Transferable Adversarial Attacks on SAM and Its Downstream Models

Song Xia[1], Wenhan Yang[2], Yi Yu[1], Xun Lin[3], Henghui Ding[4],
Lingyu Duan[2,5], Xudong Jiang[1]*

[1]Nanyang Technological University, [2]Pengcheng Laboratory,
[3]Beihang University, [4]Fudan University, [5]Peking University
{xias0002,yuyi0010,exdjiang}@ntu.edu.sg, yangwh@pcl.ac.cn,
linxun@buaa.edu.cn, hhding@fudan.edu.cn, lingyupku@edu.cn

Abstract

The utilization of large foundational models has a dilemma: while fine-tuning downstream tasks from them holds promise for making use of the well-generalized knowledge in practical applications, their open accessibility also poses threats of adverse usage. This paper, for the first time, explores the feasibility of adversarial attacking various downstream models fine-tuned from the segment anything model (SAM), by solely utilizing the information from the open-sourced SAM. In contrast to prevailing transfer-based adversarial attacks, we demonstrate the existence of adversarial dangers even without accessing the downstream task and dataset to train a similar surrogate model. To enhance the effectiveness of the adversarial attack towards models fine-tuned on unknown datasets, we propose a universal meta-initialization (UMI) algorithm to extract the intrinsic vulnerability inherent in the foundation model, which is then utilized as the prior knowledge to guide the generation of adversarial perturbations. Moreover, by formulating the gradient difference in the attacking process between the open-sourced SAM and its fine-tuned downstream models, we theoretically demonstrate that a deviation occurs in the adversarial update direction by directly maximizing the distance of encoded feature embeddings in the open-sourced SAM. Consequently, we propose a gradient robust loss that simulates the associated uncertainty with gradient-based noise augmentation to enhance the robustness of generated adversarial examples (AEs) towards this deviation, thus improving the transferability. Extensive experiments demonstrate the effectiveness of the proposed universal meta-initialized and gradient robust adversarial attack (UMI-GRAT) toward SAMs and their downstream models. Code is available at https://github.com/xiasong0501/GRAT.

1 Introduction

Large foundation models that are trained on a broad scale of data have gained massive success in various applications [6], such as vision-language chatbot [1], text-image generation [42, 44, 46], image-grounded text generation [2, 31], and anything segmentation [26]. The segment anything model (SAM) [26], trained on vast amounts of data from the SA-1B dataset, is capable of handling diverse and complex visual tasks. The open accessibility of SAM makes it a promising foundation model, serving as the starting point for fine-tuning analytics models in certain domains and downstream applications, *e.g.,* medical segmentation [61, 54, 39], 3D object segmentation [9], camouflaged object segmentation [11], overhead image segmentation [45], and high-quality segmentation [25]. However, many studies [5, 18, 57, 38, 27, 55, 58, 66, 65, 59, 48] have highlighted the secure issues of deep learning models towards adversarial attacks. By corrupting the clean input with a finely crafted and

---

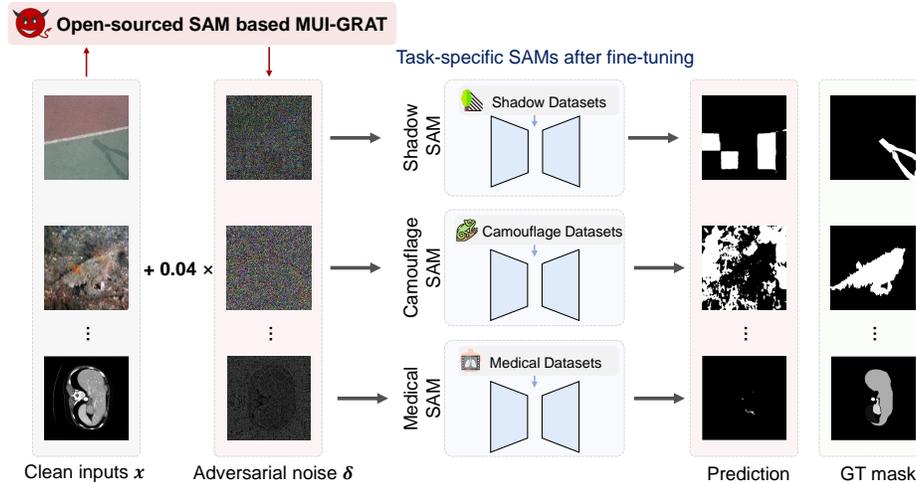*Corresponding author (exdjiang@ntu.edu.sg)

Figure 1: An illustration of UMI-GRAT towards SAM and its downstream tasks. The UMI-GRAT can mislead various downstream models by solely utilizing information from the open-sourced SAM.

nearly imperceptible adversarial perturbation, the attacker can mislead the well-trained model at a high success rate, with limited information available (*e.g.,* the surrogate model or limited queries). Consequently, significant concerns arise regarding fine-tuning open-sourced models, as it inevitably leaks critical information on downstream models, increasing their vulnerability to adversarial attacks.

Existing adversarial attacks can be roughly categorized into white-box attacks [3, 18] and black-box attacks [56, 23], based on whether the attacker can fully access the victim model. The pre-requisite of fully accessing the victim model complicates the practical deployment of white-box attacks. Conversely, transfer-based black-box attacks that require less information pose a substantial threat to real-world applications. The prevalent transfer-based black-box adversarial approaches [8, 12, 17, 22, 32, 34, 36, 15, 60] typically suppose strong prior knowledge of the victim model's task and training data, such as a 1000-class classification task in ImageNet, thereby facilitating the training of a similar surrogate model to generate potent adversarial examples (AEs). However, few studies [64, 63] consider a more practical and challenging scenario wherein the attacker is unaware of the victim model's tasks and the associated training data, due to stringent privacy and security protection policies (*e.g.,* datasets containing medical or human facial information). Moreover, the increasing size of large foundation models significantly amplifies the costs of training effective task-specific surrogate models. Thus, a more general and practical security concern is to explore the capability of attackers to mount adversarial attacks on any victim model even without the need of accessing to its downstream tasks-specific datasets to train a closely aligned surrogate model.

Given the practical security concerns of utilizing large foundation models, this paper investigates the potential risks associated with fine-tuning the open-sourced SAM on a private and encrypted dataset. We introduce a strong transfer-based adversarial attack called universal meta-initialized and gradient robust adversarial attack (UMI-GRAT), which effectively mislead SAMs and their various fine-tuned models without accessing the downstream task and training dataset, as illustrated in Figure 1.

The contributions of our paper are summarized as follows:

- We begin an investigation into a more practical while challenging adversarial attack problem: attacking various SAMs' downstream models by solely utilizing the information from the open-sourced SAMs. We provide the theoretical insights and build the experimental setting and benchmark, aiming to serve as a preliminary exploration for future research.
- We propose an offline universal meta-initialization (UMI) algorithm to extract the intrinsic vulnerability inherent in the foundation model, which is utilized as prior knowledge to enhance the effectiveness of adversarial attacks through meta-initialization.
- We theoretically formulate that, when using the open-sourced SAM as the surrogate model, a deviation occurs inevitably from the optimal direction of updating the adversary. Correspondingly, we propose a gradient robust loss to mitigate this deviation.
- Extensive experiments demonstrate the effectiveness of the proposed UMI-GRAT toward SAMs and its downstream models. Moreover, UMI-GRAT can serve as a plug-and-play strategy to significantly improve current state-of-the-art transfer-based adversarial attacks.

2 Related Work

The adversarial attacks aim to mislead the target (or victim) model by adding a small adversarial perturbation in the clean input. Existing black-box attacks can be broadly categorized into query-based and transfer-based adversarial attacks.

2.1 Query-based black box attacks

The query-based attacks consider the scenarios where the attacker does not have enough information to train a satisfying surrogate model, thus generating adversarial examples by interacting and analyzing the outputs from the victim model. This kind of attack can be divided into score-based query attacks (SQAs) [12, 24, 47] that update the AEs by observing the change of the model's prediction (*e.g.,* the logits or softmax probability) and decision-based query attacks (DQAs) [7, 10] that only rely on the model's top-1 prediction to update the AEs. However, this black-box search is naturally the NP-hard problem, and solving the optimal update strategy is non-differentiable. This makes query-based attack requires thousands of interactions with the victim model, making query-based attacks characterized by low throughput, high latency, and marked conspicuousness to attack real-world deployed systems.

2.2 Transfer-based black box attacks

The transfer-based adversarial attacks generate the AEs to mislead the victim model based on a similar surrogate model. Existing work mainly focuses on improving the transferability of AEs, which can be categorized into four groups: input-augmentation-based attacks [8, 56, 52] that enhances the effectiveness of generated AEs by augmenting the clean input (*e.g.,* using crop or rotation), optimization-based attacks [13, 35, 51, 62] that utilizes a better optimization strategy to guide the update of AEs, model modification-based attacks [53, 4] or ensemble-based attacks [19, 43, 37, 33] that enhances the AEs by utilizing a more powerful surrogate model, and feature-based attacks [32, 34] that attack the extracted feature in the intermediate layer. However, most of the work makes a strong assumption that the surrogate and victim models are optimized for the same task with identical data distribution, for example, both surrogate and victim models are optimized on the ImageNet dataset to complete the classification task. In real-world deployed systems, due to privacy and security concerns, attackers typically cannot access the training data (*e.g.,* datasets containing private information) or obtain the optimization objectives of the victim model, making training a similar surrogate model exceedingly difficult and unfeasible.

To address the challenge of deploying transfer-based black-box attacks without knowing the victim model's task and training dataset, this paper investigates the feasibility of attacking any victim model optimized for unknown tasks and distributions that significantly differ from the open-sourced surrogate model. We provide both theoretical and analytical evidence demonstrating that our proposed method can enhance the transferability and effectiveness of the generated AEs. Moreover, the proposed MUI-GRAT can serve as a plug-and-play adversarial generation strategy to enhance most existing transfer-based adversarial attacks for this challenging task.

3 Preliminaries

3.1 Adversarial attacks

Let f be any deep learning model and $\mathcal{L}$ be the loss function (*e.g.,*, the cross-entropy loss) that evaluates the quality of the model's prediction. Let $\mathcal{B}_\epsilon(\boldsymbol{x}) = \left\{ \boldsymbol{x}' : \|\boldsymbol{x}' - \boldsymbol{x}\|_p \leq \epsilon \right\}$ be an ℓ_p-norm ball centered at the input $\boldsymbol{x}$, where ϵ is a pre-defined perturbation bound. For each input $\boldsymbol{x}$, the untargeted adversarial attacks aim to find an adversarial perturbation $\boldsymbol{\delta}$ by solving:

$$\max_{\boldsymbol{x}+\boldsymbol{\delta}\in\mathcal{B}_\epsilon(\boldsymbol{x})} \mathcal{L}\left(f\left(\boldsymbol{x}\right), f\left(\boldsymbol{x}+\boldsymbol{\delta}\right)\right). \tag{1}$$

An effective solution to Equation 1 is iteratively updating the adversarial perturbation $\boldsymbol{\delta}$ based on the gradient of the loss function, for example, the iterative fast sign gradient method (I-FGSM) [28], which iteratively updates $\boldsymbol{\delta}$ by:

$$\boldsymbol{\delta}_{t+1} = clip_{\mathcal{B}_\epsilon}\{\boldsymbol{\delta}_t + \alpha \cdot sign\left(\nabla_{\boldsymbol{\delta}_t}\mathcal{L}\left(f\left(\boldsymbol{x}\right), f\left(\boldsymbol{x}+\boldsymbol{\delta}_t\right)\right)\right)\}, \tag{2}$$

where ∇ calculates the gradient and $sign$ returns the sign (*i.e.,*-1 or +1). α is a pre-defined step size to update the adversarial perturbation. $clip$ constrains the magnitude of the perturbation by projecting δ into the boundary of the ℓ_p-norm ball $\mathcal{B}_\epsilon$.

3.2 Segment anything model

The SAM consists of three parts: an image encoder $f_{\phi_{im}}$, a lightweight prompt encoder $f_{\phi_{pt}}$, and a lightweight mask decoder $f_{\phi_{mk}}$. SAM gives the mask prediction based on the image input x and prompt input p, which is expressed as:

$$y = \mathcal{SAM}(x, p) = f_{\phi_{mk}}\left(f_{\phi_{im}}(x), f_{\phi_{pt}}(p)\right), \tag{3}$$

where $f_{\phi_{im}}$ is the image encoder that provides the fundamental understanding by converting natural images into feature embeddings and $f_{\phi_{pt}}$ extracts prompt embeddings. $f_{\phi_{mk}}$ is the mask decoder that gives the mask prediction by fusing the information from both feature and prompt embeddings.

4 Methodology

4.1 Problem formulation

Let f_{ϕ_s} denote the foundation model trained on a general dataset D, and f_{ϕ_τ} denote the victim model fine-tuned on any downstream dataset D_τ, the parameters of those two models typically satisfy that:

$$\phi_s = \arg\min_{\phi_s} \mathbb{E}_{(x,y)\sim D}\left[\mathcal{L}\left(f_{\phi_s}(x), y\right)\right]; \phi_\tau = \arg\min_{\phi_\tau} \mathbb{E}_{(x_\tau, y_\tau)\sim D_\tau}\left[\mathcal{L}_\tau\left(f_{\phi_\tau}(x_\tau), y_\tau\right)\right], \phi_\tau \overset{initia}{\longleftarrow} \phi_s \tag{4}$$

Definition 1 (Transferable adversarial attack via open-sourced SAM). *For any SAM's downstream model f_{ϕ_τ} and the clean input x_τ, without any further information on the downstream task and dataset, the attacker aims to find the adversarial perturbation δ_s such that:*

$$\max_{\delta_s \in \mathcal{B}_\epsilon} \mathcal{L}_\tau\left(f_{\phi_\tau}(x_\tau), f_{\phi_\tau}(x_\tau + \delta_s)\right) \text{ s.t. } \{\delta_s = \mathcal{AT}(f_{\phi_s}, x_\tau), Private(D_\tau)\}, \tag{5}$$

where $\mathcal{AT}$ is the adversarial attack strategy and f_{ϕ_s} is the open-sourced SAM. A solution to that is fine-tuning an optimal surrogate model $f_{\phi_s^*}$ that closely aligns with the victim model. However, this approach becomes extremely challenging, when the downstream dataset is inaccessible to the attacker. Alternatively, an effective solution is to design an optimal attack strategy $\mathcal{AT}^*$ such that:

$$\mathcal{AT}^* = \arg\max_{\mathcal{AT}} \mathbb{E}_{(x_\tau, y_\tau)\sim D_\tau}\left[\mathcal{L}_\tau\left(f_{\phi_\tau}(x_\tau), f_{\phi_\tau}(x_\tau + \mathcal{AT}^*(f_{\phi_s}, x_\tau))\right)\right]. \tag{6}$$

Notably, f_{ϕ_s} and f_{ϕ_τ} are optimized on two distinctive distributions D and D_τ with losses $\mathcal{L}$ and $\mathcal{L}_\tau$, leading to a significant input-output mapping gap and gradient disparity, such as $Cosine\_similarity(\nabla f_{\phi_s}(x_\tau), \nabla f_{\phi_\tau}(x_\tau)) \ll 1$. This misalignment critically undermines the effectiveness of current gradient-based adversarial attack strategies.

Further analysis on attacking SAMs. The standard operation to deploy SAM on downstream tasks τ involves fine-tuning the image encoder $f_{\phi_{im}}$ to inherit some well-generalized knowledge. Concurrently, the lightweight prompt encoder $f_{\phi_{pt}}$ and the mask decoder $f_{\phi_{mk}}$ are trained from scratch to better accommodate the task. Considering the pivotal importance of feature embeddings and the substantial variation caused by full retraining, an intuitive approach to generate effective adversarial perturbation δ_s is utilizing the common information in $f_{\phi_{im}}$ to achieve:

$$\max_{\delta_s \in \mathcal{B}_\epsilon} \mathcal{L}\left(f_{\phi_{im}^\tau}(x_\tau), f_{\phi_{im}^\tau}(x_\tau + \delta_s)\right) \text{ s.t. } \delta_s = \mathcal{AT}^*(f_{\phi_{im}}, x_\tau), \tag{7}$$

where ϕ_{im}^τ denotes the updated parameters for the downstream model's image encoder after fine-tuning. Unless otherwise specified, we denote ϕ_{im} as ϕ in our subsequent content and take the general SAM image encoder f_ϕ as the surrogate model f_{ϕ_s}. We aim to generate the transfer-based AEs to attack any fine-tuned image encoder f_{ϕ_τ} on task τ, thereby misleading the entire prediction.

4.2 Extract the intrinsic vulnerability via universal meta initialization

Considering the great variation brought by fine-tuning the model on a new task τ, we aim to extract the intrinsic vulnerability of the foundational model that remains invariant after fine-tuning. Subsequently, this extracted vulnerability is leveraged as prior knowledge to initialize and enhance the adversarial

attack $\mathcal{AT}^*$. Inspired by the universal adversarial perturbation [40] that maintains effectiveness across various inputs, we propose the universal meta initialization (UMI) algorithm, which optimizes the initialization of adversarial perturbation to ensure both effectiveness and fast adaptability by meta-learning [41, 16]. We define the universal and meta-initialized perturbation δ as follows.

Definition 2 (**Universal and meta-initialized perturbation δ**). *Given the foundation model f_ϕ and its fine-tuned models f_{ϕ_τ} on downstream tasks τ, the universal and meta-initialized perturbation δ that extracts the intrinsic vulnerability ensures both effectiveness and fast adaptability, which are:*

1. ***Effectiveness (universal adversarial perturbation):*** *δ extracts the intrinsic vulnerability in the foundation model, which can mislead the f_ϕ successfully on most natural inputs x, which is:*

$$\max_{\delta \in \mathcal{B}_\epsilon} \mathbb{E}_{x \sim D} \left[\mathbb{I} \left\{ \mathcal{L} \left(f_\phi \left(x \right), f_\phi \left(x + \delta \right) \right) > \lambda \right\} \right], \tag{8}$$

   where λ is a pre-defined threshold for one successful attack, and $\mathbb{I} \{\cdot\}$ is the indicator function that returns 1 if the inside condition is satisfied, else 0.

2. ***Fast adaptability (meta-initialization):*** *for any downstream task τ with the corresponding private downstream dataset D_τ and model f_{ϕ_τ}, the attackers can maximize the loss $\mathcal{L}$ on downstream model f_{ϕ_τ} by updating the initialization δ via the surrogate model f_ϕ in t steps, which is:*

$$\max_{\delta \in \mathcal{B}_\epsilon} \mathbb{E}_{x_\tau \sim D_\tau} \left[\mathcal{L}_\tau \left(f_{\phi_\tau} \left(x_\tau \right), f_{\phi_\tau} \left(x_\tau + U^t \left(\delta \right) \right) \right) \right], \tag{9}$$

   U^t is the operation to update δ for t steps based on input x_τ, which is defined as:

$$U^t \left(\delta \right) = clip_{B_\varepsilon} \left[\delta + \sum_{j=1}^{t} \Delta \delta_j \right], \tag{10}$$

   where $\Delta \delta_{j+1} = \alpha_\tau \cdot sign \left(\nabla \mathcal{L} \left(f_\phi \left(x_\tau \right), f_\phi \left(x_\tau + U^j \left(\delta \right) \right) \right) \right)$ if we attack the surrogate model f_ϕ and update the adversarial perturbation based on the first-order gradient.

Generally, Equation 8 aims to extract the intrinsic vulnerability inherent in the model, which remains effective towards the input variation. Equation 9 guarantees that utilizing the perturbation δ as the initialization can rapidly threaten strong adversarial attacks for any downstream model. However, in Equation 9, D_τ and f_{ϕ_τ} are unknown if the attacker is precluded from the downstream dataset. An approximated solution to that involves using the dataset D that covers the distribution of most natural inputs and a general model f_ϕ that can approximately represent the expectation of f_{ϕ_τ}, which is:

$$\max_{\delta \in \mathcal{B}_\epsilon} \mathbb{E}_{x \sim D} \left[\mathcal{L} \left(f_\phi \left(x \right), f_\phi \left(x + U^t \left(\delta \right) \right) \right) \right]. \tag{11}$$

To optimize the above two objectives simultaneously, our learning aims to move towards the direction that maximizes the inner product of the gradients computed on both objectives. We utilize a first-order meta-learning algorithm called Reptile [41], which defines the noise δ update in each round as:

$$\delta = \delta + \eta \cdot \frac{1}{n} \sum_{i=1}^{n} \left(\tilde{\delta}_{\mu_i} - \delta \right), \tag{12}$$

where η is the update step size, and $\tilde{\delta}_{\mu_i} = U_{\mu_i}^t \left(\delta \right)$ is the updated perturbation on objective μ_i after optimizing t iterations. Here we set $n = 2$, corresponding to the two objectives in Equation 8 and 11. For μ_1 that aims to optimize Equation 11, we set $t = 5$ and $U_{\mu_1}^t \left(\delta \right)$ the same as U^t defined in Equation 10. For μ_2 that aims to optimize Equation 8, we set $U_{\mu_2}^t \left(\delta \right)$ as:

$$U_{\mu_2}^t \left(\delta \right) \leftarrow \arg \min_{\delta + \Delta \delta} \| \Delta \delta \|_\infty, \ \ \text{s.t.} \mathcal{L} \left(f_\phi \left(x \right), f_\phi \left(x + \delta + \Delta \delta \right) \right) > \lambda, \ \Delta \delta = \sum_{j=1}^{t} \Delta \delta_j. \tag{13}$$

Equation 13 aims to find a minimal update $\Delta \delta$ nearby δ to mislead the model f_ϕ. This can be achieved by using enough iterations t and a small but gradually increased norm-ball boundary ϵ. While finding an effective UMI requires a substantial number of inputs and iterations, this process can be conducted fully offline, thus not hindering real-time adversarial attacks.

4.3 Enhance the transferability via gradient robust loss

Besides the utilization of intrinsic weakness inherent in the foundation model to enhance the adversarial attack $\mathcal{AT}^*$, another method involves generating the adversarial perturbation that sustains robustness against the deviation arising from updates through a surrogate model that exhibits significant gradient disparity compared to the fine-tuned downstream model.
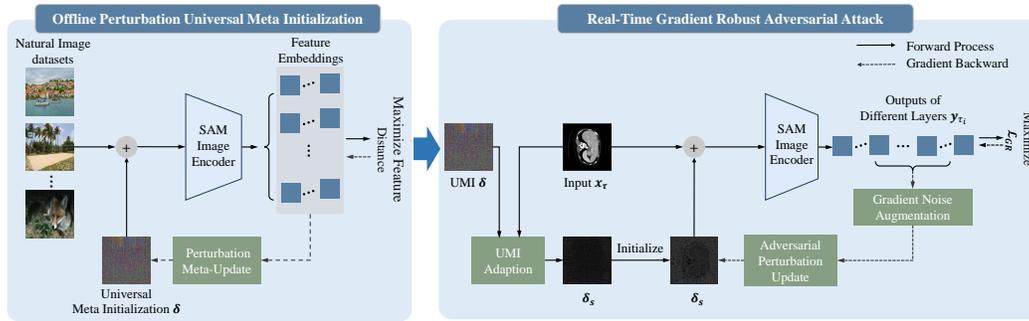
Figure 2: The data flow of our UMI-GRAT, consisting of an offline learning process of UMI and a real-time gradient robust adversarial attack.

Let us first assume that the surrogate model f_ϕ consists of m sequentially connected modules, denoted as $\{f^1_{\phi^1}, \ldots, f^m_{\phi^m}\}$. The outputs of those modules are denoted as $\{y^1, \ldots, y^m\}$, with $y^i = f^i_\phi(y^{i-1})$. For the victim model f_{ϕ_τ} with updated parameter $\Delta\phi_\tau$, the modules are denoted as $\{f^1_{\phi^1+\Delta\phi^1_\tau}, \ldots, f^m_{\phi^m+\Delta\phi^m_\tau}\}$. The output y_τ of each module with the input x_τ is denoted as:

$$y^i_\tau = f^i_{\phi^i+\Delta\phi^i_\tau}\left(y^{i-1}_\tau\right) = f^i_{\phi^i}\left(y^{i-1}_\tau\right) + h^i_{\Delta\phi^i_\tau}\left(y^{i-1}_\tau\right), \tag{14}$$

where $y^0_\tau = x_\tau$ and $h^i_{\Delta\phi^i_\tau}$ is a hypothetical function that characterizes the update brought by $\Delta\phi^i_\tau$.

Proposition 1 (Deviation in updating adversarial perturbation). *Let f_{ϕ_τ} be the victim model fine-tuned on any unknown task τ, the deviation in the direction of updating the adversarial perturbation by maximizing a predefined loss $\mathcal{L}$ in the surrogate model f_ϕ can be formulated as:*

$$\Delta\delta_\tau - \Delta\delta_s = \nabla\mathcal{L}(y^m_\tau) \cdot \left(\prod_{i=1}^m \left(\nabla f^i_{\phi^i}\left(y^{i-1}_\tau\right) + \nabla h^i_{\Delta\phi^i_\tau}\left(y^{i-1}_\tau\right)\right) - \prod_{i=1}^m \nabla f^i_{\phi^i}\left(y^{i-1}_\tau\right)\right). \tag{15}$$

In Equation 15, $\Delta\delta_\tau \leftarrow \nabla\mathcal{L}(y^m_\tau) \cdot \prod_{i=1}^m \left(\nabla f^i_{\phi^i}\left(y^{i-1}_\tau\right) + \nabla h^i_{\Delta\phi^i_\tau}\left(y^{i-1}_\tau\right)\right)$ is the update of adversarial perturbation if white-box attack the victim model and $\Delta\delta_s \leftarrow \nabla\mathcal{L}(y^m_\tau) \cdot \prod_{i=1}^m \nabla f^i_{\phi^i}\left(y^{i-1}_\tau\right)$ is the update of adversarial perturbation by maximizing the pre-defined $\mathcal{L}$ on feature embeddings of the surrogate model. Proposition 1 establishes by simultaneously considering the white-box scenarios for both surrogate and victim models to derive the gradient using the chain rule. It claims that $h^i_{\Delta\phi^i_\tau}$ leads to a great deviation in updating the adversarial perturbation δ_s towards the optimal solution in attacking victim model if directly maximizing the feature embedding distance of the surrogate model, thus degrading the effectiveness of the generated AEs.

Mitigate the deviation caused by gradient disparity. To enhance effectiveness of the generated adversarial perturbation δ_s under the hypothetical update $\nabla h_{\Delta\phi_\tau}$, we propose a gradient robust loss $\mathcal{L}_{GR}$, that aims to mitigate the deviation in Equation 15 by gradient-based noise augmentation. Denote $\mathcal{N}(\varepsilon; \mu, \sigma^2 I)$ as the isotropic Gaussian noise with mean μ and variance σ^2, which has the same dimension as $\nabla h_{\Delta\phi_\tau}$. The robust update of adversarial perturbation $\Delta\delta^*_s$ on the surrogate model based on the noised augmented gradient is:

$$\Delta\delta^*_s \leftarrow \nabla\mathcal{L}(y^m_\tau) \cdot \prod_{i=1}^m \left(\nabla f^i_{\phi^i}\left(y^{i-1}_\tau\right) + \varepsilon_i \cdot \nabla f^i_{\phi^i}\left(y^{i-1}_\tau\right)\right). \tag{16}$$

By ignoring higher-order uncertain terms in Equation 16, we can simplify it as:

$$\Delta\delta^*_s \leftarrow \nabla\mathcal{L}(y^m_\tau) \cdot \left(\prod_{i=1}^m \nabla f^i_{\phi^i}\left(y^{i-1}_\tau\right) + \sum_{i=1}^m \varepsilon_i \cdot \prod_{j=1}^{i-1} \nabla f^j_{\phi^j}\left(y^{j-1}_\tau\right)\right). \tag{17}$$

Following the adversarial perturbation update guidance in Equation 17, the corresponding gradient robust loss $\mathcal{L}_{GR}$ is defined as :

$$\mathcal{L}_{GR} = \left\| f^m_{\phi^m}\left(y^{m_{adv}}_\tau\right) - f^m_{\phi^m}\left(y^m_\tau\right) + \frac{1}{m-1}\sum_{i=1}^{m-1} \varepsilon_i \cdot \left(f^i_{\phi^i}\left(y^{i_{adv}}_\tau\right) - f^i_{\phi^i}\left(y^i_\tau\right)\right) \right\|_p, \tag{18}$$

Algorithm 1 Generating adversarial examples by UMI-GRAT

1: **Input:** open-sourced foundation model f_ϕ, natural image dataset D, task-specific image $\boldsymbol{x}_\tau$, number of meta-iterations T_m, universal step size η, attack iterations T_a, attack step size α.

2: # Off-line learning a UMI

3: **function** UNI_META_INI(f_ϕ, D, T_m, η)

4:     **Initialize** $\boldsymbol{\delta}, \boldsymbol{\delta}_{\mu_1}, \boldsymbol{\delta}_{\mu_2}$ with $\mathbf{0}$

5:     **for** t $\leftarrow$ 1 to T_m **do**

6:       assign $\boldsymbol{\delta}_{\mu_1}, \boldsymbol{\delta}_{\mu_2} = \boldsymbol{\delta}$

7:       **for** each $\boldsymbol{x}$ in D **do**

8:         update $\tilde{\boldsymbol{\delta}}_{\mu_1}$ by $U_{\mu_1}^t(\tilde{\boldsymbol{\delta}}_{\mu_1})$ in Equation 10

9:         update $\tilde{\boldsymbol{\delta}}_{\mu_2}$ by $U_{\mu_2}^t(\tilde{\boldsymbol{\delta}}_{\mu_2})$ in Equation 13

10:    update $\boldsymbol{\delta} = \boldsymbol{\delta} + \eta \cdot \frac{1}{2} \sum_{i=1}^{2} (\tilde{\boldsymbol{\delta}}_{\mu_i} - \boldsymbol{\delta})$

11:    **return** $\boldsymbol{\delta}$

12: # Real-time attack by UMI-GRAT

13: **function** GR_ATTACK(f_ϕ, $\boldsymbol{x}_\tau$, T, η, $\boldsymbol{\delta}$)

14:    # adapt the universal perturbation $\boldsymbol{\delta}$

15:    $\boldsymbol{\delta}_{adp} = \alpha_{adp} \cdot sign\left(\nabla \mathcal{L}_{adp}\left(f_\phi\left(\boldsymbol{x}_\tau\right), \tilde{\boldsymbol{y}}\right)\right)$

16:    **Initialize** $\boldsymbol{\delta}_s^* = clip_{B_\varepsilon}\left[\boldsymbol{\delta} + \boldsymbol{\delta}_{adp}\right]$

17:    # threaten GRAT

18:    **for** t $\leftarrow$ 1 to T_a **do**

19:       $\boldsymbol{g}_s \leftarrow \nabla \mathcal{L}_{GR}\left(f_\phi\left(\boldsymbol{x}_\tau + \boldsymbol{\delta}_s^*\right), f_\phi\left(\boldsymbol{x}_\tau\right)\right)$

20:       $\boldsymbol{\delta}_s^* = clip_{B_\varepsilon}\left[\boldsymbol{\delta}_s^* + \alpha \cdot sign(\boldsymbol{g}_s)\right]$

21:    $\boldsymbol{x}_\tau^{adv} = \boldsymbol{x}_\tau + \boldsymbol{\delta}_s^*$

22:    **return** $\boldsymbol{x}_\tau^{adv}$

---

where $\boldsymbol{y}^{i_{adv}}$ is the extracted adversarial feature by layer i and $\|\cdot\|_p$ is a predefined norm-based measure that is decided by the $\mathcal{L}$ (*e.g.*, $p = 1$ for L1 loss).

Discussion with the intermediate-level attacks. The intermediate-level attacks (ILAs) [34, 22, 32] also aim to maximize the dissimilarity of feature embeddings between the clean and adversarial inputs. However, the main concern in ILAs is how to find a directional vector $\boldsymbol{v}$ to guide the update direction of $f_\phi(\boldsymbol{x}) - f_\phi(\boldsymbol{x}^{adv})$, thus assuring that this feature-wised dissimilarity can maximally mislead the final prediction. Different from that, our $\mathcal{L}_{GR}$ considers the problem one step further: given an optimal directional vector $\boldsymbol{v}$, how to generate

Figure 3: The cosine similarity of white-box generated perturbations on surrogate and victim models.

the adversarial perturbation that is roust towards the potential gradient variation in the victim model, thus maximally misleading the victim model along the direction of $\boldsymbol{v}$. Our core idea is hence in parallel with ILAs and can be well combined with them to enhance the attack's effectiveness. We experimentally analyze and visualize the cosine similarity of the generated perturbations on CT-scan images by white-box attacking open-sourced SAM and medical SAM using MI-FGSM [13], ILPD [34], and our gradient robust attack in Figure 3, illustrating that the proposed $\mathcal{L}_{GR}$ effectively reduce the deviation caused by gradient variation and achieves a much transferability.
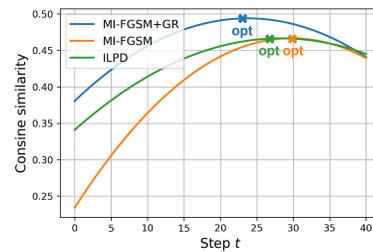
5 Implementation of the proposed MUI-GRAT

The detailed implementation of our proposed MUI-GRAT is illustrated in Algorithm 1 and Figure 2. Our UMI-GRAT method consists of two stages. The initial stage involves the offline learning of a universal meta-initialization (UMI), which aims to find the intrinsic vulnerability inherent in the foundation model. In the subsequent stage, we utilize the learned UMI as the prior knowledge to enhance the gradient-variation robust adversarial attack.

We utilize the image encoder from Vit-H-based SAM as the general foundation model f_ϕ for generating the UMI. The natural image dataset D consists of a total of 20,000 images, with 10,000 from ImageNet and 10,000 from the SA-1B dataset. We set the meta iterations T_m as 7 and the universal step size η as 1. The function Uni_Meta_Ini returns the learned UMI $\boldsymbol{\delta}$ that can be used to enhance the generation of subsequent input-specific adversarial perturbation.

In GR_Attack, we first adapt the calculated UMI $\boldsymbol{\delta}$ with the task-specific image $\boldsymbol{x}_\tau$ by one-step update using FGSM [18] with the step size $\alpha_{adp} = 4$. Assume that $\tilde{\boldsymbol{y}}$ represents the mean of the feature embedding of natural images calculated by f_ϕ, our $\mathcal{L}_{adp}$ is defined as:

$$\mathcal{L}_{adp} = \|mean(f_\phi(\boldsymbol{x}_\tau)) - \tilde{\boldsymbol{y}}\|_p. \tag{19}$$

Equation 19 aims to minimize the domain difference between $v x_\tau$ and the natural images by a specific generated perturbation. The UMI $\boldsymbol{\delta}$ is then added by $\boldsymbol{\delta}_{adp}$ and utilized to initialize $\boldsymbol{\delta}_s^*$. We set T_a to 10, and update the adversarial perturbation $\boldsymbol{\delta}_s^*$ by maximizing the gradient robust loss $\mathcal{L}_{GR}$.

Table 1: Comparison results of transfer-based adversarial attacks on different models. The surrogate model is the open-sourced SAM.

Model	Medical SAM [61]		Shadow-SAM [11]	Camouflaged-SAM [11]					
Dataset	CT-Scans		ISTD	COD10K		CAMO		CHAME	
Metrics	mDSC↓	mHD↑	BER↑	S_α↓	MAE↑	S_α↓	MAE↑	S_α↓	MAE↑
Without attacks	81.88	20.64	1.43	0.883	0.025	0.847	0.070	0.896	0.033
MI-FGSM [13]	40.83	64.42	4.31	0.372	0.214	0.331	0.286	0.352	0.250
DMI-FGSM [56]	34.51	74.20	4.39	0.435	0.134	0.395	0.210	0.416	0.164
PGN [17]	43.15	58.03	5.16	0.368	0.230	0.336	**0.318**	0.340	0.275
BSR [50]	25.7	94.48	5.20	0.414	0.146	0.372	0.226	0.402	0.178
ILPD [34]	33.65	65.98	4.40	0.366	0.245	**0.310**	0.316	0.327	0.287
MUI-GRAT	**5.22**	**111.87**	**12.46**	**0.360**	**0.248**	0.329	0.308	**0.332**	**0.293**
MUI-GRAT+DMI-FGSM	5.28	114.68	5.48	0.409	0.198	0.417	0.267	0.406	0.228
MUI-GRAT+PGN	9.62	115.87	**33.98**	0.358	0.262	0.353	0.306	0.332	0.296
MUI-GRAT+BSR	3.61	105.31	7.00	0.385	0.219	0.398	0.277	0.387	0.245
MUI-GRAT+ILPD	**3.52**	**121.89**	15.56	**0.349**	**0.263**	**0.321**	**0.311**	**0.315**	**0.317**

6 Experimental Results

6.1 Experiment setup

Evaluation details: we conduct experiments on SAMs' downstream models including, medical image segmentation SAM [61], shadow segmentation SAM [11], and camouflaged object segmentation SAM [11]. The datasets include: the synapse multi-organ segmentation dataset [29] that contains 3779 abdominal CT scans with 13 types of organs annotated, the ISTD dataset [49] that contains 1870 image triplets of shadow images, the COD10K dataset [14] that contains 5066 camouflaged object images, the CHAMELEON dataset that contains 76 camouflaged images, and the CAMO dataset [30], that contains 1500 camouflaged object images. We report the mean dice similarity score (mDSC) and mean Hausdorff distance (mHD) for evaluating medical segmentation, mean absolute error (MAE) and structural similarity (S_α) for camouflaged object segmentation, and the bit error rate (BER) for shadow segmentation. In medical SAM, the image encoder is based on SAM-Vit-B and fine-tuned with LoRA [21]. In shadow segmentation and camouflaged object segmentation SAM, the image encoders are based on SAM-Vit-H and fine-tuned with the adapter [20]. The decoders in those models are all fully retrained.

Compared methods: We mainly compare and evaluate our method with current transfer-based adversarial attacks including gradient-based attacks called MI-FGSM [13] and PGN [17], input-augmentation based attacks called DMI-FGSM [56] and BSR [50], and intermediate-level feature based attack called ILPD [34].

Implementation details: we use the MI-FGSM [13] as our basic attack method. For all methods reported, we set the attack update iterations T_a as 10, with the l_∞ bound $\epsilon = 10$ and the step size $\alpha = 2$. For our UMI, we set the meta iterations $T_m = 7$, universal step size $\eta = 1$. For PGN and BSR, we set the number of examples as 8 for efficiency.

6.2 Main results

We report our main results in Table 1. The first row of the data presents the model performance with clean inputs. The second part of the data shows the model performance under different adversarial attacks, where the data with the strongest attack is bold. The results demonstrate that the adversarial examples generated by our proposed MUI-GRAT are more effective and generalizable than others, consistently posing significant adversarial threats across various downstream models. In medical segmentation and shadow segmentation tasks that share a great difference with the natural segmentation tasks, our proposed MUI-GRAT greatly surpasses others (*e.g.,* MUI-GRAT reduces the mDSC from 81.88 to 5.22 while the previous best is 25.73.). This demonstrates the exceptional effectiveness of our proposed MUI-GRAT when the attacker lacks information about the victim model, thereby generating AEs using a surrogate model distinct from the victim model. In the camouflaged object segmentation task where the data closely resembles natural images, all methods exhibit strong transferability. Our MUI-GRAT achieves the best performance on the COD10K and CHAME datasets and performs comparably to the SOTA method on the CAMO dataset.
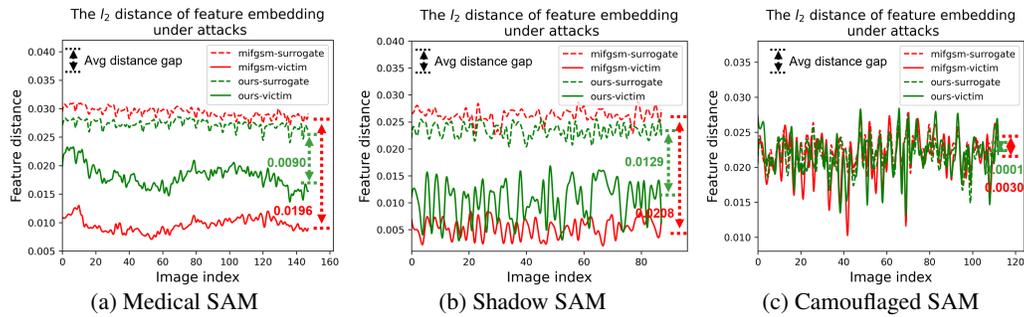
Figure 4: The l_2 distance of feature embedding from clean inputs and adversarial examples. The small distance gap between the surrogate and victim models indicates better transferability.

In the last part of Table 1, we analyze the performance of our proposed MUI-GRAT when combined with other SOTA transfer-based attacks. Notably, all methods achieve an overall performance gain after combining with ours. Though a slight performance drop occurs when attacking the CAMO dataset, combining the MUI-GRAT brings great performance gain on attacking other tasks. This demonstrates the versatility of our MUI-GRAT, which can be seamlessly applied in a plug-and-play manner to bolster existing transfer-based attacks. We report the experiment results for attacking open-sourced SAMs in Appendix A.2 and analyze the real-time attack efficiency for each method in Appendix A.3.

6.3 Analysis of the transferability

The transferability property across diverse models ensures the effectiveness of the adversarial example to mislead an unknown victim model. As discussed in Section 4.1, an intuitive objective for attacking SAM's downstream models is to maximize the dissimilarity of feature embedding extracted from the clean input x and adversarial input x^{adv}. Based on this, to numerically evaluate the improvement of transferability brought by MUI-GRAT, we present the l_2 distance of clean and adversarial feature embeddings attacked by MI-FGSM and ours and then analyze the distance gap between the surrogate and victim models. We propose that a viable transferable attack methodology should induce a substantial feature distance Δf in the victim model, while simultaneously ensuring minimal performance degradation ϵ during the transition from the surrogate to the victim model.

We show this comparison result in Figure 4, where we randomly pick a subset of inputs and show the distance of feature embedding for each clean-adversarial input pair. The average distance gap between the surrogate and victim models indicates the overall transferability of the attack method. In the medical and shadow segmentation SAM, where the data and task are distinct from the original SAM, we find a great performance drop for MI-FGSM when transferred from the surrogate model to the victim model. Though the adversarial examples generated by MI-FGSM induce a large feature distance in the surrogate model, their effect on the victim model is relatively minor. Conversely, the adversaries generated by MUI-GRAT maintain much better transferability, suffering from a small performance drop when transferred from the surrogate model to the victim model. In the camouflage object segmentation task, where the data are natural images and the segmentation objective is similar to the original SAM, both attack algorithms show good transferability (nearly no performance drop when transferred from the surrogate model to the victim models). Our MUI-GRAT shows better transferability with nearly no performance drop.

6.4 Ablation study

In this section, we explore the contribution of the proposed MUI and GRAT by integrating them with MI-FGSM and PGN attacks. The ablation results are shown in Table 2. In scenarios where the task and dataset distributions of surrogate and victim models differ markedly (e.g., medical image segmentation and natural image segmentation), we observe that the GR loss significantly enhances effectiveness. Meanwhile, across all scenarios, the proposed MUI consistently contributes to enhancing the adversarial attacks. Particularly in camouflaged object segmentation when the surrogate and victim models exhibit close similarities, the MUI yields substantial benefits.

Table 2: The performance of methods combined by MUI and GR.

Basic strategy	MUI	GR	Medical SAM [61]		Shadow-SAM [11]	Camouflaged-SAM [11]	
			mDSC↓	mHD↑	BER↑	S_α↓	MAE↑
MI-FGSM [13]	✗	✗	40.83	64.42	4.31	37.17	21.41
	✓	✗	37.54	70.13	4.72	**36.11**	**24.74**
	✗	✓	**6.34**	**100.52**	**8.07**	36.90	21.78
PGN [17]	✗	✗	43.15	58.03	5.16	36.84	23.04
	✓	✗	41.13	66.11	6.46	**35.23**	**27.68**
	✗	✓	**15.51**	**93.98**	**10.36**	36.25	24.10

This observation aligns with our analysis in Section 4. By assuming a hypothetical update $h_{\Delta\phi_\tau}$ in the victim model, the proposed gradient robust loss greatly enhances the effectiveness of the generated AEs towards the gradient variation, thus benefiting more for medical and shadow segmentation tasks. Moreover, the MUI aims to find the intrinsic vulnerability inherent in the basic foundation model through a broad general dataset, which is then provided as the prior knowledge for generating a more effective adversary. Therefore, in scenarios where the victim model inherits substantial information from the surrogate model, this prior knowledge becomes increasingly reliable and effective.

7 Conclusion

The security of utilizing large foundation models is a critical issue for deploying them in real-world applications. This paper, for the first time, considers a more challenging and practical attack scenario where the attacker executes a potent adversarial attack on SAM-based downstream models without prior knowledge of the task and data distribution. To achieve that, we propose a universal meta-initialization (UMI) algorithm to uncover the intrinsic vulnerabilities inherent in the foundation model. Moreover, by theoretically formulating the adversarial update deviation during the attacking process between the open-sourced SAM and its fine-tuned downstream models, we propose a gradient robust loss that simulates the corresponding uncertainty with gradient-based noise augmentation and analytically demonstrates that the proposed method effectively enhances the transferability. Extensive experiments validate the effectiveness of the proposed UMI-GRAT toward SAM and its downstream tasks, highlighting the vulnerabilities and potential security risks of the direct utilization and fine-tuning of open-sourced large foundation models.

Acknowledgment

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Proc. Annual Conf. Neural Information Processing Systems*, 2022.

[3] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok. Synthesizing robust adversarial examples. In *Proc. Int'l Conf. Machine Learning*, 2018.

[4] P. Benz, C. Zhang, and I. S. Kweon. Batch normalization increases adversarial vulnerability and decreases adversarial transferability: A non-robust feature perspective. In *Proc. IEEE Int'l Conf. Computer Vision*, 2021.

[5] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, 2013.

[6] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[7] W. Brendel, J. Rauber, and M. Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *Proc. Int'l Conf. Learning Representations*, 2018.

[8] J. Byun, S. Cho, M.-J. Kwon, H.-S. Kim, and C. Kim. Improving the transferability of targeted adversarial examples through object-based diverse input. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2022.

[9] J. Cen, Z. Zhou, J. Fang, W. Shen, L. Xie, D. Jiang, X. Zhang, Q. Tian, et al. Segment anything in 3d with nerfs. *Proc. Annual Conf. Neural Information Processing Systems*, 2023.

[10] J. Chen, M. I. Jordan, and M. J. Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *IEEE symposium on security and privacy*, 2020.

[11] T. Chen, L. Zhu, C. Deng, R. Cao, Y. Wang, S. Zhang, Z. Li, L. Sun, Y. Zang, and P. Mao. Sam-adapter: Adapting segment anything in underperformed scenes. In *Proc. IEEE Int'l Conf. Computer Vision*, 2023.

[12] F. Croce, M. Andriushchenko, N. D. Singh, N. Flammarion, and M. Hein. Sparse-rs: a versatile framework for query-efficient sparse black-box adversarial attacks. In *Proc. AAAI Conf. on Artificial Intelligence*, pages 6437–6445, 2022.

[13] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li. Boosting adversarial attacks with momentum. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 9185–9193, 2018.

[14] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao. Camouflaged object detection. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 2777–2787, 2020.

[15] Z. Fang, R. Wang, T. Huang, and L. Jing. Strong transferable adversarial attacks via ensembled asymptotically normal distribution learning. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2024.

[16] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proc. Int'l Conf. Machine Learning*, 2017.

[17] Z. Ge, H. Liu, W. Xiaosen, F. Shang, and Y. Liu. Boosting adversarial transferability by achieving flat local maxima. *Proc. Annual Conf. Neural Information Processing Systems*, 2023.

[18] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[19] M. Gubri, M. Cordy, M. Papadakis, Y. L. Traon, and K. Sen. Lgv: Boosting adversarial example transferability from large geometric vicinity. In *Proc. IEEE European Conf. Computer Vision*, 2022.

[20] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly. Parameter-efficient transfer learning for nlp. In *Proc. Int'l Conf. Machine Learning*, 2019.

[21] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al. Lora: Low-rank adaptation of large language models. In *Proc. Int'l Conf. Learning Representations*, 2021.

[22] Q. Huang, I. Katsman, H. He, Z. Gu, S. Belongie, and S.-N. Lim. Enhancing adversarial example transferability with an intermediate level attack. In *Proc. IEEE Int'l Conf. Computer Vision*, 2019.

[23] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin. Black-box adversarial attacks with limited queries and information. In *Proc. Int'l Conf. Machine Learning*, 2018.

[24] A. Ilyas, L. Engstrom, and A. Madry. Prior convictions: Black-box adversarial attacks with bandits and priors. In *Proc. Int'l Conf. Learning Representations*, 2019.

[25] L. Ke, M. Ye, M. Danelljan, Y.-W. Tai, C.-K. Tang, F. Yu, et al. Segment anything in high quality. *Proc. Annual Conf. Neural Information Processing Systems*, 2024.

[26] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *Proc. IEEE Int'l Conf. Computer Vision*, 2023.

[27] C. Kong, A. Luo, S. Wang, H. Li, A. Rocha, and A. C. Kot. Pixel-inconsistency modeling for image manipulation localization. *arXiv preprint arXiv:2310.00234*, 2023.

[28] A. Kurakin, I. J. Goodfellow, and S. Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*. 2018.

[29] B. Landman, Z. Xu, J. Igelsias, M. Styner, T. Langerak, and A. Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, 2015.

[30] T.-N. Le, T. V. Nguyen, Z. Nie, M.-T. Tran, and A. Sugimoto. Anabranch network for camouflaged object segmentation. *Computer vision and image understanding*, 2019.

[31] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proc. Int'l Conf. Machine Learning*, 2023.

[32] Q. Li, Y. Guo, and H. Chen. Yet another intermediate-level attack. In *Proc. IEEE European Conf. Computer Vision*, 2020.

[33] Q. Li, Y. Guo, W. Zuo, and H. Chen. Making substitute models more bayesian can enhance transferability of adversarial examples. In *Proc. Int'l Conf. Learning Representations*, 2022.

[34] Q. Li, Y. Guo, W. Zuo, and H. Chen. Improving adversarial transferability via intermediate-level perturbation decay. *Proc. Annual Conf. Neural Information Processing Systems*, 2023.

[35] J. Lin, C. Song, K. He, L. Wang, and J. E. Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. *arXiv preprint arXiv:1908.06281*, 2019.

[36] Q. Lin, C. Luo, Z. Niu, X. He, W. Xie, Y. Hou, L. Shen, and S. Song. Boosting adversarial transferability across model genus by deformation-constrained warping. In *Proc. AAAI Conf. on Artificial Intelligence*, 2024.

[37] Y. Liu, X. Chen, C. Liu, and D. Song. Delving into transferable adversarial examples and black-box attacks. In *Proc. Int'l Conf. Learning Representations*, 2016.

[38] W. Ma, Y. Li, X. Jia, and W. Xu. Transferable adversarial attack for both vision transformers and convolutional networks via momentum integrated gradients. In *Proc. IEEE Int'l Conf. Computer Vision*, 2023.

[39] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 2023.

[40] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2017.

[41] A. Nichol, J. Achiam, and J. Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

[42] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. Mcgrew, I. Sutskever, and M. Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proc. Int'l Conf. Machine Learning*, 2022.

[43] Y. Qian, S. He, C. Zhao, J. Sha, W. Wang, and B. Wang. Lea2: A lightweight ensemble adversarial attack via non-overlapping vulnerable frequency regions. In *Proc. IEEE Int'l Conf. Computer Vision*, 2023.

[44] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[45] S. Ren, F. Luzi, S. Lahrichi, K. Kassaw, L. M. Collins, K. Bradbury, and J. M. Malof. Segment anything, from space? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

[46] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2022.

[47] V. Q. Vo, E. Abbasnejad, and D. C. Ranasinghe. Brusleattack: A query-efficient score-based black-box sparse adversarial attack. In *Proc. Int'l Conf. Learning Representations*, 2024.

[48] C. Wang, Y. Yu, L. Guo, and B. Wen. Benchmarking adversarial robustness of image shadow removal with shadow-adaptive attacks. In *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, pages 13126–13130. IEEE, 2024.

[49] J. Wang, X. Li, and J. Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2018.

[50] K. Wang, X. He, W. Wang, and X. Wang. Boosting adversarial transferability by block shuffle and rotation. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2024.

[51] X. Wang and K. He. Enhancing the transferability of adversarial attacks through variance tuning. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2021.

[52] X. Wang, X. He, J. Wang, and K. He. Admix: Enhancing the transferability of adversarial attacks. In *Proc. IEEE Int'l Conf. Computer Vision*, 2021.

[53] D. Wu, Y. Wang, S.-T. Xia, J. Bailey, and X. Ma. Skip connections matter: On the transferability of adversarial examples generated with resnets. In *Proc. Int'l Conf. Learning Representations*, 2019.

[54] J. Wu, R. Fu, H. Fang, Y. Liu, Z. Wang, Y. Xu, Y. Jin, and T. Arbel. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023.

[55] S. Xia, Y. Yu, X. Jiang, and H. Ding. Mitigating the curse of dimensionality for certified robustness via dual randomized smoothing. In *Proc. Int'l Conf. Learning Representations*, 2024.

[56] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. L. Yuille. Improving transferability of adversarial examples with input diversity. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2019.

[57] X. Xu, K. Kong, N. Liu, L. Cui, D. Wang, J. Zhang, and M. Kankanhalli. An llm can fool itself: A prompt-based adversarial attack. In *The Twelfth International Conference on Learning Representations*, 2023.

[58] Y. Yu, Y. Wang, S. Xia, W. Yang, S. Lu, Y.-p. Tan, and A. Kot. Purify unlearnable examples via rate-constrained variational autoencoders. In *Proc. Int'l Conf. Machine Learning*, 2024.

[59] Y. Yu, Y. Wang, W. Yang, S. Lu, Y.-P. Tan, and A. C. Kot. Backdoor attacks against deep image compression via adaptive frequency trigger. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 12250–12259, June 2023.

[60] Y. Yu, W. Yang, Y.-P. Tan, and A. C. Kot. Towards robust rain removal against adversarial attacks: A comprehensive benchmark analysis and beyond. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2022.

[61] K. Zhang and D. Liu. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023.

[62] Z. Zhao, Z. Liu, and M. Larson. On success and simplicity: A second look at transferable targeted attacks. *Proc. Annual Conf. Neural Information Processing Systems*, 2021.

[63] Z. Zhou, S. Hu, M. Li, H. Zhang, Y. Zhang, and H. Jin. Advclip: Downstream-agnostic adversarial examples in multimodal contrastive learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2023.

[64] Z. Zhou, S. Hu, R. Zhao, Q. Wang, L. Y. Zhang, J. Hou, and H. Jin. Downstream-agnostic adversarial examples. In *Proc. IEEE Int'l Conf. Computer Vision*, 2023.

[65] Z. Zhou, M. Li, W. Liu, S. Hu, Y. Zhang, W. Wan, L. Xue, L. Y. Zhang, D. Yao, and H. Jin. Securely fine-tuning pre-trained encoders against adversarial examples. In *Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP'24)*, 2024.

[66] Z. Zhou, Y. Song, M. Li, S. Hu, X. Wang, L. Y. Zhang, D. Yao, and H. Jin. Darksam: Fooling segment anything model to segment nothing. In *Proc. Annual Conf. Neural Information Processing Systems*, 2024.

A Appendix

A.1 Randomness test

Table A.3: Experimental randomness of transfer-based adversarial attacks on SAMs (subset).

Model	Medical SAM		Shadow-SAM	Camouflaged-SAM					
Dataset	CT-Scans		ISTD	COD10K		CAMO		CHAME	
Metrics	mDSC↓	mHD↑	BER↑	S_α↓	MAE↑	S_α↓	MAE↑	S_α↓	MAE↑
MI-FGSM	31.35$_{(\pm1.11)}$	73.26$_{(\pm7.67)}$	3.05$_{(\pm0.26)}$	0.37$_{(\pm4e-3)}$	0.23$_{(\pm4e-3)}$	0.33$_{(\pm7e-3)}$	0.27$_{(\pm6e-3)}$	0.37$_{(\pm1e-3)}$	0.24$_{(\pm4e-3)}$
DMI-FGSM	27.19$_{(\pm1.00)}$	87.54$_{(\pm8.085)}$	2.53$_{(\pm0.16)}$	0.45$_{(\pm5e-3)}$	0.11$_{(\pm3e-3)}$	0.39$_{(\pm4e-3)}$	0.20$_{(\pm3e-3)}$	0.43$_{(\pm3e-3)}$	0.14$_{(\pm2e-3)}$
PGN	40.64$_{(\pm1.23)}$	40.76$_{(\pm8.95)}$	3.05$_{(\pm0.11)}$	0.36$_{(\pm5e-3)}$	0.25$_{(\pm8e-3)}$	0.32$_{(\pm6e-3)}$	**0.33**$_{(\pm5e-3)}$	0.36$_{(\pm3e-3)}$	0.23$_{(\pm7e-3)}$
BSR	28.62$_{(\pm1.41)}$	93.41$_{(\pm5.82)}$	2.58$_{(\pm0.15)}$	0.45$_{(\pm6e-3)}$	0.11$_{(\pm3e-3)}$	0.40$_{(\pm3e-3)}$	0.20$_{(\pm2e-3)}$	0.43$_{(\pm4e-3)}$	0.14$_{(\pm3e-3)}$
ILPD	28.68$_{(\pm2.18)}$	63.53$_{(\pm9.22)}$	2.74$_{(\pm0.15)}$	0.35$_{(\pm2e-3)}$	0.25$_{(\pm7e-3)}$	**0.33**$_{(\pm2e-3)}$	0.28$_{(\pm3e-3)}$	0.37$_{(\pm2e-3)}$	0.24$_{(\pm4e-3)}$
MUI-GRAT	**2.01**$_{(\pm0.28)}$	**148.18**$_{(\pm10.20)}$	**10.10**$_{(\pm0.28)}$	0.29$_{(\pm0.01)}$	**0.40**$_{(\pm5e-3)}$	0.31$_{(\pm5e-3)}$	0.33$_{(\pm4e-3)}$	**0.33**$_{(\pm5e-3)}$	**0.28**$_{(\pm8e-3)}$
MUI-GRAT+DMI	2.23$_{(\pm0.37)}$	123.76$_{(\pm24.88)}$	4.52$_{(\pm0.20)}$	0.40$_{(\pm4e-3)}$	0.20$_{(\pm4e-3)}$	0.37$_{(\pm4e-3)}$	0.27$_{(\pm2e-3)}$	0.38$_{(\pm5e-3)}$	0.23$_{(\pm8e-3)}$
MUI-GRAT+PGN	15.64$_{(\pm0.63)}$	112.38$_{(\pm10.55)}$	**16.41**$_{(\pm0.79)}$	0.33$_{(\pm7e-3)}$	0.33$_{(\pm0.01)}$	0.31$_{(\pm0.01)}$	0.36$_{(\pm0.01)}$	0.34$_{(\pm4e-3)}$	0.27$_{(\pm8e-3)}$
MUI-GRAT+BSR	6.25$_{(\pm0.58)}$	82.56$_{(\pm10.75)}$	3.87$_{(\pm0.17)}$	0.41$_{(\pm4e-3)}$	0.21$_{(\pm0.01)}$	0.38$_{(\pm6e-3)}$	0.27$_{(\pm6e-3)}$	0.40$_{(\pm0.01)}$	0.22$_{(\pm0.01)}$
MUI-GRAT+ILPD	**0.95**$_{(\pm0.27)}$	**165.30**$_{(\pm16.28)}$	10.67$_{(\pm0.52)}$	**0.28**$_{(\pm7e-3)}$	**0.40**$_{(\pm0.01)}$	**0.30**$_{(\pm4e-3)}$	**0.33**$_{(\pm7e-3)}$	**0.34**$_{(\pm4e-3)}$	**0.28**$_{(\pm0.01)}$

We evaluated 10 attack methods presented in our paper over 5 random seed runs on the subset of SAM's downstream tasks and reported the mean performance with its standard deviation. We use the same experimental setting provided in Section 6.1. The results indicate that the randomness is small and similar among all attacking methods. The uncertainty level of UMI-GRAT in mean Hausdorff Distance (mHD) is marginally higher compared to the other methods. This can account for the higher mHD value achieved by the UMI-GRAT.

A.2 Attacking open-sourced SAMs

Table A.4: The performance of open-sourced SAMs under attacks.

Surrogate model	Attack	SAM-Vit-B		SAM-Vit-L		SAM-Vit-H	
		mAP↓	mIOU↓	mAP↓	mIOU↓	mAP↓	mIOU↓
SAM-Vit-B	MI-FGSM	1.24	5.23	14.17	24.12	24.20	34.96
	MUI-GRAT	**0.65**	**2.21**	**3.67**	**9.27**	**8.43**	**15.41**
SAM-Vit-L	MI-FGSM	**14.18**	**21.11**	1.11	4.03	**16.42**	**25.10**
	MUI-GRAT	16.41	23.05	1.39	**2.40**	18.12	28.67
SAM-Vit-H	MI-FGSM	21.74	29.35	15.58	24.71	1.46	6.51
	MUI-GRAT	**18.50**	**25.26**	**10.00**	**18.12**	**0.58**	**2.25**

We report the mean average precision (mAP) and mean intersection-over-union (mIOU) metrics to evaluate the performance of attacking open-sourced SAMs. We compare the performance of MI-FGSM and ours MUI-GRAT. The implementation of these attacks adheres to the details described in Section 6.1. We randomly select 500 images from the SA-1B dataset and evaluate the performance of SAMs under the 'AutomaticMaskGenerator' mode. We white-box attacks SAM-Vit-B, SAM-Vit-L, and SAM-Vit-H models. Meanwhile, we also discuss the black-box transferability of generated AEs across different models. The results are shown in Table A.4.

Our findings reveal that both MI-FGSM and MUI-GRAT achieve comparably strong attack performance in the white-box scenario. In the black-box scenario, where AEs generated on one surrogate model are transferred to a different victim model, our results indicate that the surrogate model is critical for the transferability of generated AEs. The AEs with the strongest transferability are generated by our MUI-GRAT on attacking SAM-Vit-B, which exhibits a great performance gain compared with others. However, when employing SAM-ViT-L as the surrogate model, there is a slight reduction in the transferability of AEs produced by our MUI-GRAT. Conversely, when the surrogate model is SAM-Vit-H, the transferability of AEs generated by our MUI-GRAT surpasses MI-FGSM by a large margin.

A.3 Analysis of the attack efficiency

We analyze the real-time attack efficiency of the methods mentioned above. We report the average time required for generating one AE when the input resolution is 512×512 on the SAM-Vit-B model using one RTX 4090 GPU. The results are shown in Table A.5.

We find that our proposed MUI-GRAT achieves second-high efficiency when compared with others. The ILPD requires extra attack iterations (e.g., 10-step MI-FGSM) to find a directional guide vector v. The input

Table A.5: Analysis of the attack efficiency

Method	Num. of samples needed per iteration	Avg. time (s)
MI-FGSM	1	**0.32**
DMI-FGSM	2	0.69
PGN	8	3.96
BSR	8	2.91
ILPD	1	0.66
MUI-GRAT	1	0.36

augmentation-based attack methods, such as BSR and DMI-FGSM, require multiple samples in each iteration to do the augmentation and the PGN also requires multiple samples to obtain a stable gradient direction. However, our method utilizes an offline generated MUI and only conducts one-time gradient augmentation in each iteration, thus achieving a much higher efficiency than others.

A.4 Hardware Setup

We run our experiments for attacking the medical segmentation model using one RTX 4090 GPU with 24 GB memory. We run the rest of the experiments using one RTX A6000 GPU with 48 GB memory.

A.5 Visualization of the adversarial examples and the prediction

We visualize adversarial examples generated by MUI-GRAT utilizing solely the open-sourced SAM and the corresponding adversarial predictions in Figure A.5 and Figure A.6. The images are randomly selected. This visualization provides a more straightforward demonstration of the impact of the adversarial attack threatened by MUI-GRAT, showing how it significantly compromises the reliability of the large foundation model's predictions with just a single independent input image.
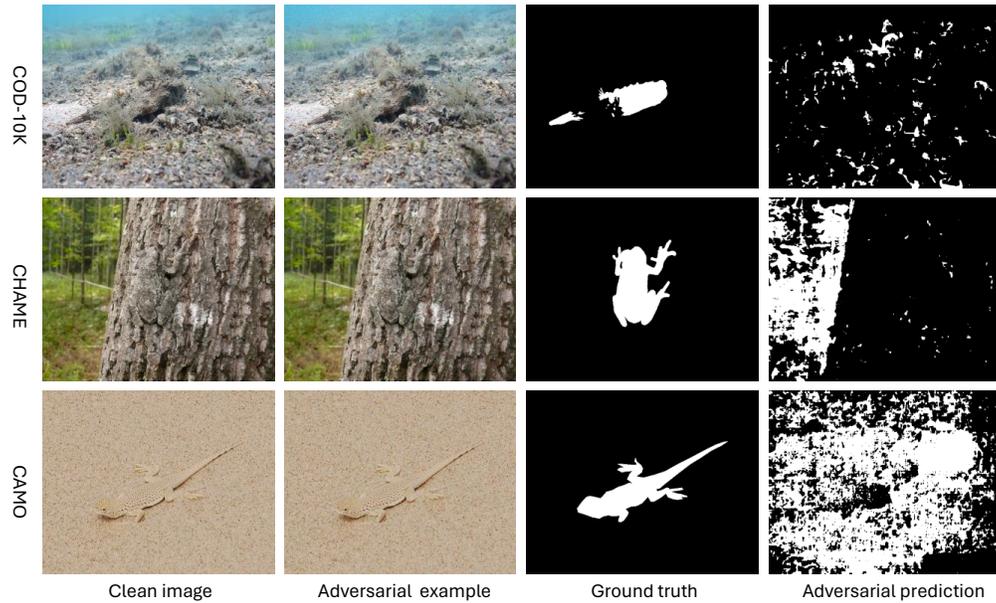

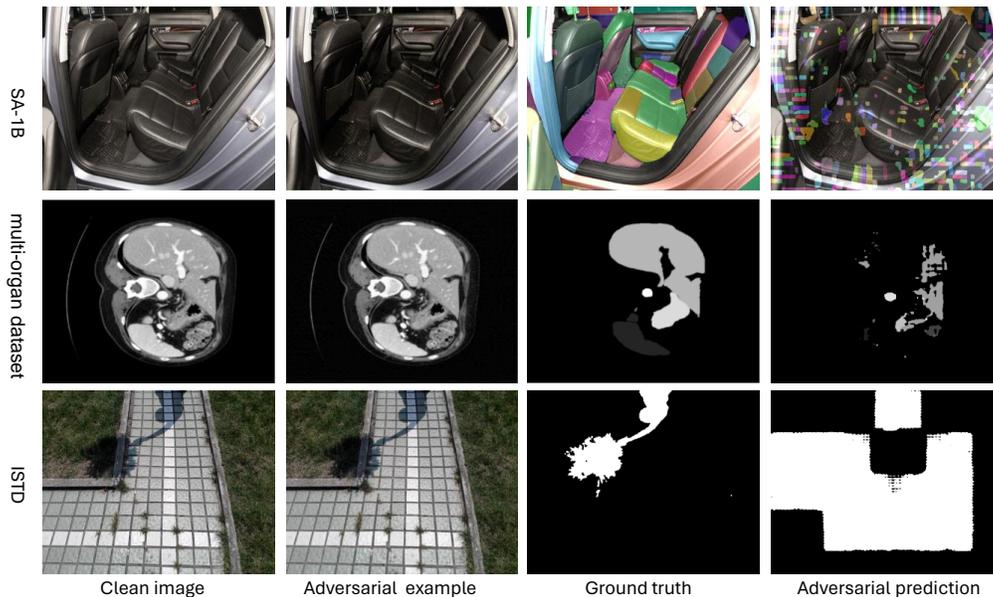Figure A.5: The visualized adversarial attack results in camouflaged object segmentation task.

Figure A.6: The visualized adversarial attack results in natural, medical, and shadow image segmentation tasks.

A.6 Impact Statement

The increasing reliance on large foundation models in various real-world applications amplifies the critical importance of ensuring their secure utilization. This paper mainly discusses the potential risks of adversarial threats associated with the direct utilization and fine-tuning of the open-sourced model even on private and encrypted datasets. To accomplish that, we begin an investigation into a more practical while challenging adversarial attack problem: attacking various SAM's downstream models by solely utilizing the information from the open-sourced SAM. We then provide the theoretical insights and build the experimental setting and benchmark, aiming to serve as a preliminary exploration for future research in this area. Experimentally, we validate the vulnerability of SAM and its downstream models under the proposed MUI-GRAT, indicating the security risk inherent in the direct utilization and fine-tuning of open-sourced large foundation models, thus highlighting the urgent need for robust defense mechanisms to protect these models from adversarial threats.

A.7 Limitations

The limitations of our paper are:

- The proposed UMI-GRAT is not contingent upon a prior regarding the model's architecture, suggesting its potential applicability across various model paradigms. However, the experiments only tested MUI-GRAT on the prevalent SAMs and their downstream models. The capability of UMI-GRAT to pose a threat to other large foundation models remains a topic for further exploration.

- While this paper highlights the risk of direct utilization of SAM and fine-tuning it on the downstream task, this paper does not provide and validate an effective solution for this secure concern.

87562

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We demonstrate the contributions and the scope of our paper in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our work in Appendix A.7

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: For each theoretical result in Section 4, we give the detailed proof and illustration.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We present the details of the implementation of our method in Section 6.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We intend to release the code associated with this work subsequent to the paper's acceptance. This will allow us to confirm that the code is stable and thoroughly tested prior to its public dissemination. Our goal in releasing the code is to enable replication of our findings, foster collaboration with fellow researchers, and uphold the principles of open science.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We give the implementation details and explain every result we got in the experiment to help readers understand.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Error bars are not reported because it would be too computationally expensive. But we keep the random seed fixed for all competing methods.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We list the equipment needed for running and reproduce our experiment in Appendix A.4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer:[Yes]

   Justification: We have reviewed and met the code of ethics for our research.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the social impact of our research in Section A.6.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This research does not have this kind of risk.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We have cited the original paper that produced the code package or dataset.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: This paper does not introduce new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper does not describe potential risks incurred by study participants.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.