. [2017]) stands out as a major milestone, enabling NLP research to surpass the results achieved by traditional RNN and LSTM models. Notably, OpenAI's Generative Pre-trained Transformer (GPT) (Radford *et al*. [2018]) is a powerful NLP model behind ChatGPT.

According to Uppal *et al*. [2022], the evolution of DL models has driven research and the development of modern applications at the intersection of computer vision and NLP. In this context, Vision Language Models (VLM) have emerged and attracted substantial interest from the scientific community

Figure 1. The left panel shows the clean image; the right panel applies additive Gaussian noise with variance $\sigma = 0.5$, which is used during fine-tuning, certification, and smoothed prediction under the randomized smoothing technique. For adversarial attacks, this clean–noisy pair can initialize an iterative image-space optimization that learns a perturbation to alter a model's output, including VLMs such as MiniGPT-4, for specified tasks. Experiments demonstrate that a VQA task can be attacked in a targeted manner, forcing the model to produce an attacker specified answer while keeping the perturbation visually subtle.

Zhu *et al*. [2023]; Wei *et al*. [2023]; Zhao *et al*. [2023]; Sun *et al*. [2024].

Despite the impressive performance of VLMs and the increasing interest they have attracted from both industry and academia, their susceptibility to noise in the input images remains insufficiently explored (Yin *et al*. [2023]; Gao *et al*. [2024]). This type of issue is often leveraged to create adversarial examples, which are subtly altered inputs that are

nearly invisible to the human eye but can lead models to make incorrect predictions (Yigit and Amasyali [2023]). Such adversarial examples are commonly employed in adversarial attack strategies to manipulate the output of the target model.

Adversarial attacks can be categorized according to the attacker's level of access to the target model. In white-box attacks, the adversary has full knowledge of the model architecture, parameters, and gradients that can ben explored to maximize the attack success rate. In contrast, the black-box attacks assume no access to target model. This latter setting is particularly relevant in real-world scenarios, where proprietary VLM systems typically expose only prediction interfaces. Furthermore, adversarial attacks can be conducted using two distinct strategies: untargeted and targeted. In untargeted attacks, the objective is simply to cause the model to generate an incorrect answer. In targeted attacks, the objective is to induce the model to produce a predefined answer (Yigit and Amasyali [2023]; Fares *et al*. [2024]; Zhao *et al*. [2023]). According to Wang *et al*. [2023], targeted attacks raise significant security concerns for VLMs due to their potential impact on applications such as medical diagnosis of specific diseases.

To address this issue, we investigate robustness certification to increase the resilience of a VLM against targeted black-box attacks by adapting the randomized smoothing technique of Cohen *et al*. [2019], originally developed for classification, which provides probabilistic guarantees that the output remains unchanged under ℓ_2-bounded input perturbation induced by isotropic Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. We extend this technique to the multimodal VQA tasks, enabling a VLM to generate coherent text responses even when the input images associated with the question are perturbed.

In our work, we present *CertifiedGPT*, a framework designed to fine-tune and certify the robustness of a VLM. While our method can be applied to any VLM in future work, we begin our exploration with MiniGPT-4 (Zhu *et al*. [2023]), based on the following considerations: i) We had limited GPU resources to all stages of our work, including fine-tuning, certifying, and evaluating the model against adversarial attacks. ii) MiniGPT-4 appeared to be a cost-effective choice for our purposes due to its efficient fine-tuning and its reported 90% of ChatGPT's quality as measured by GPT-4's evaluation Zhu *et al*. [2023]. iii) A pretrained 7B checkpoint was released by its authors, which allowed us to replace the original second-stage fine-tuning with our own version using a small version of VQAv2 dataset to evaluate MiniGPT-4 learning and certify its robustness. iv) MiniGPT-4 is also known to be vulnerable to noise in the input image as shown in studies such as Zhao *et al*. [2023] and Fares *et al*. [2024].

1.1 Key Contributions

This work makes the following contributions to certify the robustness of VLMs in VQAv2 with textual generation:

- **VLM Certification Framework**: We present an adaptation of randomized smoothing certification specifically for VLMs performing VQAv2 tasks.
- **Decoder-to-Label Certification**: We provide a procedure that maps the MiniGPT-4 decoder's textual outputs to discrete labels and apply the certification algorithm of Cohen *et al*. [2019] with proper confidence bounds, preserving theoretical guarantees while operating on generated answers.
- **Robustness Guarantees**: The certified model ensures that small, imperceptible perturbations on input images do not change the predicted answer within an ℓ_2 radius determined by the smoothing parameters.
- **Evaluation Protocol for VQA**: We establish an evaluation protocol for VQA under certification that reports certified accuracy, coverage, and abstention over the normalized answer set, providing a reproducible setup for future studies.
- **Targeted Black-Box Evaluation**: We evaluate robustness under targeted black-box adversarial attacks, reporting attack success rate.
- **Practical Scalability**: We demonstrate a computationally feasible approach that operates with limited GPU resources.

1.2 Research Questions

According to prior work on multimodal attacks [Yin *et al*. [2023]], black-box adversarial attacks are more realistic since the adversary lacks access to the victim model's trained parameters, and targeted attacks enable the adversary to control the model's generated output. To address this issue, the study has two objective. (i) apply the certification via randomized smoothing to the standard VLM predictor, and (ii) conduct an empirical evaluation under targeted black-box attacks. We focused on the following research questions (RQs).

RQ1: What is the performance gain of incorporating robustness certification through Randomized Smoothing to a VLM predictor?

RQ2: To what extent does incorporating randomized smoothing improve a VLM's empirical robustness to targeted adversarial attacks that induce predefined question outputs in a black-box setting?

1.3 Hypotheses

Based on the issues discussed above, we realized that the application of robustness certification with randomized smoothing to VLM might produce appropriate answers to our research question, which resulted in two main hypotheses:

H_1**:** Applying randomized smoothing to a VLM increases its certified accuracy at a fixed robustness radius.

H_2**:** Incorporating randomized smoothing into a VLM improves its empirical robustness (resilience) against targeted black-box adversarial attacks.

2 Background

In the following subsections, the discussion reviews adversarial attacks and defenses, outlines robustness certification, and

details randomized smoothing together with its theoretical guarantees, providing the background needed for how this work builds upon and extends prior approaches. [Fares *et al*. [2024]; Cohen *et al*. [2019]].

The reliability of modern deep learning models is increasingly challenged by adversarial attacks, which are subtle yet maliciously crafted perturbations to input data that cause models to make incorrect predictions while remaining imperceptible to humans [Zhao *et al*. [2023]; Yin *et al*. [2023]]. To counter these threats, the research community has proposed a broad spectrum of defense strategies, ranging from heuristic detectors and input purification pipelines to more principled approaches like adversarial training and robustness certification [Fares *et al*. [2024]]. While detection and purification methods offer lightweight defenses in practice, they often fail against adaptive adversaries [Fares *et al*. [2024]].

Certification methods have traditionally struggled to scale to large modern architectures, limiting their applicability in real-world systems. However, the introduction of randomized smoothing has marked a turning point in this area, offering a scalable, distribution-driven approach to certifiable robustness [Cohen *et al*. [2019]]. Randomized smoothing provides this by injecting Gaussian noise at inference to include a smoothed classifer that is more stable to perturbations and admits probabilistic ℓ_2 robustness guarantees Cohen *et al*. [2019].

2.1 Adversarial Attacks and Defense Methods

According to Fares *et al*. [2024], adversarial attacks exploit vulnerabilities (VLMs) by introducing imperceptible perturbations to the input data. Due to their inherent characteristics, tasks involving images tend to amplify these vulnerabilities. To address such threats in neural networks, several defense strategies have been developed: (i) Detectors, which identify adversarial examples in otherwise clean images; (ii) Purifiers, designed to remove adversarial features from the images; and (iii) Ensemble Methods, which combine detection and purification techniques. Additionally, other defense mechanisms include (iv) Adversarial Training Methods and (v) Robustness Certification Fares *et al*. [2024]. Among the aforementioned approaches, despite their high computational cost, adversarial training and robustness certification typically offer stronger guarantees for improving resilience against adversarial attacks.

2.2 Robustness Certification

According to Cohen *et al*. [2019], robustness certification refers to methods for ensuring that the outputs of a neural network remain consistent within certain bounds, even in the presence of noise in the input data. For example, a neural network is considered provably robust if, for any input x, it is possible to guarantee that its prediction remains constant within a neighborhood around x, often defined under ℓ_2 and ℓ_∞ norms. One of the problems reported by Cohen *et al*. [2019] is that certified defenses doesn't scale to large networks, except when combined with randomized smoothing.

2.3 Randomized Smoothing

The Randomized Smoothing technique, applied by Cohen *et al*. [2019] as a robustness certification strategy, involves constructing a smoothed classifier $g(x)$ from the base classifier of a neural network $f(x)$. The classifier $g(x)$ determines, among a set of input x perturbed by Gaussian noise, which class has the highest probability of being predicted by f. In other words, $g(x)$ returns the class that f is most likely to assign, considering all the noisy versions of x. Formally, as defined by Cohen *et al*. [2019], the smoothed classifier is given by:

$$g(x) = \arg\max_{c \in Y} P(f(x+\epsilon) = c) \text{ where } \epsilon \sim \mathcal{N}(0, \sigma^2 I) \tag{1}$$

Thus, $P(f(x+\epsilon) = c)$ is the probability that the base classifier f, when given a input x perturbed by ϵ, predicts class c. Here, $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ indicates that ϵ is sampled from a normal distribution with mean 0 and variance σ^2, where σ is the noise-level hyperparameter used by the smoothed classifier g.

2.4 Robustness Guarantee

The Robustness certification through randomized smoothing is based on theoretical results reported by Cohen *et al*. [2019], which define the decision regions of the base classifier f. These regions guide the smoothed classifier f to keep the predictions with high probabilities even in the presence of noise in the input image. In our work, we rely on the guarantees established by Cohen *et al*. [2019] as a foundation, without reproducing or extending the mathematical proofs. However, we present Theorem 1 by Cohen *et al*. [2019] as an introduction.

Theorem 1 by Cohen *et al*. [2019]. Let $f : \mathbb{R}^d \to \mathcal{Y}$ be any deterministic or random function, and let $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. Let g be defined as in (1). Suppose $c_A \in \mathcal{Y}$ and $p_A, p_B \in [0, 1]$ satisfy:

$$\mathbb{P}(f(x+\epsilon) = c_A) \geq \underline{p}_A \geq \overline{p}_B \geq \max_{c \neq c_A} \mathbb{P}(f(x+\epsilon) = c) \tag{2}$$

$$\begin{aligned} &\textit{Then } g(x+\delta) = c_A \text{ for all } \|\delta\|_2 < R, \text{where} \\ &R = \frac{\sigma}{2}\left(\Phi^{-1}(p_A) - \Phi^{-1}(\overline{p}_B)\right) \end{aligned} \tag{3}$$

Assume that the base classifier f, when applied to inputs drawn from the distribution $\mathcal{N}(x; \sigma^2 I)$, predicts the most likely class c_A with probability p_A, while the second most likely class is predicted with probability p_B. the main result shows that the smoothed classifier g is certifiably robust in a neighborhood around x, specifically within an ℓ_2 ball of radius R, where Φ^{-1} denotes the inverse cumulative distribution function (CDF) of the standard Gaussian. This guarantee still holds if p_A is replaced by a lower bound $\underline{p}_A$, and p_B is replaced by an upper bound $\overline{p}_B$.

3 Related Work

This section summarizes three recent studies that address methods for VLMs aimed at increasing resilience against

adversarial attacks, including targeted black-box attacks. Table 1 summarizes their defense class, training requirements, and key notes. Section 3.1 presents an adversarial example detection method proposed by Fares *et al*. [2024], which requires no training or fine-tuning and leverages open-source diffusion models to reconstruct input images and calculate similarity with the original images, which may be clean or adversarial. Section 3.2 reviews the work of Gan et al. (2020), which focuses on large-scale adversarial training by augmenting training data with adversarial perturbations to improve representation learning quality. Section 3.3 discusses the study by Zhang and Li [2019], which proposes a novel method to enhance the adversarial robustness of the image encoder in VLMs by modifying only the learned parameters in the text modality, keeping the image encoder parameters frozen, and aligning textual representations with adversarial images.

3.1 Efficient adversarial detection

The method proposed by Fares *et al*. [2024] detects adversarial examples in VLMs by leveraging open-source diffusion models, such as UniDiffuser (Bao et al., 2023), to reconstruct images from text descriptions generated by the target VLM's image captioning task. Importantly, this method does not rely on any training to detect adversarial inputs. The approach works as follows: (i) during inference, it generates a text description from the original input image (whether clean or adversarial); (ii) it uses this generated text to produce a second image via UniDiffuser; and (iii) it employs a feature extractor to compare the similarity between the original image and the reconstructed image. Original images exhibiting low similarity to the UniDiffuser-generated images are flagged as adversarial by the method. Furthermore, Fares *et al*. [2024] evaluated their detection approach against targeted black-box attacks developed by Zhao *et al*. [2023]. According to the authors, this approach is resilient to adaptive attacks, positioning it as an excellent defense mechanism against adversarial threats.

3.2 Large-scale adversarial pretraining

Gan *et al*. [2020] proposed an adversarial data augmentation method for pre-training VLMs that enhances the quality of learned representations by adding perturbations to both image and text inputs during training. Their approach employs a generator network that produces perturbations and a discriminator network that evaluates the quality of the generated data compared to the original data. During training, the generator continuously updates its perturbation strategy to fool the discriminator, while the discriminator learns to distinguish between real and perturbed data. Through this adversarial process, the pre-trained model is exposed to diverse multimodal patterns, improving its robustness across downstream tasks. The authors argue that this adversarial data augmentation enables the model to learn from a wide variety of multimodal patterns, making it more robust for multiple subsequent tasks. In contrast, traditional pre-training methods, such as Bidirectional Encoder Representations for Transformers (BERT) Devlin *et al*. [2018], rely solely on unsupervised learning from raw text and are not specialized for vision-language tasks. Other pre-training methods incorporate images but do not employ adversarial data augmentation.

Table 1. Representative defenses for VLMs.

Work	Defense	Training	Key idea
Efficient Adversarial Defense Fares *et al*. [2024]	Detector	Inference	Detects adaptive attacks.
Large-Scale Adversarial Training Gan *et al*. [2020]	Adv. train.	Train	Improves robustness at high cost.
Adversarial Prompt Tuning Zhang *et al*. [2023]	Prompt	Fine-tune	Low-cost robustness improvement.

3.3 Adversarial prompt tuning

Zhang *et al*. [2023] aim to improve the adversarial robustness of VLM image encoders without altering the original images or requiring extensive parameter training or architectural modifications. Their method introduces a new paradigm focused on enhancing resistance to adversarial images by modifying the textual input while keeping the image encoder parameters fixed. In their approach, adversarial images are generated from existing clean images in a dataset containing paired images and texts. Unlike Gan *et al*. [2020], adversarial examples are generated prior to fine-tuning and stored in a new dataset, reducing computational cost compared to the adversarial training approach. During fine-tuning, the similarity between each adversarial image and the original clean image's associated text is computed. Since there is naturally some divergence between the adversarial image and the clean text, the method modifies the original text via backpropagation to create a textual representation aligned with the adversarial image. According to the authors, this alignment between the learned text and adversarial image increases adversarial robustness.

4 Methodology

This section details the methodology for certifying and evaluating the adversarial robustness of MiniGPT-4. The procedure integrates noise-augmented fine-tuning, randomized smoothing for statistical robustness certification, and adversarial attack evaluation using both white-box and black-box strategies. Parameter choices for noise, sampling, and computational resources are explicitly reported to support experimental rigor and facilitate future validation. The certification workflow (upper panel) and smoothed prediction procedure (lower panel) are summarized in Figure 2, and the adversarial evaluation pipeline is shown in Figure 3.

4.1 Preprocessing and Data Selection

We utilized the VQAv2 dataset, with image-question-answer annotations indexed by COCO image IDs to ensure consistent referencing throughout the pipeline. To maximize sample diversity and support robust evaluation, annotation files are subsampled using a stratified selection method across question types. This approach balances keyword, reasoning, and factual questions within each subset, mitigating language and

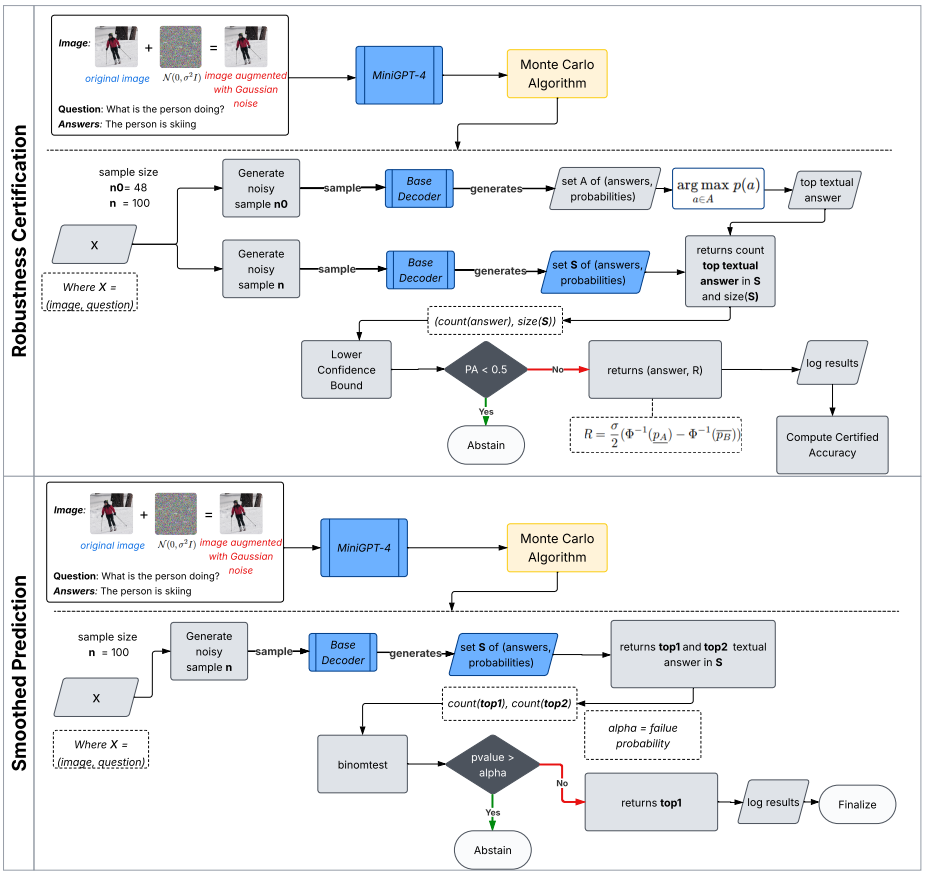

Figure 2. Workflow diagram illustrating the methodology to statistically certify the robustness of MiniGPT-4 against adversarial perturbations in image-question pairs. The upper panel ("Robustness Certification") details Gaussian noise injection, Monte Carlo sampling, answer aggregation, and certification via confidence bounds and abstain logic. The lower panel ("Smoothed Prediction") shows repeated noisy sampling and statistical hypothesis testing for robust answer selection. Color-coded blocks indicate main process steps: input handling, noise generation, base decoding, probabilistic aggregation, decision boundaries, and accuracy computation.

visual bias. The resulting samples are partitioned into distinct splits tailored to specific experimental phases: 10,000 samples are used for model fine-tuning (train split), 5,000 samples serve as the primary validation set (val split), and three additional 5,000-sample subsets (`val1`, `val2`, `val3`) are dedicated to evaluation, robustness certification, and smoothed prediction tasks, respectively.

4.1.1 Dataset

The VQAv2 dataset [Goyal *et al*. [2017]] addressed the language bias issues found in its predecessor (VQA) by introducing a balanced structure, adding complementary images for each question and allowing multiple answers. The original dataset samples are denoted as:

$$\mathcal{D} = \left\{ \left(x_i, q_i, \left\{ a_i^{(j)} \right\}_{j=1}^{10} \right) \right\}_{i=1}^{N} \tag{4}$$

where $x \in \mathcal{X} \subset \mathbb{R}^{W \times H \times 3}$ is an RGB image of width W and height H, and q_i is a language question associated with the image. Each example includes 10 ground-truth answers $\{a_i^{(j)}\}_{j=1}^{10}$, provided by human annotators. During the fine-tuning, the clean images x are augmented with Gaussian noise, resulting in perturbed inputs $(x + \epsilon)$, where $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$.

4.2 Fine-Tuning

During fine-tuning, MiniGPT-4 is trained on the selected subset using image-question-answer triplets in two stages: (i) a pre-training stage using the LAION (Schuhmann *et al*. [2021]) and CC (Changpinyo *et al*. [2021]) datasets to align vision and language, and (ii) a fine-tuning stage on a small, high-quality dataset created by the authors to enable conversational capabilities. In our strategy, we replaced the original second-stage fine-tuning with a different dataset and exposed the model to noisy images during our own fine-tuning process. According to Cohen *et al*. [2019], if a base classifier (the decoder in our case) is not exposed to noisy images during training, it will not learn how to classify (generate responses in our context) inputs sampled from $\mathcal{N}(x, \sigma^2 I)$, where x denotes a clean input.

For each image, Gaussian noise with varying standard deviations (σ) is injected to simulate adversarial perturbations and enhance the model's resilience. The image augmentation is achieved by compositing noise into the original image tensor, producing a new input that is fed into the model alongside its paired question and potential answers. The response is stored, and the loss (e.g., cross-entropy) is computed and optimized. Checkpoints are saved at each epoch for subsequent evaluation and certification

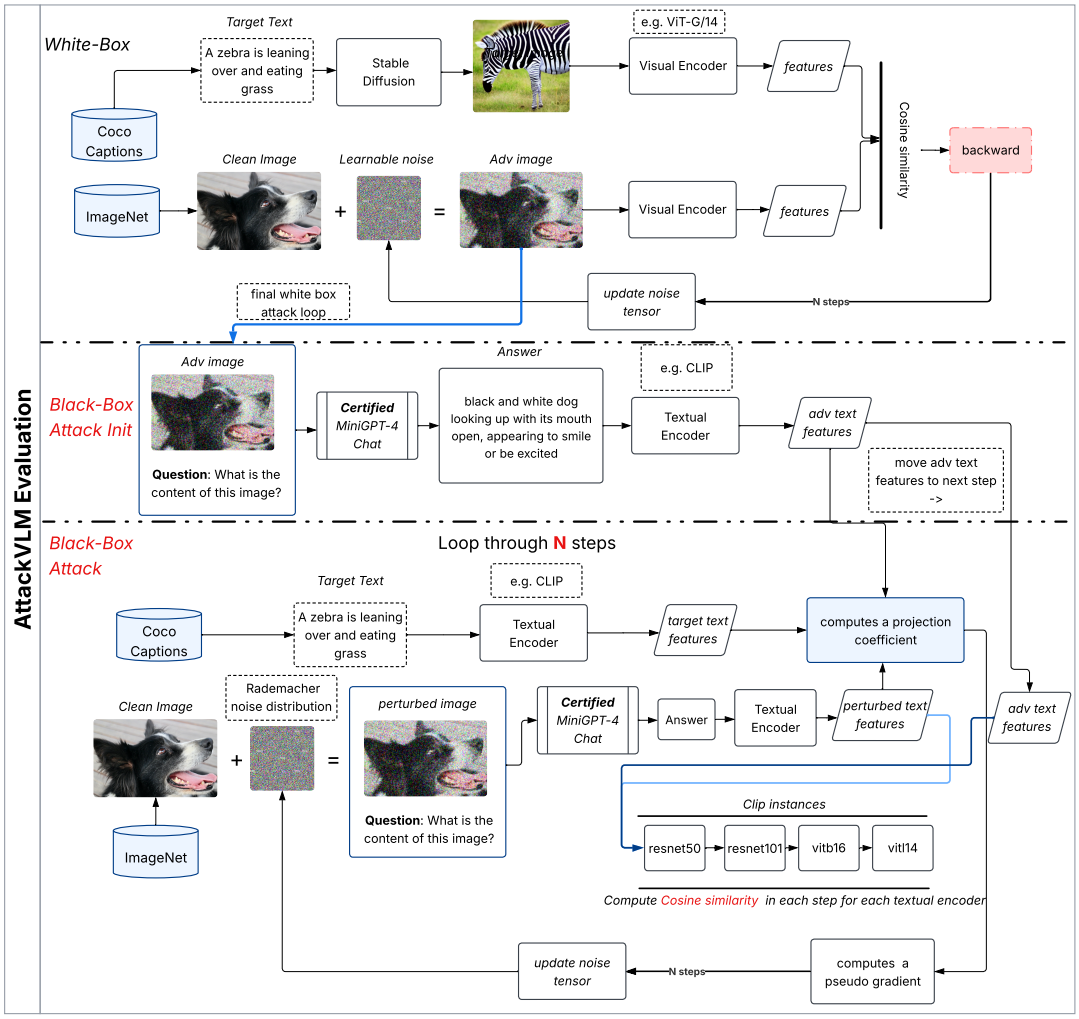

Figure 3. The diagram summarizes the three-stage adversarial evaluation protocol. The top section illustrates white-box attack initialization, where learnable noise is optimized to match the target image features using visual encoders and cosine similarity. The middle section demonstrates black-box attack initiation using adversarial images and MiniGPT-4 certification, followed by CLIP-based feature extraction. The bottom stage presents the iterative black-box attack process, where clean images are perturbed, textual and visual features are compared across multiple backbone encoders, and attack parameters are updated over multiple steps.

4.3 Robustness Certification via Randomized Smoothing

The certification procedure, illustrated in the upper panel of the Figure 2, consists of using a Monte Carlo to estimate the most frequent answer (top-1), and a lower confidence bound for its probability. If this bound exceeds the certification threshold ($p_A > 0.5$), the robustness radius is computed as: $R = \frac{\sigma}{2}\left(\Phi^{-1}(p_A) - \Phi^{-1}(p_B)\right)$ where p_B is the maximum upper confidence bound over all competitors ct_A. Otherwise, the model abstains. Certified accuracy and robustness metrics are logged for all validation samples.The complementary smoothed prediction procedure, depicted in the lower panel of Figure 2, performs repeated noisy sampling, identifies the top-2 labels, and applies a binomial test at level α to decide between returning the top label or abstaining. Certified accuracy, abstention rate, and radius statistics are logged for all evaluated samples.

According to Cohen *et al.* [2019], randomized smoothing has a known limitation when applied to neural networks. Since the base decoder f cannot reliably predict the output for a given input, it is not possible to evaluate or exactly compute the radius in which $g(x)$ prediction is certifiably robust. To address this, Cohen *et al.* [2019] proposed the use of a Monte Carlo algorithm for both prediction and certification, with both processes guaranteed to succeed with arbitrarily high probability. In our study, we applied a Monte Carlo Algorithm as shown in Algorithm 1, which is closely based on the one proposed by Cohen *et al.* [2019], with minor modifications to make it compatible with decoder rather than a classifier.

As defined in Algorithm 1, **Predict** draws n noisy samples via $\texttt{SampleUnderNoise}(f, x, n, \sigma)$, normalizes each generation, and maps outputs into $Y \cup \{\text{UNK}\}$. If a normalized generation does not match any label in Y, it is mapped to the token UNK(Unknown), ensuring every sample contributes one categorical vote. After that, it identifies the top-2 labels c_A, c_B by count, with counts n_A, n_B, and performs a binomial test on n_A successes out of $n_A + n_B$ trials against the null proportion 0.5. If the p-value $\leq \alpha$, it returns c_A; other-

Algorithm 1 Certification and Prediction

/ all generations are normalized and mapped to the discrete space $Y \cup \{UNK\}$ */*

/ evaluate g at x */*
function Predict(f, σ, x, n, α)
  $samples \leftarrow$ SampleUnderNoise(f, x, n, σ)
  $c_A, c_B \leftarrow$ top-2 labels by count in $samples$
  $n_A \leftarrow$ Count($samples, c_A$)
  $n_B \leftarrow$ Count($samples, c_B$)
  if BinomPValue($n_A, n_A + n_B, 0.5$) $\leq \alpha$ **then**
    return c_A
  else
    return ABSTAIN
  end if
end function

*/ certify robustness of g around x */*
function Certify($f, \sigma, x, n_0, n, \alpha$)
  $S_0 \leftarrow$ SampleUnderNoise(f, x, n_0, σ)
  $t_A \leftarrow$ MajorityLabel(S_0)
  $S \leftarrow$ SampleUnderNoise(f, x, n, σ)
  $n_A \leftarrow$ Count(S, t_A)
  $p_A \leftarrow$ LowerConfBound($n_A, n, 1 - \alpha$)
  $p_B \leftarrow$ UpperConfBound(Count(S, c), $n, 1 - \alpha$)
  if $t_A =$ UNK or $p_A \leq \frac{1}{2}$ **then**
    return ABSTAIN
  end if
  $R \leftarrow \frac{\sigma}{2}\left(\Phi^{-1}(p_A) - \Phi^{-1}(p_B)\right)$
  if $R \leq 0$ **then**
    return ABSTAIN
  else
    return prediction t_A and radius R
  end if
end function

wise, it returns ABSTAIN. **Certify** first draws n_0 samples to select the majority label t_A, then draws n further samples to compute the Clopper–Pearson lower bound p_A for t_A and the maximum upper bound p_B over all competitors $c \neq t_A$. It abstains if $t_A =$ UNK or $p_A \leq \frac{1}{2}$. Otherwise, it computes the certified radius, and returns ABSTAIN if $R \leq 0$; else it returns the prediction t_A and radius R. This implementation applies Gaussian noise only to images (questions are fixed) and operates on normalized generations mapped into a discrete label space.

4.4 Evaluation

Evaluation comprises three complementary protocols: *(i)* standard VQAv2 evaluation, *(ii)* randomized-smoothing-based certification, and *(iii)* smoothed prediction analysis. Dedicated validation splits are used for each procedure (`val1` for standard evaluation, `val2` for certification, `val3` for smoothed prediction). All protocols rely on the same MiniGPT-4 conversational prompting and deterministic decoding with `max_new_tokens = 20` for randomized-smoothing-based certification, smoothed prediction analysis, and log per-sample outputs for reproducibility.

Standard VQAv2 Evaluation

For each example in the validation split, the image–question pair is forwarded through the fine-tuned MiniGPT-4 decoder to produce a single answer (no sampling). When configured, additive Gaussian noise with magnitude parameter ε is applied to the input images prior to decoding; the main configuration uses $\varepsilon = 0$ (no noise). Answers are collected and scored using the official VQAv2 API (`VQA/VQAEval`) with the provided question and annotation files. Overall VQAv2 accuracy is computed exactly as in the official evaluation, which scores an answer by human agreement over the 10 reference annotations (effectively $\min(k/3, 1)$ where k is the number of annotators matching the prediction), and aggregates results over the validation set. In multi-device runs, accuracy is averaged across devices.

Certification Evaluation

We certify robustness of the predicted label to ℓ_2-bounded perturbations via Monte Carlo sampling under a Gaussian smoothing scheme, yielding a top-1 prediction, a certified radius R, and an abstention option when statistical conditions are not met. For every k-th validation example, the smoothed decoder first performs a selection phase with n_0 samples, followed by an estimation phase with n samples to compute the top label and its certified radius at failure probability α. The configuration uses hyperparameter ε for image noise, $n_0 = 48$, $n = 250$, $\alpha = 0.001$, and batches samples for efficiency.

Because VQA references include paraphrases, a prediction is deemed correct if its cosine similarity with any normalized ground-truth answer, computed using sentence-transformer embeddings (`all-MiniLM-L6-v2`), is at least $\tau = 0.85$. This mitigates strict string-matching biases in open-ended VQA. The metrics in this step are described as follows:

- **Certified accuracy:** fraction of certified predictions that are correct under the semantic criterion.
- **Abstention rate:** proportion of examples where the smoothed classifier abstains.
- **Certified radius statistics:** summary of R (e.g., mean/median) across evaluated items, optionally stratified by abstention. Average radius is reported with caveats regarding its comparability across settings.

Smoothed Prediction Analysis

The goal is to estimate the performance of the smoothed decoder under input noise without computing certified radius, using larger Monte Carlo budgets to stabilize top-1 predictions, while applying the same semantic correctness criterion as above. For every k-th example, the smoothed decoder produces a single prediction using n samples at noise level ε and failure probability α. The configuration uses $\sigma = 0.5$, $n = 500$, $\alpha = 0.001$, and periodic checkpointing to persistent logs. The metrics in this step is described as follows:

- **Smoothed accuracy:** fraction of predictions deemed correct by the semantic similarity criterion at the configured σ.

- **Abstention rate:** proportion of ABSTAIN outputs from the smoothed predictor.

Notes on Symbols and Implementation

ε denotes the magnitude of the noise injected into the input image in all evaluation steps and analyses, whereas σ refers to the Gaussian noise level applied to the input images during fine-tuning. Nonetheless, Monte Carlo sample generation uses the σ hyperparameter as the noise level. Furhtermore, due to the high computational cost of certification and smooth prediction, we set `batch_size` = 1. Consequently, noisy samples are generated and processed internally, one at a time, during Monte Carlo sampling.

4.4.1 Adversarial Attack Evaluation

To evaluate the adversarial robustness of Certified MiniGPT-4, the AttackVLM [Zhao *et al*. [2023]] framework is adopted and integrated into the adversarial evaluation pipeline. The setup combines complementary white-box and black-box attack protocols to systematically probe resilience against multimodal adversarial threats, as shown in Figure 3.

With respect to the threat model, the boundaries of the proposed approach should be clearly stated when interpreting the reported robustness results. Although the adversarial evaluation includes both white-box surrogate optimization and targeted black-box transfer attacks, all certification guarantees provided by CertifiedGPT are restricted to image-space perturbations. The certification procedure does not provide guarantees against textual perturbations, prompt injection attacks or coordinated multimodal attacks involving both image and text. Therefore, the certified robustness results reported in this work should be interpreted exclusively within the scope of image-based adversarial perturbations

Phase-1: White-Box Attack Generation In the White-Box initialization as shown in the upper panel of the Figure 3, a textual target is used to generate a target image through Stable Diffusion Rombach *et al*. [2022], then both the target image and an clean image are used to learn a noise that maximize the similarities between clean and target image features through backpropagation with projected gradient descent (PGD) [Madry *et al*. [2017]]. Because real-world attackers typically lack gradient access to the victim, the attack proceeds via a surrogate model; in this case, a CLIP[Radford *et al*. [2021]] based encoder is used as the surrogate since MiniGPT-4's visual backbone belongs to the CLIP family, which improves transferability of adversarial perturbations. This process is described by Zhao *et al*. [2023] as Matching image-image feature (**MF-ii**)

Phase-2: Black-Box Attack Initialization As the second stage in Figure 3, the adversarial image crafted in the white-box phase is used to initialize the black-box attack. The adversarial image is then fed to MiniGPT-4 Chat to obtain a textual response. A CLIP-based encoder is employed as a surrogate to extract embeddings from the returned text, producing adversarial text features that seed the subsequent black-box optimization step toward the targeted response.

Phase-3: Black-Box Attack The third and final phase shown in the Figure 3, an iterative attack runs through n steps using the **adversarial text features**, **target text** and **original(clean) image**.

Through textual encoder (CLIP), the features of the target text is extracted. Furthermore, the clean image is initially perturbed by Rademacher noise distribution, whose noise is in the discrete space. The image is sent to the MiniGPT-4 chat with a default question(e.g., What is the content of the image?). The answer is then, used by textual encoder to generate the perturbed text features. For **adversarial text features**, different CLIP instances (e.g., resnet50, resnet101, vitb16, vitl14) are used to improve robustness of the similarity signal. The resulting similarity feedback induces a pseudo□gradient that updates the noise tensor within the specified budget; this loop repeats for N steps, progressively steering the model's responses toward the target while maintaining black□box access only. This process is described by Zhao *et al*. [2023] as Matching text-text feature (**MF-tt**)

By integrating both **white-box** and **black-box** attack modes and leveraging multiple backbone encoders for analysis, this evaluation ensures a thorough and state-of-the-art assessment of adversarial robustness. The methodology provides empirical insights into the vulnerabilities and strengths of Certified MiniGPT-4 in the presence of complex multimodal perturbations, supporting quantitative and qualitative evaluation of its security posture.

4.5 Hypothesis Test

In order to determine whether there were statistically significant differences in the results, we applied ANCOVA to test the hypothesis.

5 Experiments

All experiments, except the adversarial-attack studies, use dedicated VQAv2 validation splits (val1 for standard evaluation, val2 for certification, val3 for smoothed prediction). Fine-tuning is performed with Gaussian noise at levels $\sigma \in \{0.25, 0.50, 1.00\}$; after each run, the MiniGPT-4 state is saved as a checkpoint (e.g.,`checkpoint.0.50.pth.tar`) for subsequent evaluations. Before certification, model performance after noise-aware fine-tuning is evaluated using the official VQAv2 accuracy. Certification uses Monte Carlo budgets $(n_0, n) = (48, 250)$, unless otherwise noted. For smoothed prediction, two budgets are executed, $n \in \{250, 500\}$, to analyze model outputs when the image x is perturbed by Gaussian noise. Finally, the adversarial-attack experiments load COCO captions together with a subsample of ImageNet-1k as sources for target texts and clean images.

5.1 Experiments Execution

We implemented our framework using PyTorch with XLA for TPU support during fine-tuning and evaluation. Due to the limited budget of our research project, we relied on the free TPU quotas provided by Google Cloud instead of paid compute resources. For certification and adversarial

attacks, we used an NVIDIA L4 GPU (24GB) hosted on Google Cloud. Because of financial constraints restricting extensive GPU usage, we minimized hyperparameter tuning to keep the experiments feasible within our budget.

The base architecture employed is MiniGPT-4, initialized with Vicuna as the language model and BLIP2-based vision encoders. The fine-tuning process uses 8-bit precision for memory efficiency, with all Vision Transformer (ViT) and Q-Former parameters frozen. The fine-tuning and certification, and smoothed prediction evaluations are conducted under different Gaussian noise levels ($\sigma \in \{0.25, 0.50, 1.00\}$).

Fine-tuning We fine-tuned the model on the VQAv2(train/val) dataset using 48-image batches with an image resolution of 448×448, applying standard scaling-based augmentations. For optimization, we used the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay of 0.05. The learning rate follows a linear warmup followed by cosine decay schedule, with a warmup phase of 53 steps, a maximum learning rate of 1×10^{-5}, and a minimum learning rate of 1×10^{-6}. The finetuning was conducted for a maximum of 11 epochs, with early stopping (patience = 1) based on validation performance.

Evaluation before certification On the val2 split, MiniGPT-4 is first evaluated prior to running the certification pipeline. Official VQAv2 accuracy is computed on model checkpoints fine-tuned at different noise levels, allowing comparison across states distinguished by σ.

Certification On the `val2` split, certification uses $n_0 = 48$ noisy samples for top-1 selection and $n = 250$ noisy samples for confidence estimation, yielding a probabilistic guarantee at level $1 - \alpha$. The procedure abstains when the selected class is `UNK` or when $p_A \leq 0.5$ or the resulting radius is non-positive.

Smoothed Prediction On the `val3` split, we considered $n \in \{250, 500\}$ noisy samples to stabilize the top-1 decision, and the procedure returns either the predicted label or ABSTAIN when the test is not significant or the top label is UNK.

Adversarial attacks We evaluate AttackVLM on MiniGPT-4 both without smoothing ($\sigma = 0$, original decoder) and with smoothing using checkpoints fine-tuned at $\sigma \in \{0.25, 0.50, 1.00\}$; for the latter, inference employs the smoothed-prediction wrapper when reporting attack success. The workflow follows a two-stage protocol: a white-box MF-ii initialization that optimizes a learnable perturbation with PGD on visual-feature cosine similarity, followed by a query-based black-box loop that perturbs inputs, queries MiniGPT-4 Chat, embeds answers and targets with multiple CLIP text encoders, and updates noise via a random gradient-free estimate over a small number of steps. Due to budget constraints, attack hyperparameters are intentionally conservative (e.g., n=100 samples and 2 iterations per input), so reported success rates likely underestimate what stronger settings would achieve.

Table 2. VQA performance under Gaussian noise: overall accuracy and category-wise scores (Yes/No, Number, Other) for noise levels $\sigma \in 0, 0.25, 0.5, 1.0$, showing an initial drop at 0.25 and partial recovery at higher σ.

Noise Level	Overall	Yes/No	Number	Other
0	32.39	48.17	23.44	21.70
0.25	27.84	38.44	17.71	21.70
0.5	29.04	42.26	26.56	18.82
1.0	29.86	42.74	27.60	19.87

6 Results and Discussion

This section presents the empirical findings from noise-aware fine-tuning, randomized smoothing certification, smoothed prediction, and targeted black-box attacks, and interprets their implications for VLM. Results are organized to answer research questions RQ1 and RQ2, as presented in the Section 1.2. The discussion contrasts trends across noise levels and budgets, relates robustness gains to abstention and UNK mapping, and examines where certificate-aligned resilience persists against targeted attacks versus where guarantees do not apply; limitations arising from conservative attack budgets and resource-constrained settings are also noted to contextualize conclusions.

6.1 Results concerning (RQ1-H_1)

As illustrated in Figure 4 and shown in Table 2 together show that noise-aware finetuning stabilizes optimization and modestly reshapes task performance, setting the stage for certification and attack analyses. The learning curves converge rapidly across σ, with validation loss closely tracking training loss and exhibiting the smallest gap at moderate noise, consistent with regularization effects from Gaussian perturbations during fine-tuning. Table 2 reflects the expected trade-off: accuracy dips at $\sigma = 0.25$ relative to $\sigma = 0$, then partially recovers as σ increases, with "Number" questions benefiting the most at higher noise while "Other" degrades slightly, patterns that suggest noise helps numeric robustness but can penalize open-ended phrasing.

Table 3 indicates how robustness guarantees scale with both radius and smoothing strength. At small radius (ℓ_2=0.5-1.0), the best certificates arise from the strongest noise level (σ=1.0), delivering certified top-1 accuracies of 35–33% while the corresponding natural accuracy remains 38% for that checkpoint, reflecting the expected gain in certifiable stability with moderate impact on nominal performance. At larger radius (ℓ_2=2.0-3.0), the best available guarantees come from $\sigma = 0.25$ and plateau at 25% certified accuracy despite higher standard accuracy (41%), showing that mild smoothing preserves nominal accuracy but cannot sustain large certified neighborhoods. Overall, the table quantifies the classic trade-off: increasing σ improves the certifiable radius at small to medium r, whereas excessive radius targets require accepting lower certified coverage even when clean accuracy is higher, guiding selection of σ based on the desired robustness budget.

As shown in the Table 4 At σ=0.50, increasing the Monte Carlo budget from N=250 to N=500 shifts mass from *Abstain* (0.34 $\rightarrow$ 0.30) into *Correct, Accurate* (0.37 $\rightarrow$ 0.39), indicating that additional samples mainly resolve borderline

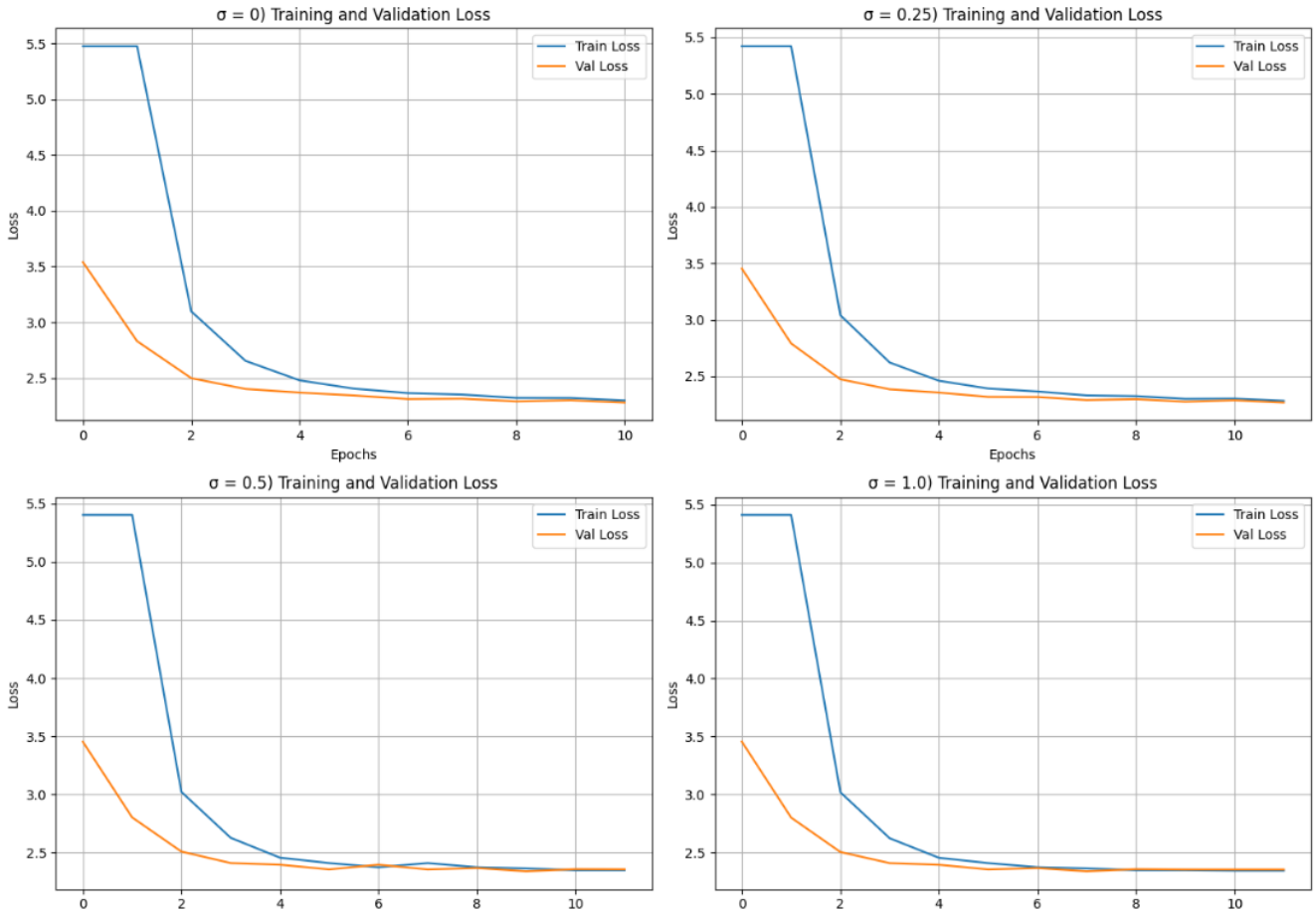

Figure 4. Regularized learning curves: training and validation loss across epochs for Gaussian noise levels $\sigma \in \{0, 0.25, 0.5, 1.0\}$ show rapid convergence and closely tracked reduction in error, indicating improved generalization with moderate noise and minimal overfitting across settings.

Table 3. Certified and standard accuracy for different ℓ_2 certification radii.

ℓ_2 Radius	Best σ	Cert. Acc. (%)	Std. Acc. (%)
0.5	**1.00***	35.0	38.0
1.0	1.00	33.0	38.0
2.0	0.25	25.0	41.0
3.0	0.25	25.0	41.0

cases in favor of confident, correct decisions. The parallel rise in *Incorrect, Accurate* (0.29 → 0.31) is modest, suggesting only a small redistribution of previously uncertain items into confidently wrong outcomes, a typical byproduct of tighter sampling variance. Notably, both "Inaccurate" categories remain at 0.00, reinforcing that the decision thresholding avoids low-confidence outputs; overall, σ=0.50 with N=500 offers a favorable balance—reduced abstention with a net gain in correct, confident predictions while keeping erroneous commitments limited.

Analyzing the results for **RQ1**, we observed that certification with randomized smoothing yields measurable gains in certified performance at modest radius; for example, at ℓ_2=0.5 with σ=100, the best certified top-1 accuracy reaches 35%, which is the closest to the correspondent standard certification accuracy(38%) among the evaluated radius, as shown in Table 3. These findings answer **RQ1** by quantifying the performance impact of certification on the VLM and support H_1 under the stated smoothing and sampling settings.

Table 4. Decomposition of smoothed predictions at $\sigma = 0.50$ as the number of Monte Carlo samples N increases from 250 to 500. The mass of "Correct, Accurate" rises modestly (0.37 → 0.39), "Incorrect, Accurate" also increases slightly (0.29 → 0.31), and the overall abstention rate decreases (0.34 → 0.30); no "Inaccurate" outcomes are observed (0.00 in both rows). This pattern indicates that enlarging N primarily converts abstentions into confident decisions, with a small redistribution toward accurate errors, consistent with reduced sampling uncertainty in randomized smoothing. .

Category	$N = 250$	$N = 500$
Correct, Accurate	0.37	0.39
Correct, Inaccurate	0.00	0.00
Incorrect, Accurate	0.29	0.31
Incorrect, Inaccurate	0.00	0.00
Abstain	0.34	0.30

6.2 Results concerning (RQ2-H_2)

This section evalutes the experimental results with respect to RQ2 to evaluate whether incorporating randomized smoothing improves the empirical robustness of a VLM against targeted black-box adversarial attacks. In addition, it analyzes the influence of the hyperparameters used during both the fine-tuning and the certification stages.

Table 5. Targeted black-box attack success rates (lower is better) for **AttackVLM** evaluated on **MiniGPT-4** fine-tuned with different Gaussian noise levels (σ) under the randomized smoothing setup. Lower values indicate greater adversarial robustness.

Backbone	$\sigma = 0$	$\sigma = 0.25$	$\sigma = 0.5$	$\sigma = 1.0$
RN50	0.585	0.583	0.579	0.579
RN101	0.562	0.554	0.551	0.552
ViT-B/16	0.595	0.587	0.583	0.581
ViT-B/32	0.620	0.623	0.620	0.620
ViT-L/14	0.470	0.454	0.449	0.448

As shown in Table 5, the application of randomized smoothing technique at each noise level modestly reduced the success rate of targeted black-box adversarial attacks, thereby improving adversarial robustness, with the clearest reductions for **ViT-L/14** and consistent but smaller reductions for **RN50**/**RN101** and **ViT-B** variants. **ViT-L/14** shows the largest decrease as σ increases, dropping from 0.470 at $\sigma = 0$ to 0.448 at $\sigma = 1.0$, indicating the greatest robustness improvement against transfer-based attacks. **RN50** and **RN101** exhibit steady, incremental decreases (e.g., RN50: 0.585 $\rightarrow$ 0.579; RN101: 0.562 $\rightarrow$ 0.551–0.552), while **ViT-B/16** declines from 0.595 to 0.581 and **ViT-B/32** remains essentially unchanged around 0.620, suggesting architecture-dependent sensitivity to smoothing. Overall, randomized smoothing yields small but consistent improvements in adversarial robustness, with the largest gains observed for **ViT-L/14** and the smallest for **ViT-B/32**, indicating that the effectiveness of randomized smoothing varies across the evaluated backbones.

6.2.1 Hypothesis Validation

To evaluate whether the observed differences in attack success rates are statistically significant, the following hypotheses, introduced in Section 1.3, are evaluated:

- $\mathbf{H_0}$ **for RQ2:** There is no statistically significant relationship between the application of randomized smoothing and the attack success rate of adversarial attacks.
- $\mathbf{H_1}$ **for RQ2:** There is a statistically significant relationship between the application of randomized smoothing and the attack success rate of adversarial attacks.

To evaluate whether the observed differences in attack success rates were statistically significant, we employed a statistical analysis based on an Ordinary Least Squares (OLS) linear model that included the backbone used by AttackVLM, the noise level, randomized smoothing, and their interactions across the evaluated experimental conditions. The hypothesis validation was based on the adversarial attack results presented in Table 5. The following variables were considered:

- **Continuous variable:** Gaussian noise level, modeled as a continuous variable and evaluated at the discrete levels $\sigma = 0.0, 0.25, 0.50$, and 1.0.;
- **Binary variable:** smoothing, treated as a categorical factor with two levels: *presence*, corresponding to the certified model using randomized smoothing with the proposed smoothing predictor, and *absense*, corresponding to the original MiniGPT-4 predictor
- **Categorical variable:** backbone, treated as a factor representing different network architectures (e.g., RN101, ViT-B/32, and ViT-L/14).
- **Continuous response variable:** attack success rate (ASR).

Important: The *backbone* variable considered in the statistical analysis refers exclusively to the CLIP backbone employed by **AttackVLM** to generate transferable adversarial perturbations. It does not refer to the MiniGPT-4 model. Therefore, the estimated backbone coefficients should be interpreted as the effect of the surrogate backbone used by the attacker on the attack success rate against the certified MiniGPT-4 model.

The estimated coefficients presented in Table 6 quantify the effects of each predictor variable (backbone and noise level) on the response variable (attack success rate). The first five rows present the intercept and the backbone coefficients under the experimental condition without Gaussian noise. Specifically, the intercept represents the estimated attack success rate for the reference backbone (RN101), whereas the remaining coefficients indicate the estimated increase or decrease in attack success rate relative to this reference.

Among the remaining surrogate backbones, RN50, ViT-B/16, and ViT-B/32 exhibited positive coefficients, indicating higher estimated attack success rates than the reference backbone. Regarding the coefficient associated with **Backbone[T.ViT-B/32]**, it indicates that, when adversarial attacks are generated using the ViT-B/32 surrogate backbone, the estimated attack success rate under the baseline condition (i.e., without Gaussian noise) is 0.0637 higher than that of the reference backbone. The only exception was the **Backbone[T.ViT-L/14]** surrogate backbone, whose coefficient (-0.0945) indicates a lower estimated attack success rate relative to the reference backbone. This result suggests greater adversarial robustness of the fine-tuned MiniGPT-4 model against attacks generated using the ViT-L/14 surrogate backbone under the baseline condition (i.e., without Gaussian noise).

The remaining rows in Table 6 represent the coefficients interaction between Gaussian noise level and the surrogate backbone on the attack success rate. The results show that the NoiseLevelCoef:Backbone[T.RN50] and NoiseLevelCoef:Backbone[T.ViT-B/32] backbones exhibit positive interaction coefficients, with 0.0026 and 0.0067 values respectively, indicating that increasing the noise level is associated with higher attack success rates. In contrast, the coefficients associated with NoiseLevelCoef:Backbone[T.ViT-B/16] and NoiseLevelCoef:Backbone[T.ViT-L/14] indicate that increasing the noise level is associated with lower attack success rates and, consequently, with improved adversarial robustness of the certified model. Although the coefficient associated with NoiseLevelCoef:Backbone[T.ViT-B/16] is negative, only NoiseLevelCoef:Backbone[T.ViT-L/14] exhibits the largest reduction in attack success rate (-0.0108) and it is statistically significant ($p < 0.001$), indicating that increasing the noise level provides the greatest improvement in the adversarial robustness of the certified MiniGPT-4 against attacks generated using the ViT-L/14 surrogate backbone. For the remaining surrogate backbones, increasing the noise

Table 6. Estimated coefficient statistics from the linear model. Positive coefficients are associated with higher attack success rates (greater vulnerability), whereas negative coefficients indicate lower attack success rates as the noise level increases.

	Coef.	$P > \|z\|$
Backbone[T.RN101] (Intercept)	0.5581	0.000
Backbone[T.RN50]	0.0255	0.000
Backbone[T.ViT-B/16]	0.0358	0.000
Backbone[T.ViT-B/32]	0.0637	0.000
Backbone[T.ViT-L/14]	**-0.0945**	**0.000**
NoiseLevelCoef:Backbone[T.RN101]	-0.0080	0.000
NoiseLevelCoef:Backbone[T.RN50]	0.0026	0.136
NoiseLevelCoef:Backbone[T.ViT-B/16]	-0.0066	0.008
NoiseLevelCoef:Backbone[T.ViT-B/32]	0.0067	0.000
NoiseLevelCoef:Backbone[T.ViT-L/14]	**-0.0108**	**0.000**

Table 7. Estimated coefficient statistics from the linear model. Positive coefficients are associated with higher attack success rates (greater vulnerability), whereas negative coefficients, such as that observed for the ViT-L/14 backbone, indicate lower attack success rates under randomized smoothing.

	Coef.	$P > \|z\|$
Backbone[T.RN101] (Intercept)	0.5613	0.000
Backbone[T.RN50]	0.0222	0.000
Backbone[T.ViT-B/16]	0.0357	0.000
Backbone[T.ViT-B/32]	**0.0590**	**0.000**
Backbone[T.ViT-L/14]	-0.0915	0.000
SmoothingCoef:Backbone[T.RN101]	-0.0090	0.000
SmoothingCoef:Backbone[T.RN50]	0.0059	0.005
SmoothingCoef:Backbone[T.ViT-B/16]	-0.0037	0.019
SmoothingCoef:Backbone[T.ViT-B/32]	0.0101	0.000
SmoothingCoef:Backbone[T.ViT-L/14]	**-0.0103**	**0.000**

level did not improve the adversarial robustness of the victim model (MiniGPT-4).

The analysis of the effects of randomized smoothing and surrogate backbone architecture, as presented in Table 7, follows the same rationale as the analysis based on Gaussian noise levels. The first five rows present the estimated coefficients associated with the surrogate backbones employed by AttackVLM to attack the MiniGPT-4 model without randomized smoothing. The remaining rows correspond to the interaction coefficients between randomized smoothing and the surrogate backbones. As in the previous analysis, the coefficient associated with **Backbone[T.ViT-B/32]** indicates the highest estimated attack success rate under the baseline condition (i.e., without randomized smoothing), whereas the coefficient associated with **SmoothingCoef:Backbone[T.ViT-L/14]** indicates lower attack success rates and, consequently, improved adversarial robustness of the certified MiniGPT-4 model. Both coefficients are statistically significant ($p < 0.001$). For the remaining surrogate backbones, randomized smoothing did not improve the adversarial robustness of the MiniGPT-4 model.

In summary, the results presented and discussed in this subsection provide sufficient empirical evidence to reject the null hypothesis (H_0) and support the alternative hypothesis (H_1) for RQ2, indicating a statistically significant result in the attack success rate for at least one backbone architecture. Although the attacks were performed using reduced hyperparameter settings due to limited access to high-performance GPUs, the results indicate that both the introduction of Gaussian noise and the application of randomized smoothing contribute to mitigating the success of adversarial attacks.

7 Conclusion

In this study, we introduced CertifiedGPT, a certified robustness framework based on a randomized smoothing technique that builds upon the contribution of Cohen *et al*. [2019]. The procedure operates over a finite, normalized answer space $Y \cup \{\text{UNK}\}$ and computes the certified radius from the estimated bounds on p_A and p_B. Within this discrete evaluation, we observed reduced targeted black-box attack success for some backbones (e.g., ViT-L/14) and noise levels, conditioned on non-abstention and nonzero certified radius R. We reported attack success both in the discrete label space aligned with certification, and we recommend stratifying results by R bins and reporting coverage and UNK rates for complete context.

Rather than proposing a new certification theory, this work demonstrates how the randomized smoothing can be adapted and empirically evaluated in the context of VLMs and VQA tasks. This adaptation represents a practical step toward certifiable robustness in multimodal systems and establishes a foundation for future research on robustness certification for generative VLMs.

The reported robustness improvements should be interpreted within the evaluated **threat model**, **certification as-**

sumptions, and **attack budget**. Additional studies employing stronger adversaries, larger optimization budgets, and alternative attack strategies are necessary to further assess the resilience of certified VLMs under more challenging adversarial settings.

7.1 Limitations and Future Work

Several limitations should be considered when interpreting the results of this study. First, the proposed robustness certification framework relies on mapping VLM outputs to a finite answer space, since randomized smoothing requires a finite set of prediction classes. Although this adaptation enables the application of randomized smoothing to VQA tasks, open-ended responses may be mapped to *UNK* token, increasing abstention rates. This limitation is particularly relevant for complex VQA scenarios in which semantically valid answers may not align perfectly with predefined answer classes. In such cases, the discretization required for certification may sacrifice part of the semantic richness of generative outputs in exchange for formal robustness guarantees. In addition, high abstention rates may introduce an availability risk. An adversary could potentially exploit conditions that increase the likelihood of *UNK* outputs, causing the model to abstain repeatedly rather than provide an answer. Future work should investigate more advanced robustness certification strategies for complex VQA tasks, such as those requiring visual reasoning.

A second limitation concerns the threat model considered in this work. The proposed certification framework assumes Gaussian noise applied exclusively to the image modality. Consequently, the certified guarantees do not extend to textual perturbations or coordinated multimodal attacks involving both image and text. While this assumption is consistent with the attack setting evaluated in this study, it significantly narrows the real-world applicability of the reported robusteness guarantees, since modern VLMs may be vulnerable through both visual and textual modalities. Therefore, the robustness certificates presented here should be interpreted strictly within the image-space threat model considered in this work. Future work should explore certification approaches capable of handling multimodal perturbations and alternative noise distributions.

The experimental evaluation was also constrained by computational resources. Restricted access to high-performance GPUs and computational resource limitations affected both fine-tuning, certification and smoothed prediction evaluation. PyTorch XLA was not fully compatible with some multimodal components and self-attention layers, requiring code migration and additional engineering effort. Furthermore, fine-tuning and adversarial evaluation had to be conducted using different computational infrastructures, which limited broader hyperparameter exploration, more extensive attack configurations, and broader experimental settings. Future work should investigate larger-scale experimental settings enabled by improved computational resources, allowing broader hyperparameter exploration, more extensive attack evaluations, and larger certification studies.

Dataset size and Monte Carlo sampling constraints represent an additional limitation. Fine-tuning was performed using a maximum of 10k samples, while evaluation was limited to approximately 5k samples. In addition, the computational cost of Randomized Smoothing restricted the number of Monte Carlo samples (n_0 and n) used during certification. These constraints may affect both the statistical confidence of the estimated certified radii and the extent to which the reported robustness results generalize beyond the evaluated subsets. Moreover, the relatively small Monte Carlo sample sizes used during certification may influence the tightness of the reported certificates.

Larger sampling configurations could produce different estimates of certified accuracy and certified radii by reducing estimation uncertainty and providing more stable probability bounds. According to Cohen *et al*. [2019], larger Monte Carlo sample sizes generally produce tighter confidence bounds, potentially resulting in larger certified radii. Therefore, the robustness guarantees reported in this study should be interpreted in light of the computational constraints that shaped the certification procedure. Consequently, the reported robustness estimates provide evidence within the evaluated sample rather than a complete characterization of the model's behavior across the full VQAv2 distribution. Future work should investigate larger-scale evaluations using more extensive datasets and larger Monte Carlo sample sizes.

Finally, adversarial attacks were executed using conservative hyperparameter configurations due to computational limitations. As a result, the reported robustness improvements should be interpreted within the evaluated threat model and attack budget. Stronger adversaries, larger query budgets, additional optimization iterations, or alternative attack strategies may lead to different attack success rates and potentially reduce the effectiveness of the proposed defense. Consequently, the robustness gains reported in this study should not be interpreted as guarantees against all targeted adversarial attacks, and the defense may be bypassed under more aggressive attack settings. Future work should therefore assess the proposed framework under stronger attack configurations, larger optimization budgets, and alternative adversarial strategies.

Despite these limitations, the results indicate that adapting Randomized Smoothing to VLMs is a feasible direction for improving robustness against image-based adversarial perturbations. The primary contribution of this work lies in the adaptation and empirical evaluation of certified robustness techniques within the VLM and VQA domains rather than in proposing a new certification theory. Future work includes extending the framework to additional VLM architectures, more recent multimodal large language models, broader threat models, and larger-scale certification studies. Code and normalization resources will be released to support reproducibility.

Declarations

Authors' Contributions

Leonardo Souza: conceived the study, designed the methodology, implemented the software, conducted the experiments, and performed the validation, formal analysis, data curation, and visual-

ization. **Leonardo Souza** drafted the original manuscript. **Pedro Nuno de Souza Moura** and **Maíra Athanázio de Cerqueira Gatti**: Conceptualization, Methodology, Supervision, Writing – Review Editing. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We thank Google TRC (TPU Resource Cloud) for discussions and infrastructure support. Compute resources included TPUs for fine-tuning and GPUs (NVIDIA L4) for AttackVLM experiments.

Funding

This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Availability of data and materials

Public datasets used (COCO/VQAv2). Trained weights (if shared) and full code for certification, normalization/canonization, and evaluation scripts will be made avail-able upon request.

References

Changpinyo, S., Sharma, P., Ding, N., and Soricut, R. (2021). Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568. DOI: 10.1109/cvpr46437.2021.00356.

Cohen, J., Rosenfeld, E., and Kolter, Z. (2019). Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320. PMLR. DOI: 10.48550/arxiv.1902.02918.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. DOI: 10.48550/arxiv.1810.04805.

Fares, S., Ziu, K., Aremu, T., Durasov, N., Takáč, M., Fua, P., Nandakumar, K., and Laptev, I. (2024). Mirrorcheck: Efficient adversarial defense for vision-language models. *arXiv preprint arXiv:2406.09250*. DOI: 10.48550/arxiv.2406.09250.

Gan, Z., Chen, Y.-C., Li, L., Zhu, C., Cheng, Y., and Liu, J. (2020). Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628. DOI: 10.48550/arxiv.2006.06195.

Gao, K., Bai, Y., Bai, J., Yang, Y., and Xia, S.-T. (2024). Adversarial robustness for visual grounding of multimodal large language models. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*. DOI: 10.48550/arxiv.2405.09981.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. (2017). Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913. DOI: 10.1007/s11263-018-1116-0.

Islam, S., Elmekki, H., Elsebai, A., Bentahar, J., Drawel, N., Rjoub, G., and Pedrycz, W. (2023). A comprehensive survey on applications of transformers for deep learning tasks. *Expert Systems with Applications*, page 122666. DOI: 10.1016/j.eswa.2023.122666.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25. DOI: 10.1145/3065386.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*. DOI: 10.48550/arxiv.1706.06083.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al*. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR. DOI: 10.48550/arxiv.2103.00020.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., *et al*. (2018). Improving language understanding by generative pre-training. Available at:`https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035`.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695. DOI: 10.1109/cvpr52688.2022.01042.

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. (2021). Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*. DOI: 10.48550/arxiv.2111.02114.

Sun, J., Wang, C., Wang, J., Zhang, Y., and Xiao, C. (2024). Safeguarding vision-language models against patched visual prompt injectors. *arXiv preprint arXiv:2405.10529*. DOI: 10.48550/arxiv.2405.10529.

Uppal, S., Bhagat, S., Hazarika, D., Majumder, N., Poria, S., Zimmermann, R., and Zadeh, A. (2022). Multimodal research in vision and language: A review of current and emerging trends. *Information Fusion*, 77:149–171. DOI: 10.1016/j.inffus.2021.07.009.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30. DOI: 10.65215/r5bs2d54.

Wang, X., Ji, Z., Ma, P., Li, Z., and Wang, S. (2023). Instructta: Instruction-tuned targeted attack for large vision-language models. *arXiv preprint arXiv:2312.01886*. DOI: 10.1186/s42400-026-00612-4.

Wei, L., Jiang, Z., Huang, W., and Sun, L. (2023). Instructiongpt-4: A 200-instruction paradigm for fine-

tuning minigpt-4. *arXiv preprint arXiv:2308.12067*. Available at:https://openreview.net/forum?id=DNvzCsQG1D.

Yigit, G. and Amasyali, M. F. (2023). From text to multimodal: A comprehensive survey of adversarial example generation in question answering systems. *arXiv e-prints*, pages arXiv–2312. DOI: 10.48550/arXiv.2312.16156.

Yin, Z., Ye, M., Zhang, T., Du, T., Liu, H., Cheng, J., Wang, T., and Ma, F. (2023). Vlattack: Multimodal adversarial attacks on vision-language tasks via pre-trained models. DOI: 10.52202/075280-2303.

Zhang, A., Lipton, Z. C., Li, M., and Smola, A. J. (2023). *Dive into deep learning*. Cambridge University Press. DOI: 10.48550/arxiv.2106.11342.

Zhang, J. and Li, C. (2019). Adversarial examples: Opportunities and challenges. *IEEE transactions on neural networks and learning systems*, 31(7):2578–2593. DOI: 10.1109/tnnls.2019.2933524.

Zhao, Y., Pang, T., Du, C., Yang, X., Li, C., Cheung, N.-M. M., and Lin, M. (2023). On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems*, 36:54111–54138. DOI: 10.52202/075280-2355.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. (2023). Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*. DOI: 10.48550/arxiv.2304.10592.

A Additional Tables

This section presents supplementary tables that analyze smoothed prediction behavior under varying Monte Carlo budgets and noise levels. The focus is on how increasing the sample size shifts mass among outcome categories—*Correct*, *Incorrect*, and *Abstain*—and how these shifts affect the confidence–coverage trade-off at $\sigma = 0.25$ and $\sigma = 1.0$. Results highlight that larger budgets primarily convert abstentions into confident decisions with only a small rise in confidently wrong outcomes, clarifying the practical operating points for randomized smoothing in VQA.

At $\sigma = 0.25$ (Table 8), increasing the budget from $N = 250$ to $N = 500$ yields a small shift from abstentions into confident outcomes, with *Abstain* dropping from 0.29 to 0.26 and *Correct, Accurate* nudging from 0.38 to 0.39. The modest rise in *Incorrect, Accurate* (0.33 $\rightarrow$ 0.35) indicates that some previously undecided cases are resolved in favor of confident but wrong predictions, a typical variance-reduction effect when more samples stabilize the vote. Since both "Inaccurate" categories remain at 0.00, the decision rule continues to filter out low-confidence outputs; overall, $\sigma = 0.25$ with $N = 500$ offers slightly better coverage while keeping the increase in confidently wrong commitments limited. According to the Table 9 At $\sigma = 1.0$, enlarging the Monte Carlo budget from $N = 250$ to $N = 500$ yields a small yet consistent shift from uncertainty to confident decisions: *Correct, Accurate* nudges up from 0.38 to 0.39, while *Abstain* drops from 0.31 to 0.27, indicating that additional samples mainly resolve borderline cases rather than changing overall behavior. The parallel rise in *Incorrect, Accurate* (0.31 $\rightarrow$ 0.34) shows a modest redistribution of formerly abstained items into confident errors, a typical effect when variance in the vote estimate is reduced; importantly, both "Inaccurate" categories remain at 0.00, so the procedure avoids low-confidence outputs even as budget grows. Overall, $\sigma = 1.0$ with $N = 500$ provides a better confidence–coverage trade-off than $N = 250$ by decreasing abstentions and slightly increasing correct commitments, while keeping the increase in confidently wrong outcomes limited.

Table 8. Decomposition of smoothed predictions at $\sigma = 0.25$ as the number of Monte Carlo samples N increases from 250 to 500. The mass of "Correct, Accurate" remains essentially stable (0.38→0.39), while "Incorrect, Accurate" slightly rises (0.33→0.35), and the overall abstention rate decreases (0.29→0.26); no "inaccurate" outcomes are observed (0.00 in both rows). This pattern indicates that enlarging N primarily converts abstentions into confident decisions, with a small redistribution toward accurate errors, aligning with the expected tightening of sampling uncertainty in randomized smoothing.

Category	$N = 250$	$N = 500$
Correct, Accurate	0.38	0.39
Correct, Inaccurate	0.00	0.00
Incorrect, Accurate	0.33	0.35
Incorrect, Inaccurate	0.00	0.00
Abstain	0.29	0.26

Table 9. Decomposition of smoothed predictions at $\sigma = 1.0$ as the number of Monte Carlo samples N increases from 250 to 500. The mass of "Correct, Accurate" increases slightly (0.38 $\rightarrow$ 0.39), "Incorrect, Accurate" also rises (0.31 $\rightarrow$ 0.34), and the overall abstention rate decreases (0.31 $\rightarrow$ 0.27); no "Inaccurate" outcomes are observed (0.00 in both rows). This pattern indicates that enlarging N primarily converts abstentions into confident decisions, with a small redistribution toward accurate errors, consistent with reduced sampling uncertainty in randomized smoothing.

Category	$N = 250$	$N = 500$
Correct, Accurate	0.38	0.39
Correct, Inaccurate	0.00	0.00
Incorrect, Accurate	0.31	0.34
Incorrect, Inaccurate	0.00	0.00
Abstain	0.31	0.27

B Adversarial Attacks

Across figures, the attack dynamics show a consistent early spike in query success followed by backbone-dependent settling. Figure 5 ($\sigma = 0$) exhibits the sharpest transient, with ViT-B/32 and ViT-B/16 stabilizing highest, RN101 converging near a mid-level plateau, and RN50 trending downward; the "adv-vitl14" curve remains lowest after a brief recovery around steps 20–40. With smoothing at $\sigma = 0.25$ in Figure 6, all models again decay over steps, but ViT-B/32 and RN50 show the most pronounced drops after 60–80 steps, RN101 plateaus slightly higher, and ViT-B/16 declines more gently, while the adversarial ViT-L/14 stays consistently below the others. Increasing smoothing to $\sigma = 0.50$ (Figure 7) maintains the same ordering and decay pattern, indicating that moderate noise dampens query success uniformly while

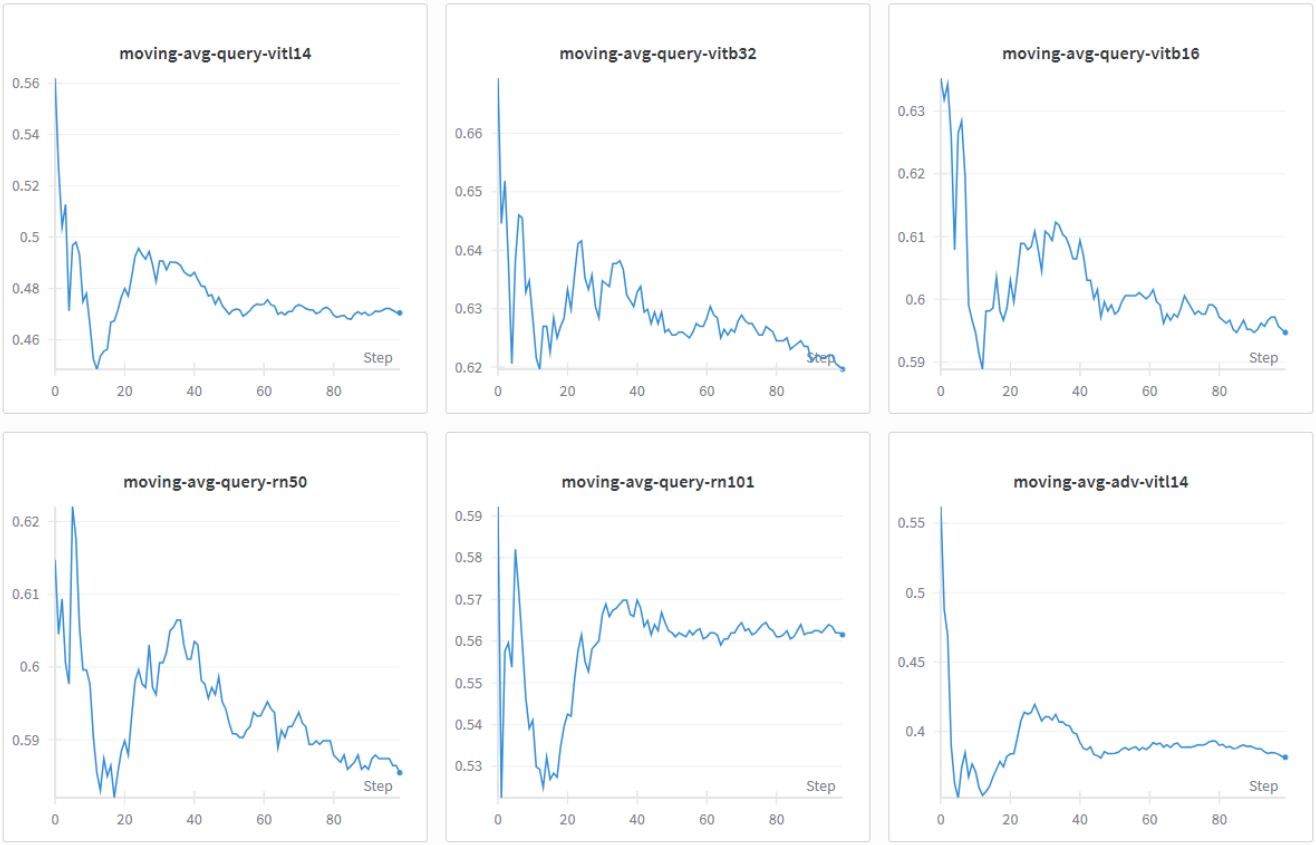

Figure 5. Moving average of query success across steps with no smoothing noise ($\sigma = 0$). The series show a sharp initial transient followed by stabilization: ViT-B/32 and ViT-B/16 maintain higher averages, RN101 stabilizes near 0.56, while RN50 shows a gradual decline over time; the "adv-vitl14" curve remains lower than the others, with a slight recovery between 20–40 steps before settling at a lower level.

preserving backbone gaps. At $\sigma = 1.0$ (Figure 8), the decay persists and the adversarial ViT-L/14 curve remains the lowest trajectory throughout, reinforcing its heightened sensitivity under the smoothed regime. The aggregate comparison in Figure 9 confirms these trends across $\sigma \in \{0, 0.25, 1.0\}$: ViT-B/32 and ViT-B/16 sustain the highest moving-average success and degrade modestly as σ increases, RN50 declines more steadily, RN101 stabilizes mid-range, and the adversarial ViT-L/14 variant consistently underperforms, highlighting architecture-dependent robustness consistent with the evaluation setup.

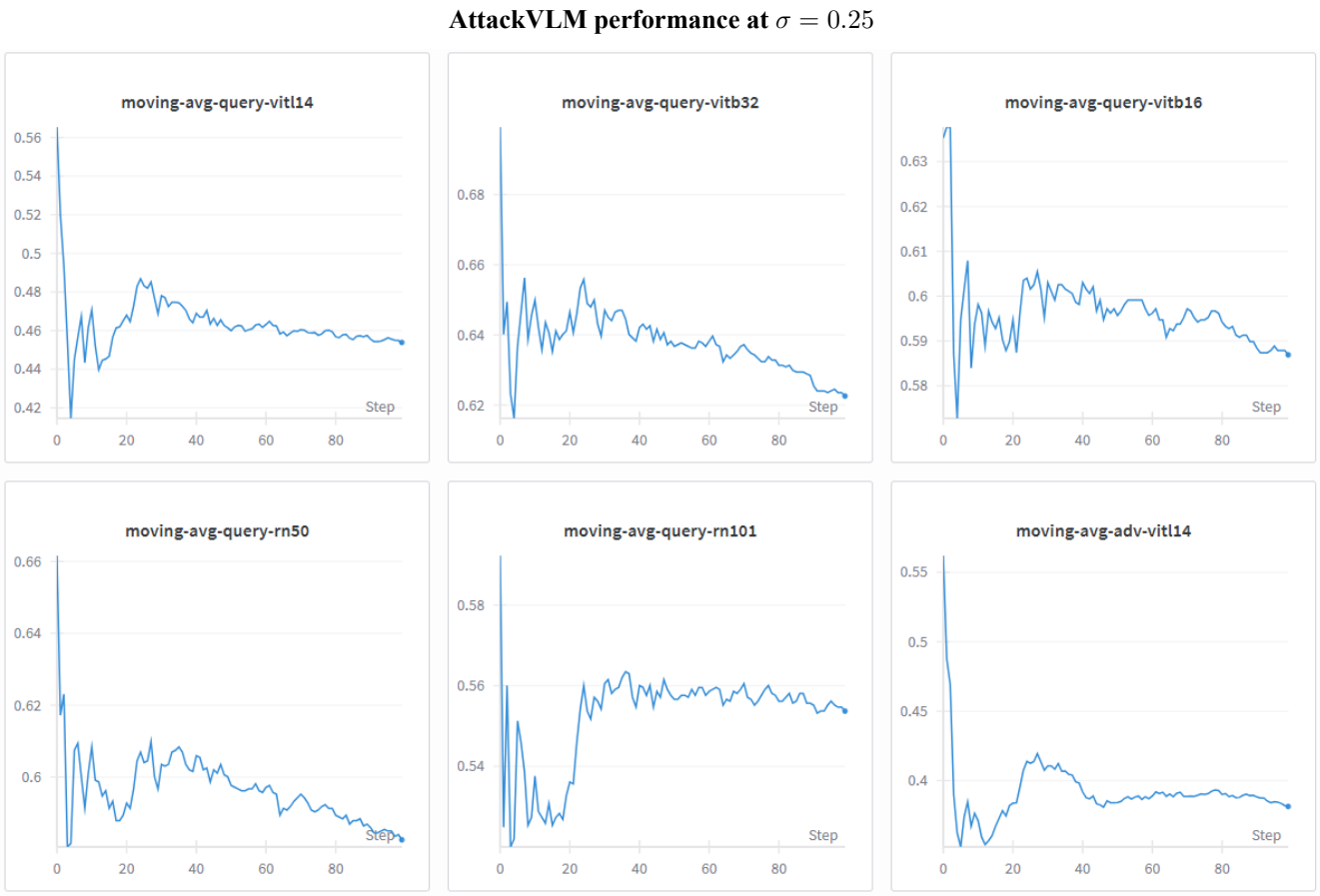

Figure 6. Adversarial attacks performance against Certified MiniGPT-4 fine-tuned with Gaussian noise $\sigma = 0.25$. The plot reports the moving-average of query success across steps under the AttackVLM protocol. All backbones exhibit an initial transient followed by a gradual decay, with ViT-B/32 and RN50 showing more pronounced drops after 60–80 steps, RN101 stabilizing at a slightly higher plateau, and the adversarial ViT-L/14 variant remaining consistently lower throughout, indicating persistent architecture-dependent sensitivity under the smoothed setting.

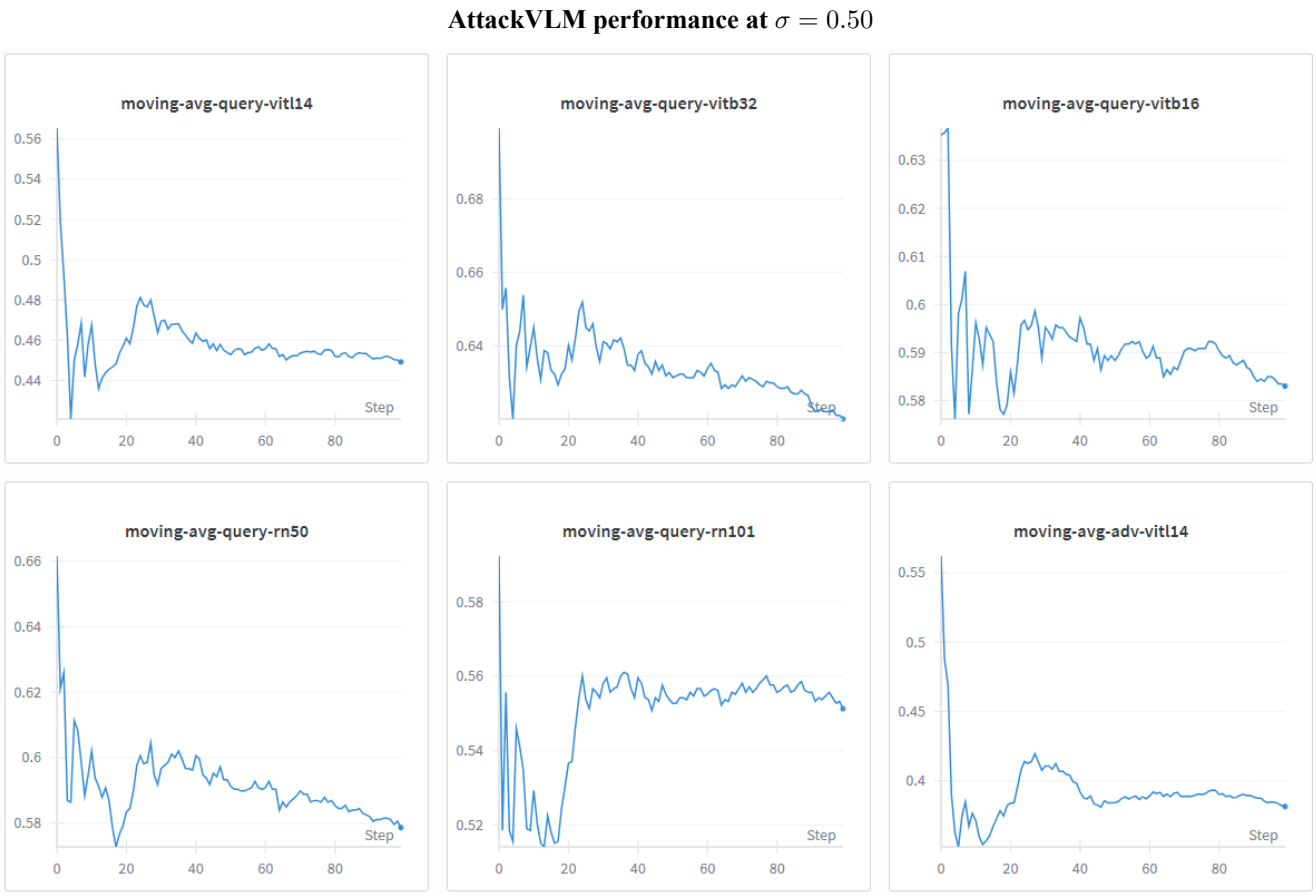

Figure 7. Adversarial attacks performance against Certified MiniGPT-4 fine-tuned with Gaussian noise $\sigma = 0.50$. The plot shows the moving-average of query success across steps under the AttackVLM protocol. All backbones exhibit an initial transient followed by a gradual decay; ViT-B/32 and RN50 display the most pronounced drops after roughly 60–80 steps, RN101 stabilizes at a slightly higher plateau, and the adversarial ViT-L/14 variant remains consistently lower throughout, indicating persistent architecture-dependent sensitivity under the smoothed setting.

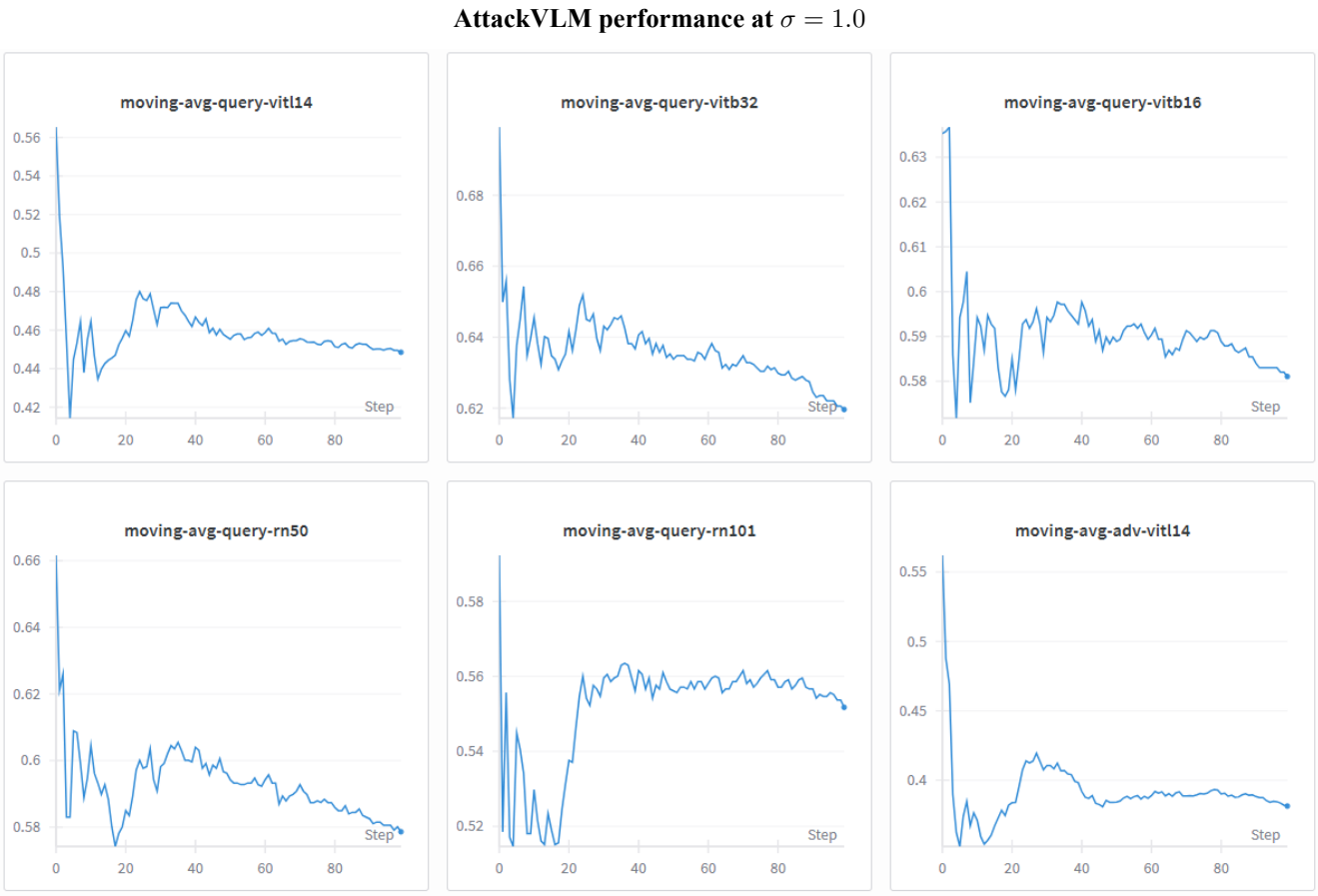

Figure 8. Adversarial attacks performance against Certified MiniGPT-4 fine-tuned with Gaussian noise $\sigma = 1.0$. The plot shows the moving-average of query success across steps under the AttackVLM protocol. All backbones exhibit an initial transient followed by a gradual decay; ViT-B/32 and RN50 display the most pronounced drops after roughly 60–80 steps, RN101 stabilizes at a slightly higher plateau, and the adversarial ViT-L/14 variant remains consistently lower throughout, indicating persistent architecture-dependent sensitivity under the smoothed setting.

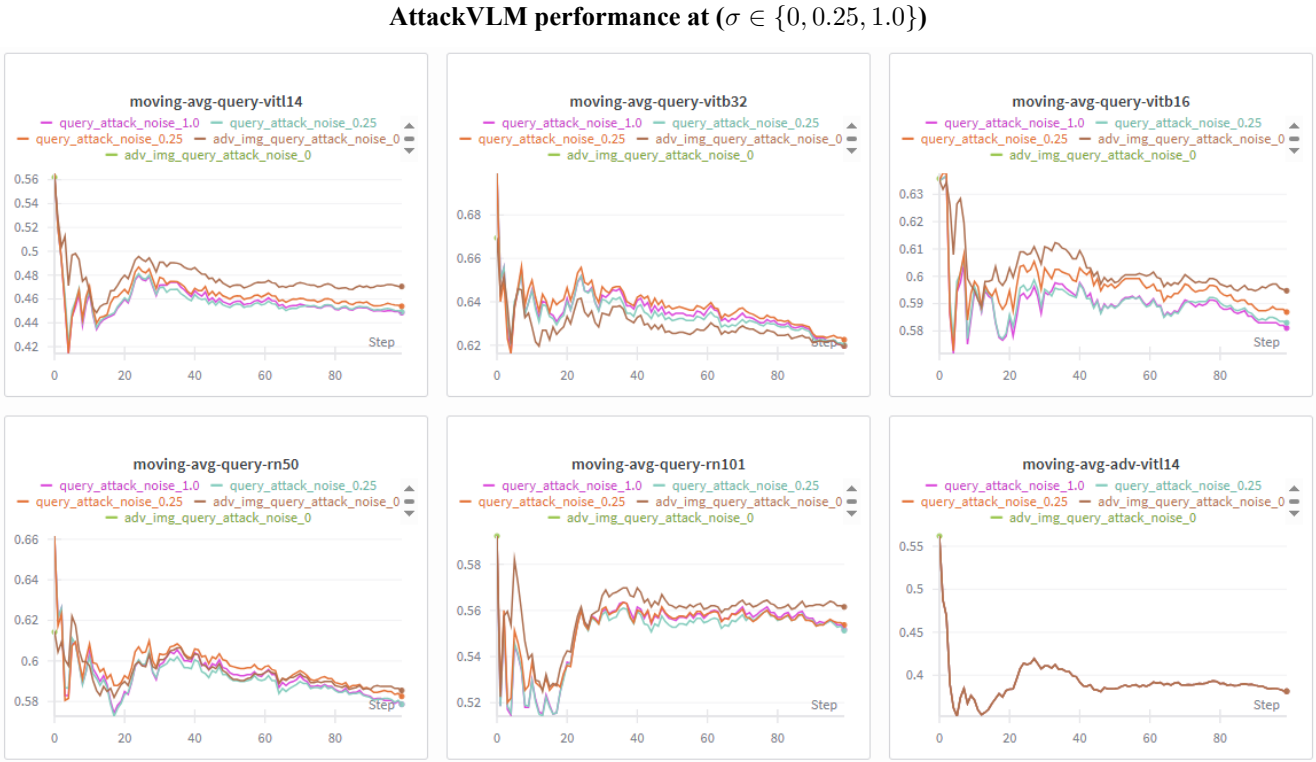

Figure 9. Comparison of moving-average query success across attack backbones (ViT-L/14, ViT-B/32, ViT-B/16, RN50, RN101, and the adversarial ViT-L/14 variant) under multiple attack noise levels ($\sigma \in \{0, 0.25, 1.0\}$). Across steps, ViT-B/32 and ViT-B/16 sustain the highest query success and show modest degradation as σ increases, while RN50 exhibits a steadier decline and RN101 stabilizes at an intermediate plateau; the adversarial ViT-L/14 remains consistently lower, indicating reduced query effectiveness across all noise regimes. The gap between backbones narrows over time but persists at convergence, evidencing architecture-dependent robustness/sensitivity to attack noise consistent with the randomized smoothing setup evaluated.