Scaling Laws for Black-box Adversarial Attacks

Chuan Liu[1], Huanran Chen[1], Yichi Zhang[1], Jun Zhu[1], and Yinpeng Dong[1,2⋆]

[1] Tsinghua University
[2] Shanghai Qi Zhi Institute

Abstract. Adversarial examples exhibit cross-model transferability, enabling threatening black-box attacks on commercial models. Model ensembling, which attacks multiple surrogates, is a known strategy to improve this transferability. However, prior studies typically use small, fixed ensembles, which leaves open an intriguing question of whether scaling the number of surrogate models can further improve black-box attacks. In this work, we conduct the first large-scale empirical study of this question. We show that by resolving gradient conflict with advanced optimizers, we overcome the quantitative limitations of idealized theoretical bounds to empirically discover a robust log-linear scaling law, demonstrating that the Attack Success Rate scales linearly with the logarithm of the ensemble size T. We rigorously verify this law across standard classifiers, SOTA defenses, and MLLMs, and find that scaling distills robust, semantic features of the target class. Consequently, we apply this fundamental insight to benchmark SOTA MLLMs. This reveals both the attack's devastating power and a clear robustness hierarchy, as evidenced by achieving an over 80% transfer attack success rate on proprietary models like GPT-4o, while also highlighting the exceptional resilience of Claude-3.5-Sonnet. Our findings urge a shift in focus for robustness evaluation: from designing intricate algorithms on small ensembles to understanding the principled and powerful threat of scaling.

Keywords: Black-box Attacks · Scaling Laws · Model Ensembling

1 Introduction

Over the past decade, deep learning has achieved remarkable advancements across various computer vision tasks, leading to its widespread applications [41]. However, deep neural networks (DNNs) remain susceptible to adversarial examples [26, 60] – perturbations that are subtle and nearly imperceptible but can significantly mislead the model, resulting in incorrect predictions. Adversarial attacks have garnered considerable interest due to their implications for understanding DNN functionality [21], assessing model robustness [76], and informing the development of more secure algorithms [52].

Adversarial attacks are typically categorized into white-box and black-box attacks. Given the greater risks with commercial services, e.g., GPT-4 [54],

⋆ ✉ Corresponding author: dongyinpeng@mail.tsinghua.edu.cn

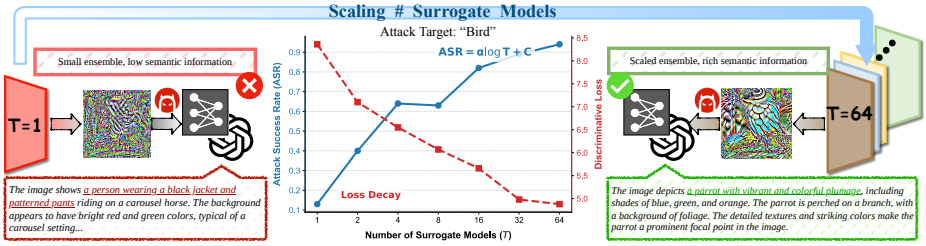


Fig. 1: The Scaling Laws of Black-Box Adversarial Attacks. (Center) The Attack Success Rate (ASR) exhibits a robust log-linear relationship with the ensemble cardinality T. **(Left)** At $T = 1$, the adversarial perturbation contains low semantic information, failing to mislead advanced MLLMs. **(Right)** By scaling to $T = 64$, the optimizer distills rich semantic features of the target class ("Bird"), successfully manipulating the generative output of state-of-the-art commercial models.

black-box attack methodologies [19, 47] have attracted growing attention. One prominent approach within black-box attacks is the transfer-based attack, which approximately optimizes the adversarial examples against the black-box model by minimizing the discriminative loss on a selection of predefined surrogate models. Researchers have devised advanced optimization algorithms [44, 52, 64] and data augmentation techniques [20, 50] in transfer-based attacks. Model ensembling [19, 47] further facilitates these methods by leveraging multiple surrogate models simultaneously to improve the transferability of adversarial examples. By averaging the gradients or predictions from different models, ensemble attack methods seek to increase the transferability across various target models [11, 19].

Despite the progress in ensemble attack algorithms, the number of surrogate models used in previous research is often limited [11, 73, 75]. Recent advancement in large language models shows that scaling training data can significantly improve the generalization ability [36]. Since the transfer-based attack is also a generalization problem that expects adversarial examples from surrogate models to generalize to target models, while the training data are surrogate models rather than text or images [11, 19], this equivalence prompts an intriguing question: *Can we achieve significant improvements in transfer-based attacks by scaling the number of surrogate models, akin to the scaling of data and model parameters in large language models like GPT?*

In this work, we conduct the first large-scale empirical investigation into the scaling laws of black-box adversarial attacks, establishing that this phenomenon emerges only when the gradient conflict issue is properly resolved (Sec. 3.2). While classical optimization theory provides a qualitative motivation for scaling the ensemble, we demonstrate that its idealized assumptions fail to quantitatively predict real-world adversarial behavior under strict boundary constraints (Sec. 4.1). Driven by this theoretical gap, our core contribution is the empirical discovery and rigorous verification of a robust log-linear scaling law: the Attack Success Rate on black-box targets scales linearly with the logarithm of the ensemble cardinality T (Sec. 4.3). We demonstrate the broad generality of this

law across diverse target families, including standard classifiers, state-of-the-art defenses, and multimodal large language models (Sec. 4.4). Furthermore, we provide a qualitative explanation for this scaling behavior (Sec. 6.2): as T increases, the optimizer is forced to discard model-specific, non-robust features and instead distills the robust, semantic features of the target class.

Inspired by this fundamental insight, we then extend our scaling laws to attack the foundational vision encoders (e.g., CLIP) of modern MLLMs (Sec. 5) to demonstrate its practical power as both a SOTA attack and a novel robustness benchmark. We achieve devastatingly high ASR on top-tier models, including 90.0% ASR on Qwen3-VL-235B and 85.0% ASR on GPT-4o. Meanwhile, our scaling benchmark reveals a stark hierarchy in SOTA model robustness. These findings show that scaling is a critical, measurable threat vector, and we urge the community to adopt large-scale robustness evaluation, rather than designing intricate algorithms on small ensembles, to build and validate genuinely robust and trustworthy foundation models.

2 Related Work

2.1 Black-Box Attacks

Deep learning models are widely recognized for their vulnerability to adversarial attacks [26, 60]. Extensive research has documented this susceptibility, particularly in image classification contexts [2, 8], under two principal settings: white-box and black-box attacks. In the white-box setting, the adversary has complete access to the target model [52]. Conversely, in the black-box setting, such information is obscured from the attacker. With limited access to the target model, black-box attacks typically either construct queries and rely on feedback from the model [17, 33] or leverage the transferability of predefined surrogate models [47]. Query-based attacks often suffer from high time complexity and substantial computational costs, making transfer-based methods more practical and cost-effective for generating adversarial examples.

Within the realm of transfer-based attacks, existing methods can generally be categorized into three types: input transformation, gradient-based optimization, and ensemble attacks. Input transformation methods employ data augmentation techniques before feeding inputs into models, such as resizing and padding [69], translations [20], or transforming them into frequency domain [50]. Gradient-based methods concentrate on designing optimization algorithms to improve transferability, incorporating strategies such as momentum optimization [19, 44], scheduled step sizes [24], and gradient variance reduction [64].

2.2 Ensemble Attacks

Inspired by model ensemble techniques in machine learning [15,25,67], researchers have developed sophisticated adversarial attack strategies utilizing sets of surrogate models. In forming an ensemble paradigm, practitioners often average over

losses [19], predictions [47], or logits [19]. Additionally, advanced ensemble algorithms have been proposed to enhance adversarial transferability [9, 11, 42, 71]. However, these approaches tend to treat model ensemble as a mere technique for improving transferability without delving into the underlying principles.

Prior studies [32] indicate that increasing the number of classifiers in adversarial attacks can effectively reduce the upper bound of generalization error in Empirical Risk Minimization. Recent research [73] has theoretically defined transferability error and sought to minimize it through model ensemble. Nonetheless, while an upper bound is given, they do not really scale up in practice and have to train neural networks from scratch to meet the assumption of sufficient diversity among ensemble models [46].

2.3 Scaling Laws

In recent years, the concept of scaling laws has gained significant attention. These laws describe the relationship between model size and generalization ability, typically reflected in evaluation loss [36]. Subsequent studies have expanded on scaling laws from various perspectives, including image classification [3, 29], language modeling [58] and mixture of experts (MoE) models [59], highlighting the fundamental nature of this empirical relationship.

In the domain of adversarial attack research, recent work has attempted to derive scaling laws for adversarially trained defense models on CIFAR-10 [6, 37]. However, the broader concept of scaling remains underexplored. While some preliminary observations on the effect of enlarging model ensembles can be found in prior studies [73], these effects have not been systematically analyzed. Recent work [65] also explore the scaling law of attack success rate when jailbreaking large language models [10, 78] by increasing number of in-context demonstrations. In this work, however, we focus on the empirical scaling laws within model ensembles, and verify it through extensive experiments on open-source pretrained models. Motivated by this perspective, we unveil vulnerabilities within commercial LLMs by targeting their vision encoders.

3 Backgrounds

In this section, we provide necessary backgrounds, including the problem formulations for different attack settings and the prerequisite algorithms.

3.1 Problem Formulations

Our work investigates scaling laws across multiple black-box attack scenarios. We formulate the settings below.

Attacks on Image Classifiers. Let $\mathcal{F}$ denote the set of all image classifiers. The vulnerability of such classifiers to adversarial examples has been a long-standing area of research [26, 60]. Given a natural image $\boldsymbol{x}_{\text{nat}}$ with its true label

$\boldsymbol{y}_{\text{nat}}$, the goal of a transfer-based attack is to craft an adversarial example $\boldsymbol{x}$ within an ℓ_∞ perturbation budget ϵ that fools black-box target models.

Targeted Attack. The primary focus of our work is the challenging targeted attack setting. Given a target class $\boldsymbol{y}$ (where $\boldsymbol{y} \neq \boldsymbol{y}_{\text{nat}}$), the goal is to craft an $\boldsymbol{x}$ that is misclassified as $\boldsymbol{y}$. This can be formulated as minimizing the expected loss over all models in $\mathcal{F}$:

$$\min_{\boldsymbol{x}} \mathbb{E}_{f \in \mathcal{F}} \left[\mathcal{L}(f(\boldsymbol{x}), \boldsymbol{y}) \right], \quad \text{s.t.} \left\| \boldsymbol{x} - \boldsymbol{x}_{\text{nat}} \right\|_\infty \leq \epsilon, \tag{1}$$

where $\mathcal{L}$ is a loss function (e.g., cross-entropy). In practice, we approximate this by ensembling a limited set of T surrogate classifiers $\{f_i\}_{i=1}^T \subset \mathcal{F}$:

$$\min_{\boldsymbol{x}} \frac{1}{T} \sum_{i=1}^{T} \mathcal{L}(f_i(\boldsymbol{x}), \boldsymbol{y}), \quad \text{s.t.} \left\| \boldsymbol{x} - \boldsymbol{x}_{\text{nat}} \right\|_\infty \leq \epsilon. \tag{2}$$

Untargeted Attack. We also validate our scaling laws in the untargeted setting in Sec. 6.1, where the goal is to mislead the classifier to **any** class other than $\boldsymbol{y}_{\text{nat}}$. The objective is to maximize the loss with respect to original $\boldsymbol{y}_{\text{nat}}$:

$$\max_{\boldsymbol{x}} \frac{1}{T} \sum_{i=1}^{T} \mathcal{L}(f_i(\boldsymbol{x}), \boldsymbol{y}_{\text{nat}}), \quad \text{s.t.} \left\| \boldsymbol{x} - \boldsymbol{x}_{\text{nat}} \right\|_\infty \leq \epsilon. \tag{3}$$

Attacks on Vision Encoders. Contrastive Language-Image Pretraining (CLIP) is a popular self-supervised framework that can pretrain large-scale language-vision models on web-scale text-image pairs via contrastive learning [40, 55, 57]. Based on the generalization ability of CLIP models, previous works have utilized them as surrogate models in model ensembles [16, 23, 31].

We extend our analysis to attack these vision encoders, which are foundational to modern MLLMs. Instead of classifying to a specific logit, the goal is to craft an image $\boldsymbol{x}$ whose embedding $f^I(\boldsymbol{x})$ is close to the text embedding $\boldsymbol{e}_{\text{tar}}$ of a target label $\boldsymbol{y}$. This target embedding is typically pre-computed using a text encoder, f^T. The objective function is adapted by replacing the classifier f_i with a surrogate image encoder f_i^I and the cross-entropy loss $\mathcal{L}$ with a similarity-based loss $\mathcal{L}_{\text{sim}}$ (e.g., cosine embedding loss):

$$\min_{\boldsymbol{x}} \frac{1}{T} \sum_{i=1}^{T} \mathcal{L}_{\text{sim}}(f_i^I(\boldsymbol{x}), \boldsymbol{e}_{\text{tar}}), \quad \text{s.t.} \left\| \boldsymbol{x} - \boldsymbol{x}_{\text{nat}} \right\|_\infty \leq \epsilon, \tag{4}$$

where $\boldsymbol{e}_{\text{tar}} = f^T(\boldsymbol{y})$ is the fixed target text embedding.

3.2 Prerequisite Algorithms

Researchers have developed sophisticated algorithms to enhance adversarial transferability, which can be categorized into gradient optimizers, input transformations, and ensemble strategies.

Gradient Optimizers. These methods refine the gradient update step itself. A pioneering example is **MI-FGSM** [19], which stabilizes gradients by integrating a momentum term:

$$\boldsymbol{g}_{t+1} = \mu \cdot \boldsymbol{g}_t + \frac{\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x_t}, \boldsymbol{y})}{\|\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x_t}, \boldsymbol{y})\|_1}, \tag{5}$$

with update $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \alpha \cdot \text{sign}(\boldsymbol{g}_{t+1})$. Other methods like VMI-FGSM [64] further refine this momentum concept.

Input Transformations. These methods apply data augmentation during the attack to find more generalizable adversarial patterns. Prominent examples include Diverse Inputs (DI) [69] and Spectrum Simulation Attack (SSA) [50]. SSA, a state-of-the-art method, applies spectrum transformations $\mathcal{T}$ and computes gradients on the transformed input $\mathcal{T}(\boldsymbol{x})$.

Ensemble Strategies and Gradient Conflict. The above components can be applied to ensemble attacks. The most straightforward method is logit ensembling [19], where the optimizer computes gradients over the sum of all models' logits. While simple, this strategy exposes a critical flaw: **gradient conflict**. We discover that as the number of surrogates increases, the gradients from different models, which often have random directions, tend to cancel each other out.

To address this, advanced optimizers are required. These methods aim to find a common direction of vulnerability rather than simply averaging gradients. They typically work by reducing gradient variance, such as SVRE [71], or by encouraging gradient alignment, which is the approach taken by the Common Weakness Attack (CWA) [11]. CWA seeks common vulnerabilities by encouraging flatter loss landscapes and minimizing the second-order Taylor series:

$$\frac{1}{T}\sum_{i=1}^{T}\left[\mathcal{L}(f_i(\boldsymbol{p}_i, \boldsymbol{y}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{p}_i)^\top H_i(\boldsymbol{x} - \boldsymbol{p}_i)\right], \tag{6}$$

where $\boldsymbol{p}_i$ is the closest optimum of model f_i to $\boldsymbol{x}$ and H_i is the Hessian.

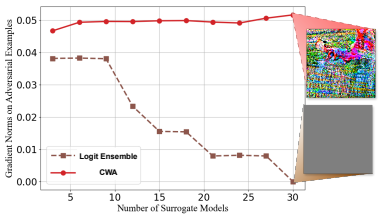


Fig. 2: Comparisons between CWA and MI-FGSM with logit averaging. On the right side, we visualize the gradient at 30 surrogate models and 10 iteration steps.

Fig. 2 presents the impact of scaling the number of surrogate models on the optimization process. In this experiment, we set the attack target to the "tabby cat" class and vary the ensemble cardinality from 1 to 30 models. We record the L_1 norm of the gradient after 10 iterations. As the plot demonstrates, the naive logit ensemble strategy suffers from severe gradient conflict. Conversely, the CWA algorithm successfully aligns the update directions, maintaining a stable gradient norm across all ensemble sizes and yielding rich, structured adversarial perturbations.

Our Chosen Algorithm: SSA-CWA. For our main experiments, we require an algorithm that both resolves the gradient conflict and achieves state-of-the-art performance. We therefore select SSA-CWA, which combines the advanced input transformation of SSA [50] with the robust gradient optimization of CWA [11] (pseudocode provided in Appendix C). While leveraging SSA-CWA as our core algorithm, we also explore the scaling laws of other optimizers and validate our gradient conflict hypothesis in Sec. 6.1.

4 Scaling Laws for Adversarial Attacks

In this section, we present our key findings on scaling laws for black-box adversarial attacks. Sec. 4.1 contrasts idealized theory to motivate our large-scale empirical study. Sec. 4.2 details our experimental setup for generating adversarial examples. Sec. 4.3 presents the initial discovery of the empirical scaling laws on standard classifiers. Finally, Sec. 4.4 verifies the broad generality of these laws by testing them against defended models and across diverse MLLMs.

4.1 Motivation

In transfer-based black-box attacks, $\boldsymbol{x}_{\text{nat}}$ and $\boldsymbol{y}$ are predefined. Consequently, the optimization objectives can be considered as functions of variable $\boldsymbol{x}$. Let $\boldsymbol{x}^*$ and $\hat{\boldsymbol{x}}$ be the minimizers of the expected population loss $\mathbb{L}(\boldsymbol{x})$ and the empirical ensemble loss $\hat{\mathbb{L}}(\boldsymbol{x})$, respectively:

$$\boldsymbol{x}^* = \arg\min_x \mathbb{E}_{f\in\mathcal{F}}[\mathcal{L}(f(\boldsymbol{x}), \boldsymbol{y})], \quad \hat{\boldsymbol{x}} = \arg\min_x \frac{1}{T}\sum_{i=1}^{T} \mathcal{L}(f_i(\boldsymbol{x}), \boldsymbol{y}). \tag{7}$$

To motivate our study, we first analyze the convergence behavior under an idealized, unconstrained Empirical Risk Minimization (ERM) framework.

Theorem 1. *Suppose the model ensemble $\{f_i\}_{i=1}^T$ is i.i.d. sampled from $\mathcal{F}$, and the optimization is unconstrained such that $\nabla\mathbb{L}(\boldsymbol{x}^*) = 0$. As $T \to \infty$, the empirical minimizer $\hat{\boldsymbol{x}}$ converges in distribution ($\xrightarrow{d}$) to $\boldsymbol{x}^*$, and the loss gap behaves asymptotically as:*

$$T(\mathbb{L}(\hat{\boldsymbol{x}}) - \mathbb{L}(\boldsymbol{x}^*)) \xrightarrow{d} \frac{1}{2}\|S\|_2^2, \tag{8}$$

with $S \sim \mathcal{N}(0, (\nabla^2\mathbb{L}(\boldsymbol{x}^*))^{1/2}\text{Cov}(\mathbb{L}(f_i))(\nabla^2\mathbb{L}(\boldsymbol{x}^*))^{1/2})$.

Remark 1. Theorem 1 suggests that under ideal conditions, the difference between the empirical loss and the population loss converges at a rate of $\mathcal{O}(1/T)$.

While Theorem 1 confirms that scaling the ensemble size T is a principled approach to bridging the generalization gap, it also highlights a crucial disconnect with practical adversarial attacks. First, since adversarial perturbations are strictly bounded by an ℓ_∞ norm constraint, the unconstrained stationarity

assumption ($\nabla \mathbb{L}(\boldsymbol{x}^*) = 0$) becomes invalid, necessitating KKT conditions instead. Second, the theorem assumes that surrogates are sampled independently. In practice, however, real-world surrogates are typically pretrained on similar datasets and share strong architectural biases, rendering them highly correlated rather than i.i.d. [34]. Consequently, our theorem cannot accurately model the convergence behavior of highly correlated ensembles under strict boundary constraints. It is precisely this theoretical disconnect that directly motivates our large-scale empirical investigation in the following sections.

4.2 Experimental Setup

To investigate the relationship between the number of surrogate models and attack efficacy, we first define our comprehensive experimental setup.

Surrogate Pool. To ensure the transferability and diversity of our attack, we select a large pool of 64 models from Torchvision [56] and Timm [66]. This pool includes a wide variety of architectures, hyperparameters, and training datasets, such as AlexNet [38], DenseNet [30], EfficientNet [61], ResNet [28], ViT [22], DeiT [62], and many others. The full list is available in Appendix B.

Attack Algorithm. We generate adversarial examples by attacking ensembles of surrogate models. For each attack, an ensemble is constructed by randomly sampling T models from our 64-model pool. We vary the ensemble cardinality T across a logarithmic scale from 2^0 to 2^6. We use the SSA-CWA attack [11] for 40 iteration steps, with a perturbation budget of $\epsilon = 8/255$ under the ℓ_∞ norm. The loss function $\mathcal{L}$ is the standard cross-entropy loss. This process yields 7 distinct sets of adversarial examples, one for each value of T.

Target Model Families. To test the universality of any observed scaling phenomena, we evaluate our pre-generated adversarial examples against three distinct families of black-box target models:

- **Family 1: Standard Image Classifiers.** We select 7 widely-used models, including Swin-Transformers [48], ConvNeXt [49], and ResNeXt [70].
- **Family 2: Defended Models.** We select 6 models representing SOTA defenses, including adversarially trained models (FGSMAT [39], EnsAT [63]) and purification defenses (HGD [43], BDR [18], JPEG, and DiffPure [53]).
- **Family 3: Multimodal Large Language Models.** We select 3 popular MLLMs: LLaVA-1.5-7B [45], Qwen-2.5-VL-7B [4], and Llama-3.2-11B-Vision [27]. These are used to further *validate the generality* of the laws across different data modalities and architectures.

Dataset and Generation Protocol. Following previous work, we leverage the 1000 images from the ImageNet-1K NIPS 2017 dataset [35, 72]. To ensure the generality of our findings and create a large-scale testbed, we target 8 distinct categories for each source image [74]. This creates a total of 8000 unique source-target pairs for each ensemble configuration. As we generate adversarial examples

Table 1: Coefficient of determination (R^2) for the log-linear scaling laws of Attack Success Rate across different target models and labels.

Target Model	Label: 150	Label: 426	Label: 507	Label: 590	Label: 62	Label: 715	Label: 843	Label: 952	Average
ConvNeXT Base	0.97	0.97	0.95	0.98	0.95	0.94	0.96	0.97	0.97
ConvNeXT Small	0.97	0.97	0.95	0.98	0.94	0.97	0.95	0.97	0.97
ResNeXt-101 32x8d	0.98	0.99	0.95	0.98	0.94	0.95	0.96	0.97	0.97
ResNeXt-50 32x4d	0.95	0.97	0.94	0.97	0.92	0.93	0.94	0.95	0.96
Swin Transformer Base	0.97	0.95	0.97	0.99	0.96	0.94	0.98	0.98	0.98
Swin Transformer Small	0.98	0.97	0.97	0.99	0.96	0.95	0.98	0.97	0.98
Swin Transformer V2 Small	0.97	0.96	0.91	0.98	0.96	0.95	0.98	0.98	0.98

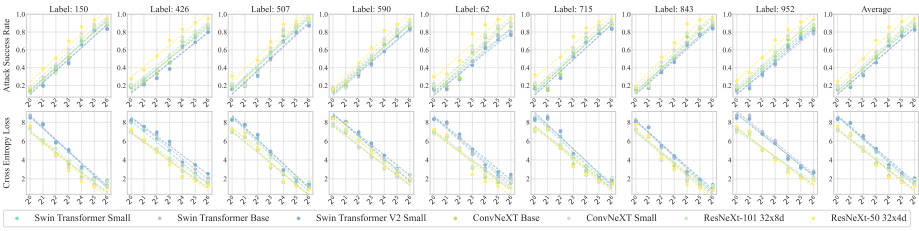


Fig. 3: Scaling Laws over model ensemble in a black-box setting. We plot the x-axis on a base-2 logarithmic scale, and we measure both the attack success rate (ASR) and average cross entropy-loss of target models. The cardinality of the model ensembles is varied from 2^0 to 2^6, with models randomly selected for each ensemble. We fit the results of selected target models by lines for better demonstration.

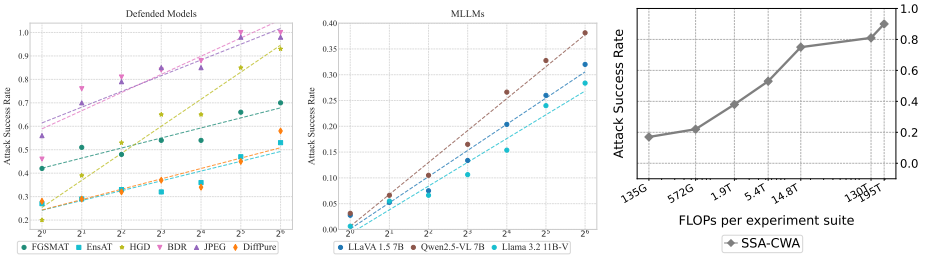
for 7 distinct ensemble cardinalities $T \in \{2^0, 2^1, \ldots, 2^6\}$, this results in a large-scale dataset of $7 \times 8000 = 56000$ generated adversarial examples.

To ensure our results are statistically robust and not biased by a single random selection of surrogate models, we employ a batched generation protocol. For any given cardinality T, we first partition the 1000 source images into 8 disjoint batches of 125 images. Then, for each batch, we randomly sample a *new* ensemble of T models from our surrogate pool. This newly sampled ensemble is used to craft the 125 adversarial examples for its assigned batch. This methodology ensures that the final metrics for each T are averaged over 8 distinct, randomly-drawn ensembles, providing a highly robust measurement of the scaling effect.

Metrics. For Families 1 and 2, we evaluate the Attack Success Rate (ASR) along with the cross-entropy loss between predictions and target labels to validate ASR as a transferability metric. For the text-generating Family 3, we use a string-matching paradigm against target labels. To accommodate the descriptive nature of MLLM outputs, a relaxed keyword map (see Appendix D) is applied to robustly identify semantic mentions of the target concept.

4.3 Log-Linear Scaling Laws in Image Classifiers

Empirical Observation. We begin our analysis by evaluating the generated adversarial examples on the standard image classifiers. As demonstrated in Fig. 3 (top row), we identify a robust empirical phenomenon: the Attack Success Rate exhibits strong *log-linear scaling laws* with the ensemble cardinality T.



(a) Scaling Laws on Family 2. **(b)** Scaling Laws on Family 3. **(c)** ASR vs. FLOPs.

Fig. 4: Generalization and tradeoff of the scaling laws. (a) The laws persist in defended models. (b) The laws hold for advanced MLLMs. (c) We profile the SSA-CWA attack with `torch.profile` and plot the average ASR over standard classifiers against total FLOPs. ASR improvement is not uniform per doubling of FLOPs, indicating that surrogate diversity is a more fundamental driver of transferability than raw computation.

As shown by the fitted lines, the ASR increases linearly as we increase T on a logarithmic scale. Concurrently, the average cross-entropy loss against the target label exhibits a mirrored log-linear decay. This strong symmetric relationship confirms that ASR is a valid and robust proxy for the true optimization objective, as the decreasing loss directly correlates with the increasing success rate.

Analysis of the Log-Linear Relationship. We empirically characterize the observed log-linear scaling laws through their mathematical parameters and valid boundaries. In this relationship, the intercept at $T = 1$ reflects a target model's baseline vulnerability to single-surrogate attacks, which heavily depends on the target's architecture and the specific adversarial class. The slope defines the attack's scaling efficiency, quantifying the marginal ASR gain per doubling of T and serving as an indicator of the model's resilience against ensemble scaling. Importantly, this predictable trend operates within a specific statistical regime. As T scales, the stable log-linear trend ultimately flattens as the ASR approaches the 1.0 saturation limit. Thus, our scaling law robustly captures the attack dynamics between initial variance stabilization and final metric saturation.

4.4 Verification of the Scaling Laws' Universality

Having established the existence of these log-linear scaling laws, we now test their universality. A critical question is whether this trend is a fragile artifact of standard classifiers or a more fundamental principle of transfer attacks. We test this by evaluating their robustness against active defenses (Family 2) and their generality across entirely different architectures and modalities (Family 3).

Robustness Against Defended Models. As demonstrated in Fig. 4a, the log-linear scaling laws remain robustly intact across a diverse set of SOTA defenses. While robust methods like DiffPure successfully lower the overall ASR, they do not eliminate the predictable relationship between $\log T$ and ASR. Notably, the scaling efficiency varies among the defense mechanisms. This confirms that active

defenses primarily increase the computational threshold required for a successful attack rather than providing absolute immunity, and demonstrates that scaling the ensemble is a viable path to breaking even SOTA defenses.

By leveraging these scaling laws, we improved the ASR of DiffPure [53] by over 30%, adversarially trained models by over 20%, and HGD [43] by over 50%.

Generalization Across Architectures. Finally, we perform the ultimate test of generality: do the adversarial examples, generated *only* from standard image classifiers, follow the same scaling laws on MLLMs?

Remarkably, as shown in Fig. 4b, the log-linear scaling laws persist even in these advanced multimodal models. This cross-modal generalization is highly non-trivial, yet they successfully transfer to MLLMs possessing vastly different architectures and trained on different data modalities. While the log-linear form is general, the specific robustness of each model varies. For instance, `LLaVA-1.5-7B` and `Qwen-2.5-VL-7B` have similar parameter counts ($\approx$ 7B) and both show high robustness to single-model attacks, yet their averaged ASR curves clearly diverge, indicating different vulnerabilities to scaling. This suggests that the relationship between model scale and adversarial robustness is more complex in MLLMs and challenges the simple 'model scale = robustness' narrative [6].

4.5 Practical Compute-Efficiency Analysis

Having verified the universality of the log-linear scaling laws, we now address a practical concern on the tradeoff between computational cost and performance.

We profile the attack using `torch.profile` and plot ASR against floating-point operations (FLOPs) in Fig. 4c. Two key observations emerge. First, ASR gain is not constant per doubling of FLOPs. Second, for a fixed computational budget, ensembling multiple lightweight models is more efficient than relying on a single heavyweight surrogate, because surrogate diversity (T) is a more fundamental driver of transferability than raw compute.

5 Extending Scaling Laws to Vision Encoders

Having established the log-linear scaling laws on standard classifiers, we now extend our investigation to a more complex and practical domain by attacking modern MLLMs by targeting their foundational vision encoders, such as CLIP [57]. We hypothesize that the same scaling principles hold when using the embedding-space attack formulation defined in Sec. 3.1.

5.1 Experimental Setup

We first introduce our settings for attacking pretrained CLIP models and evaluating them on MLLM targets.

Surrogate Models. We select 12 pretrained open-source CLIP models via OpenCLIP [12]. In contrast to previous works that typically use fewer than 4 surrogate models [77], our scaled ensemble (up to $T = 12$) is sufficient to yield robust results. The full model list is in Appendix E.

Table 2: Coefficient of determination (R^2) for the log-linear scaling laws of Attack Success Rate across SOTA MLLMs.

Target Model	sea lion	barometer	swing	pickelhaube	fig	lock	computer	snake	Average
Claude 3.5 Sonnet	0.77	0.88	0.86	0.84	0.94	0.94	0.93	0.92	0.94
DeepSeek-VL2	0.89	0.97	0.94	0.98	0.92	0.98	0.99	0.97	0.98
Doubao Seed	0.96	0.88	0.95	0.98	0.84	0.83	0.91	0.82	0.94
GPT-4o	0.96	0.89	1.00	0.92	0.97	0.98	0.92	0.96	0.97
Gemini 2.5 Pro	0.95	0.81	0.98	0.90	1.00	0.93	0.95	0.95	0.97
Llama 3.2 90B-Vision	0.92	0.85	0.97	0.89	0.97	0.93	0.95	0.96	0.96
Qwen3 VL 235B	0.97	0.99	0.97	0.99	0.97	0.97	0.91	0.91	0.97

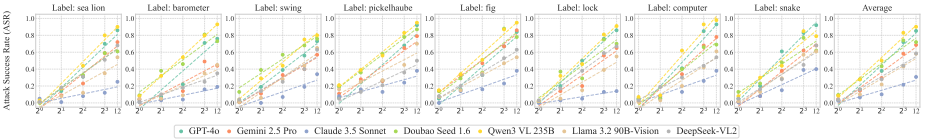


Fig. 5: Verification of the log-linear scaling laws on SOTA MLLMs. We use an ensemble of CLIP models as surrogates, scaling the cardinality T from 1 to 12. All attacks in this figure are generated using $\epsilon = 16/255$. The ASR is measured using our LLM-as-Judge metric (defined in Sec. 5.1). The log-linear scaling laws clearly persist.

Target Models. We select 7 state-of-the-art MLLMs as our black-box targets, including 4 commercial models (GPT-4o [54], Claude-3.5-Sonnet [1], Gemini-2.5-Pro [13], Doubao-Seed-1.6 [7]) and 3 open-source models (Llama-3.2-90B-Vision-Instruct [27], Qwen3-VL-235B-A22B-Instruct [5], and DeepSeek-VL2 [68]).

Dataset and Hyperparameters. We use the same SSA-CWA attack protocol as in Sec. 4.2, but with a simplified dataset for this section. We randomly sample 100 images from the NIPS17 dataset with the same 8 distinct target classes, resulting in 800 unique source-target pairs per ensemble configuration. All experiments in this section are conducted using a unified perturbation budget of $\epsilon = 16/255$ under the ℓ_∞ norm, a common and widely applied setting for practical attacks [16, 51]. Every group of our experiments are conducted under 8 NVIDIA A100 GPUs for less than 24 hours.

Evaluation Metric. We employ a single, unified evaluation metric for all MLLM targets. As MLLMs produce generative outputs instead of classification logits, we adopt the LLM-as-Judge framework from MultiTrust [76]. We leverage GPT-4o to semantically evaluate the target model's description. An attack is considered successful only if GPT-4o judges that (1) the model's description mentions the adversarial target concept and (2) the description does not mention factors indicating the image is of low quality, noisy, or modified. Because this metric requires the target model to generate a coherent description of the adversarial target, it is functionally equivalent to a zero-shot image-captioning task, and the eight target concepts provide a rigorous evaluation of cross-modal robustness. The prompt of our metric is provided in Appendix D.

5.2 Generality and Power of Scaling Laws

Our results, conducted on SOTA MLLMs, deliver two key findings.

Table 3: Attack Success Rate (ASR) on SOTA MLLMs using our full scaled ensemble ($T = 12$) and practical budget ($\epsilon = 16/255$). The ASR is measured using our unified LLM-as-Judge metric. We report the per-target ASR and the average ASR across all 8 target classes against 7 latest generation MLLMs.

Target Model - Label	Sea Lion	Barometer	Swing	Pickelhaube	Fig	Lock	Computer	Snake	Average
GPT-4o	0.86	0.76	0.73	0.92	0.84	0.86	0.93	0.92	**0.85**
Claude-3.5-Sonnet	0.25	0.19	0.34	0.38	0.38	0.14	0.38	0.41	**0.31**
Gemini-2.5-Pro	0.72	0.44	0.57	0.79	0.86	0.65	0.78	0.68	**0.69**
Doubao-Seed-1.6	0.61	0.73	0.76	0.87	0.78	0.70	0.69	0.62	**0.72**
Llama-3.2-90B-Vision	0.54	0.45	0.65	0.70	0.54	0.55	0.61	0.61	**0.58**
Qwen3-VL-235B	0.90	0.93	0.80	0.95	0.95	0.91	0.98	0.79	**0.90**
DeepSeek-VL2	0.68	0.35	0.63	0.50	0.63	0.67	0.54	0.68	**0.59**

The log-linear scaling laws persist. Fig. 5 presents the ASR results on our 7 SOTA MLLM targets as a function of ensemble cardinality (T). Our first key finding is that the log-linear scaling laws, discovered in Sec. 4.3, robustly generalize to this new domain. This outcome confirms that the principle is not an artifact of standard classifiers but a fundamental property of ensemble attacks. Even when attacking the complex, SOTA vision encoders of models, scaling the number of CLIP surrogates yields a predictable scaling law.

Scaling is a powerful and practical weapon. Having verified the scaling laws (Fig. 5), we now examine its practical impact. Table 3 reports the final ASR from our full scaled ensemble ($T = 12$). The results are twofold: they confirm the devastating power of scaling while simultaneously revealing a stark hierarchy in SOTA model robustness. The attack's power is most evident on Qwen3-VL-235B and GPT-4o, which succumb to a 90.0% and 85.0% average ASR, respectively. This performance represents a significant leap in practical capability, surpassing the 40-50% ASRs reported by previous methods on commercial models [16]. However, this effectiveness is not universal. Claude-3.5-Sonnet exhibits exceptional robustness, neutralizing the attack to a mere 31.0% ASR. These results demonstrate that scaling is a powerful and practical tool, and the resulting ASRs serve as a clear benchmark for the diverging robustness of modern MLLMs.

6 Further Analysis and Discussion

In this section, we further demonstrate the robustness of our discovered scaling laws through ablation studies (Sec. 6.1), reconcile our findings with existing benchmarks, and provide a deeper analysis into the semantic nature of the perturbations generated via scaling (Sec. 6.2).

6.1 Ablations on the Scaling Laws' Generality

While our main experiments (Sec. 4) focused on targeted attacks using SSA-CWA, here we verify that the *log-linear scaling laws* are general principles that hold under different attack settings and across various optimizers.

Untargeted Attack Setting. We first validate the scaling laws in the untargeted setting (defined by Eq. 3). To do this, we follow the core experimental settings from Sec. 4.2, using the SSA-CWA algorithm on the same 64-model surrogate pool and NIPS17 dataset. We measure the average ASR against 3 representative target models from Family 1 (Swin-B, ConvNeXt-B, and ResNeXt50). As the untargeted setting is a significantly simpler task, We use a small $\epsilon = 2/255$ to avoid the ASR quickly saturating at 100%, which would obscure the underlying trend. As shown in Fig. 6a, even under untargeted conditions, a clear log-linear trend persists. This confirms that the scaling laws are a general principle of ensemble attacks, independent of the specific attack objective.

Across Different Ensemble Strategies. We now provide a tailored experiment to validate our hypothesis from Sec. 3.2: that the log-linear scaling laws are unlocked only by advanced ensemble strategies (like CWA or SVRE) that resolve gradient conflict, while naïve logit-averaging ensembles fail to scale.

Experimental Design. To test this, we design two families of ensemble strategies and evaluate their scaling performance. All experiments follow the protocol in Sec. 4.2 (targeting `sea lion`) and results are averaged over our standard target classifiers (Family 1 of Sec. 4.2).

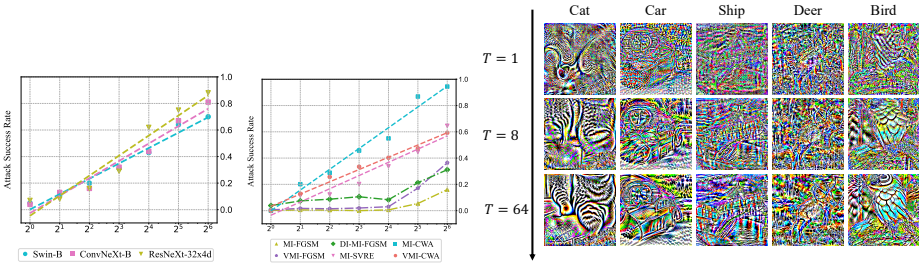
- Family 1 (Logit-Averaging Ensembles): We test the naïve logit-averaging strategy, which we hypothesize will fail to scale. We combine it with various powerful gradient optimizers and input transformations: MI-FGSM [19], DI-MI-FGSM [69], and VMI-FGSM [64].
- Family 2 (Conflict-Resolving Ensembles): We then test advanced ensemble strategies that explicitly align gradients. This family includes MI-CWA [11], VMI-CWA [11, 64], and MI-SVRE [71].

Results and Analysis. The results are shown in Fig. 6b. All strategies in Family 1 fail to exhibit the log-linear scaling laws. While performance can be boosted by components like DI or VMI, their ASR stagnates at a low value. Conversely, strategies in Family 2 demonstrate a clear log-linear scaling law. This behavior mirrors the results of our main SSA-CWA algorithm (from Sec. 4), confirming that the scaling phenomenon is robust across different conflict-resolving ensemble strategies. This experiment confirms our claim: the ability to scale an ensemble attack is contingent on the ensemble optimization strategy.

6.2 The Semantic Nature of Scaled Perturbations

Beyond attack success rates, we investigate *what* the ensemble attack learns as T increases. As shown in Fig. 6c, scaling the ensemble fundamentally changes the *nature* of the resulting perturbation. At $T = 1$, the perturbation is noisy and model-specific. However, as T increases to 64, the perturbation evolves to contain clear, interpretable, and semantic features of the target class (e.g., textures for 'Cat'), while these perturbations remain strictly within the small ℓ_∞ budget.

This is a critical finding that provides a qualitative explanation for the scaling laws. First, scaling T forces the optimizer to discard model-specific, 'non-robust



(a) Untargeted setting. **(b)** Across optimizers. **(c)** Visualizations as T scales.

Fig. 6: Ablations and visualizations of the scaling laws. (a) The log-linear scaling laws persist in the untargeted setting (Eq. 3). (b) The laws hold for advanced optimizers (MI-CWA, VMI-CWA, MI-SVRE) that mitigate gradient conflict, whereas the naïve logit ensemble strategy fails to scale. (c) Visualization of adversarial perturbations (using 5 labels from CIFAR10 [37]). As the ensemble cardinality grows, perturbations evolve from unstructured noise ($T = 1$) into semantically meaningful patterns that resemble features of the target classes ($T = 64$).

features' [34] and instead find a 'common weakness' solution that fools all models. Next, our visualization strongly suggests this common weakness is not just a mathematical artifact but is tied to the *semantic essence* of the target class. The attack learns to inject robust features of the target [14]. This reinforces that our scaling approach is a principled method to distill the shared, transferable features of a concept from a diverse set of models.

7 Conclusion

In this work, we investigate scaling laws for black-box adversarial attacks. We find that while gradient conflict stalls naïve ensembling, advanced optimizers (e.g., CWA) unlock robust log-linear scaling laws. We empirically verify our laws' universality across standard classifiers (Sec. 4.3), SOTA defenses , and MLLMs (Sec. 4.4), and show that scaling distills semantic features (Sec. 6.2). Applying this methodology to SOTA MLLMs (Sec. 5) serves as a powerful benchmark. The scaled attack achieves 85.0% ASR on GPT-4o, while Claude-3.5-Sonnet shows exceptional resilience at only 31.0% ASR. Our findings urge a shift in focus for robustness evaluation: from designing intricate algorithms on small ensembles to understanding the principled and powerful threat of scaling, which is crucial for building and verifying genuinely robust foundation models.

Acknowledgement

This work was supported by NSFC Projects (Nos. 62276149, U2341228).

References

1. Anthropic: The claude 3 model family: Opus, sonnet, haiku. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf (2024), accessed: 2026-02-20
2. Athalye, A., Carlini, N., Wagner, D.: Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In: International conference on machine learning. pp. 274–283. PMLR (2018)
3. Bahri, Y., Dyer, E., Kaplan, J., Lee, J., Sharma, U.: Explaining neural scaling laws. Proceedings of the National Academy of Sciences **121**(27), e2311878121 (2024)
4. Bai, S., et al.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923
5. Bai, S., et al.: Qwen3-vl technical report (2025), https://arxiv.org/abs/2511.21631
6. Bartoldson, B.R., Diffenderfer, J., Parasyris, K., Kailkhura, B.: Adversarial robustness limits via scaling-law and human-alignment studies (2024), https://arxiv.org/abs/2404.09349
7. ByteDance Seed: Seed1.6 tech introduction. https://seed.bytedance.com/en/seed1_6 (2025), accessed: 2026-02-20
8. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 ieee symposium on security and privacy (sp). pp. 39–57. Ieee (2017)
9. Chen, B., Yin, J., Chen, S., Chen, B., Liu, X.: An adaptive model ensemble adversarial attack for boosting adversarial transferability. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4489–4498 (2023)
10. Chen, H., Dong, Y., Wei, Z., Su, H., Zhu, J.: Towards the worst-case robustness of large language models. arXiv preprint arXiv:2501.19040 (2025)
11. Chen, H., Zhang, Y., Dong, Y., Yang, X., Su, H., Zhu, J.: Rethinking model ensemble in transfer-based adversarial attacks. In: The Twelfth International Conference on Learning Representations (2024)
12. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2818–2829 (2023)
13. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
14. Dai, Z., Gifford, D.: Fundamental limits on the robustness of image classifiers. In: The Eleventh International Conference on Learning Representations (2022)
15. Dietterich, T.G.: Ensemble methods in machine learning. In: International workshop on multiple classifier systems. pp. 1–15. Springer (2000)
16. Dong, Y., Chen, H., Chen, J., Fang, Z., Yang, X., Zhang, Y., Tian, Y., Su, H., Zhu, J.: How robust is google's bard to adversarial image attacks? arXiv preprint arXiv:2309.11751 (2023)
17. Dong, Y., Cheng, S., Pang, T., Su, H., Zhu, J.: Query-efficient black-box adversarial attacks guided by a transfer-based prior. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(12), 9536–9548 (2021)
18. Dong, Y., Fu, Q.A., Yang, X., Pang, T., Su, H., Xiao, Z., Zhu, J.: Benchmarking adversarial robustness on image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 321–331 (2020)

19. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J.: Boosting adversarial attacks with momentum. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9185–9193 (2018)
20. Dong, Y., Pang, T., Su, H., Zhu, J.: Evading defenses to transferable adversarial examples by translation-invariant attacks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4312–4321 (2019)
21. Dong, Y., Su, H., Zhu, J., Bao, F.: Towards interpretable deep neural networks by leveraging adversarial examples. arXiv preprint arXiv:1708.05493 (2017)
22. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
23. Fort, S., Lakshminarayanan, B.: Ensemble everything everywhere: Multi-scale aggregation for adversarial robustness (2024), https://arxiv.org/abs/2408.05446
24. Gao, L., Zhang, Q., Song, J., Liu, X., Shen, H.T.: Patch-wise attack for fooling deep neural network. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16. pp. 307–322. Springer (2020)
25. Gontijo-Lopes, R., Dauphin, Y., Cubuk, E.D.: No one representation to rule them all: Overlapping features of training methods. arXiv preprint arXiv:2110.12899 (2021)
26. Goodfellow, I.J.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
27. Grattafiori, A., et al.: The llama 3 herd of models (2024), https://arxiv.org/abs/2407.21783
28. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
29. Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M.M.A., Yang, Y., Zhou, Y.: Deep learning scaling is predictable, empirically. arXiv preprint arXiv:1712.00409 (2017)
30. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
31. Huang, H., Erfani, S., Li, Y., Ma, X., Bailey, J.: X-transfer attacks: Towards super transferable adversarial attacks on clip (2025), https://arxiv.org/abs/2505.05528
32. Huang, H., Chen, Z., Chen, H., Wang, Y., Zhang, K.: T-sea: Transfer-based self-ensemble attack on object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20514–20523 (2023)
33. Ilyas, A., Engstrom, L., Athalye, A., Lin, J.: Black-box adversarial attacks with limited queries and information. In: International conference on machine learning. pp. 2137–2146. PMLR (2018)
34. Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., Madry, A.: Adversarial examples are not bugs, they are features (2019), https://arxiv.org/abs/1905.02175
35. K, A., Hamner, B., Goodfellow, I.: Nips 2017: Non-targeted adversarial attack. https://kaggle.com/competitions/nips-2017-non-targeted-adversarial-attack (2017), kaggle
36. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models (2020), https://arxiv.org/abs/2001.08361

37. Krizhevsky, A.: Learning multiple layers of features from tiny images. University of Toronto (05 2012)
38. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)
39. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236 (2016)
40. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 951–958. IEEE (2009)
41. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)
42. Li, Q., Guo, Y., Zuo, W., Chen, H.: Making substitute models more bayesian can enhance transferability of adversarial examples. arXiv preprint arXiv:2302.05086 (2023)
43. Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., Zhu, J.: Defense against adversarial attacks using high-level representation guided denoiser (2018), https://arxiv.org/abs/1712.02976
44. Lin, J., Song, C., He, K., Wang, L., Hopcroft, J.E.: Nesterov accelerated gradient and scale invariance for adversarial attacks. arXiv preprint arXiv:1908.06281 (2019)
45. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023)
46. Liu, L., Wei, W., Chow, K.H., Loper, M., Gursoy, E., Truex, S., Wu, Y.: Deep neural network ensembles against deception: Ensemble diversity, accuracy and robustness. In: 2019 IEEE 16th international conference on mobile ad hoc and sensor systems (MASS). pp. 274–282. IEEE (2019)
47. Liu, Y., Chen, X., Liu, C., Song, D.: Delving into transferable adversarial examples and black-box attacks. arXiv preprint arXiv:1611.02770 (2016)
48. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
49. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
50. Long, Y., Zhang, Q., Zeng, B., Gao, L., Liu, X., Zhang, J., Song, J.: Frequency domain model augmentation for adversarial attack. In: European conference on computer vision. pp. 549–566. Springer (2022)
51. Luo, H., Gu, J., Liu, F., Torr, P.: An image is worth 1000 lies: Adversarial transferability across prompts on vision-language models (2024), https://arxiv.org/abs/2403.09766
52. Madry, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
53. Nie, W., Guo, B., Huang, Y., Xiao, C., Vahdat, A., Anandkumar, A.: Diffusion models for adversarial purification (2022), https://arxiv.org/abs/2205.07460
54. OpenAI: Gpt-4 technical report (2024), https://arxiv.org/abs/2303.08774
55. Palatucci, M., Pomerleau, D., Hinton, G.E., Mitchell, T.M.: Zero-shot learning with semantic output codes. Advances in neural information processing systems **22** (2009)
56. Paszke, A., et al.: Pytorch: An imperative style, high-performance deep learning library (2019), https://arxiv.org/abs/1912.01703

57. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021), https://arxiv.org/abs/2103.00020
58. Sharma, U., Kaplan, J.: Scaling laws from the data manifold dimension. Journal of Machine Learning Research **23**(9), 1–34 (2022)
59. Sun, X., et al.: Hunyuan-large: An open-source moe model with 52 billion activated parameters by tencent (2024), https://arxiv.org/abs/2411.02265
60. Szegedy, C.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
61. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
62. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021)
63. Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., McDaniel, P.: Ensemble adversarial training: Attacks and defenses. arXiv preprint arXiv:1705.07204 (2017)
64. Wang, X., He, K.: Enhancing the transferability of adversarial attacks through variance tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1924–1933 (2021)
65. Wei, Z., Wang, Y., Li, A., Mo, Y., Wang, Y.: Jailbreak and guard aligned language models with only few in-context demonstrations. arXiv preprint arXiv:2310.06387 (2023)
66. Wightman, R.: Pytorch image models. https://github.com/huggingface/pytorch-image-models (2019). https://doi.org/10.5281/zenodo.4414861
67. Wortsman, M., Ilharco, G., Gadre, S.Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A.S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., et al.: Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In: International conference on machine learning. pp. 23965–23998. PMLR (2022)
68. Wu, Z., et al.: Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding (2024), https://arxiv.org/abs/2412.10302
69. Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z., Yuille, A.L.: Improving transferability of adversarial examples with input diversity. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2730–2739 (2019)
70. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017)
71. Xiong, Y., Lin, J., Zhang, M., Hopcroft, J.E., He, K.: Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14983–14992 (2022)
72. Yang, X., Dong, Y., Pang, T., Su, H., Zhu, J.: Boosting transferability of targeted adversarial examples via hierarchical generative networks (2022), https://arxiv.org/abs/2107.01809
73. Yao, W., Zhang, Z., Tang, H., Liu, Y.: Understanding model ensemble in transferable adversarial attack (2025), https://arxiv.org/abs/2410.06851

74. Zhang, C., Benz, P., Imtiaz, T., Kweon, I.S.: Understanding adversarial examples from the mutual influence of images and perturbations (2020), https://arxiv.org/abs/2007.06189
75. Zhang, Y., Hu, S., Zhang, L.Y., Shi, J., Li, M., Liu, X., Wan, W., Jin, H.: Why does little robustness help? a further step towards understanding adversarial transferability. In: 2024 IEEE Symposium on Security and Privacy (SP). pp. 3365–3384. IEEE (2024)
76. Zhang, Y., Huang, Y., Sun, Y., Liu, C., Zhao, Z., Fang, Z., Wang, Y., Chen, H., Yang, X., Wei, X., Su, H., Dong, Y., Zhu, J.: Benchmarking trustworthiness of multimodal large language models: A comprehensive study (2024)
77. Zhao, Y., Pang, T., Du, C., Yang, X., Li, C., Cheung, N.M., Lin, M.: On evaluating adversarial robustness of large vision-language models (2023), https://arxiv.org/abs/2305.16934
78. Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J.Z., Fredrikson, M.: Universal and transferable adversarial attacks on aligned language models. arXiv preprint arXiv:2307.15043 (2023)