

# CAMH: Advancing Model Hijacking Attack in Machine Learning

Xing He<sup>1\*</sup>, Jiahao Chen<sup>1\*</sup>, Yuwen Pu<sup>1</sup>, Qingming Li<sup>1†</sup>, Chunyi Zhou<sup>1</sup>,  
Yingcai Wu<sup>1</sup>, Jinbao Li<sup>2,3</sup>, Shouling Ji<sup>1</sup>

<sup>1</sup>College of Computer Science and Technology, Zhejiang University

<sup>2</sup>Shandong Artificial Intelligence Institute

<sup>3</sup>School of Mathematics and Statistics, Qilu University of Technology

## Abstract

In the burgeoning domain of machine learning, the reliance on third-party services for model training and the adoption of pre-trained models have surged. However, this reliance introduces vulnerabilities to model hijacking attacks, where adversaries manipulate models to perform unintended tasks, leading to significant security and ethical concerns, like turning an ordinary image classifier into a tool for detecting faces in pornographic content, all without the model owner’s knowledge. This paper introduces Category-Agnostic Model Hijacking (CAMH), a novel model hijacking attack method capable of addressing the challenges of *class number mismatch*, *data distribution divergence*, and *performance balance between the original and hijacking tasks*. CAMH incorporates synchronized training layers, random noise optimization, and a dual-loop optimization approach to ensure minimal impact on the original task’s performance while effectively executing the hijacking task. We evaluate CAMH across multiple benchmark datasets and network architectures, demonstrating its potent attack effectiveness while ensuring minimal degradation in the performance of the original task.

## 1 Introduction

In today’s digital era, machine learning (ML) models have become essential tools in critical fields such as finance, healthcare, and autonomous driving (Li et al. 2019; Zhang et al. 2020; Liu et al. 2020a). Training machine learning models is expensive, as it requires costly computational resources and precise hyperparameter tuning by domain experts (Strubell, Ganesh, and McCallum 2020). To reduce these costs, organizations are increasingly turning to third-party services like Google AutoML (Google 2017), Amazon SageMaker (Amazon 2017), or Microsoft Azure ML (Microsoft 2014) for neural network training or downloading pre-trained models from repositories such as Hugging Face (Clement et al. 2016) or Model Zoo, especially those trained on popular datasets (Jing 2021).

However, this practice exposes models to real-world security risks, particularly model hijacking attacks. Model hijacking attacks enable adversaries to take control of a target model and compel it to perform completely different

tasks than its original one. Successfully executing a model hijacking attack can lead to two principal risks: accountability risk and parasitic computing risk (Salem, Backes, and Zhang 2022).

The accountability risk means that attackers could use the model to conduct illegal or unethical tasks, such as turning an ordinary image classifier into a tool for identifying faces in pornographic content, all without the model owner’s knowledge. This could lead to the model owner being wrongly accused of providing illegal services despite not being directly involved. Additionally, the owners of hijacked models face the risk of parasitic computing, where attackers exploit the hijacked model to run their own applications without incurring the costs of operating, and maintaining the models themselves, which can be prohibitively expensive in some regions (see detailed scenarios in Section 3.1).

Salem et al. (Salem, Backes, and Zhang 2022) first introduced the concept of model hijacking attacks. They demonstrated during training how a target model could be redirected, such as transforming an image classifier into a facial recognition tool. These attacks degrade model performance and expose model owners to legal and ethical risks. Si et al. (Si et al. 2023) further expanded this research domain by introducing new attack strategies applicable to complex scenarios beyond image classification, including data manipulation and task transfer. The hijacking attacks are an emerging threat to model security, and the only two current methods primarily focus on modifying poisoning operations within the training dataset, which means that the model’s output remains constrained to the number of categories defined by the original task. Consequently, these methods impose significant restrictions on the flexibility of hijacking tasks, such as requiring the number of categories in the hijacking task not to exceed those in the original task and ensuring that the data distribution of the original task does not differ significantly. If these conditions are not met, it becomes impossible to balance the hijacking task and the original task’s performance, resulting in loss of concealment or hijacking failure.

We hope that model hijacking attacks will no longer be limited by previous methods. In the implementation process, we summarize three major challenges:

- **C1:** *How do we solve the limitation that the number of hijacking task categories cannot exceed the number of original task categories?*

\*These authors contributed equally.

†Corresponding author

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

- **C2:** *How to ensure the effectiveness of model hijacking attacks given different data distributions?*
- **C3:** *How do we balance the performance of the original task with the execution of the hijacking task?*

To address these challenges, we propose a novel model hijacking attack method—CAMH. We’ve integrated synchronized training layers to handle different class numbers in model training. We’ve also optimized noise to adjust for data distribution variations. Finally, we use a dual-loop approach to keep the original task’s performance stable. To assess the effectiveness of the CAMH method comprehensively, we have established two evaluation metrics: Camouflage Ratio (CR) and Exploitability Ratio (ER). The Camouflage Ratio gauges the performance discrepancy between the hijacked model and the benign model on the original task, whereas the Exploitability Ratio measures the performance gap between the hijacked model and the benign model on the hijacking task. We evaluate the effectiveness of CAMH on multiple benchmark tasks (including MNIST, SVHN, GT-SRB, CIFAR10, and CIFAR100) across various network architectures (including ResNet18 and ResNet34). The experimental results demonstrate that CAMH successfully executes hijacking tasks, with the ER values exceeding 85% for most datasets and the CR maintained at approximately 98% for all datasets.

The main contributions of this paper include:

- We propose a novel model hijacking attack method, CAMH, which, for the first time, overcomes the limitation that the number of categories in the hijacked dataset cannot exceed those in the original dataset.
- We proposed synchronized training layers, random noise optimization, and a dual-loop optimization approach to address challenges such as class number mismatch, data distribution divergence, and balancing performance between the original and hijacked tasks.
- We validated CAMH’s effectiveness through experiments on various benchmark datasets and network architectures. Even when the hijacking dataset’s categories were four times greater or the hijacking data volume was as low as 30%, the ER exceeded 85%, confirming CAMH’s robustness.

## 2 Related Work

In this section, we review some of the related works. We divide the related works into backdoor attacks in machine learning and model hijacking attacks. Backdoor attacks can be seen as a specific instance of model hijacking attacks. The research into backdoor attacks helps us to deeply comprehend the intrinsic mechanisms of model manipulation and provides a basis for mastering more widespread strategies for model hijacking. We start with the backdoor attacks, followed by model hijacking attacks.

### 2.1 Backdoor Attacks in Machine Learning

Backdoor attacks pose a significant security threat in the field of machine learning, where the adversary manipulates the target model’s training to backdoor it (Li et al. 2022).

The backdooring behavior is usually assigned with a trigger, which is when inserted in any input sample, the target model predicts a specified label. Gu et al. (Gu, Dolan-Gavitt, and Garg 2017) introduced BadNets, the first backdoor attack against machine learning. BadNets uses a white square at the corner of the images as a trigger to misclassify the backdoored inputs to a specific label. Salem et al. (Salem et al. 2022) later proposed a dynamic backdoor, where instead of using a fixed trigger, they used a dynamic one. Another similar attack is the Trojan attack (Liu et al. 2017; Cheng et al. 2021). This attack simplifies the assumptions of the backdoor attack by not assuming the knowledge of any sample from the distribution of the target model’s training dataset. There also exist multiple backdoor attacks against Natural Language Processing (NLP) models (Chen et al. 2021), federated learning (Wang et al. 2020), video recognition (Zhao et al. 2020), transfer Learning (Yao et al. 2019), diffusion model (Ho, Jain, and Abbeel 2020), Chatgpt (Shi et al. 2023) and others (Saha, Subramanya, and Pirsivash 2020; Rakin, He, and Fan 2020; Liu et al. 2020b; Li et al. 2021; Hayase and Oh 2022).

The backdoor attack can be considered a specific instance of the model hijacking attack by classifying the backdoored samples as the hijacking dataset. However, our model hijacking attack is more general, i.e., it poisons the model to implement a completely different task.

### 2.2 Model Hijacking Attack

Model hijacking attacks, in which attackers attempt to manipulate a target model to perform unintended tasks, are an emerging attack that can lead to serious security and liability issues. Salem et al. (Salem, Backes, and Zhang 2022) introduced the concept of model hijacking attacks within the realm of computer vision tasks. Their work is similar to SSBA (Li et al. 2021) have proposed techniques to enhance the stealthiness of backdoor attacks, thereby reducing the effectiveness of backdoor detection mechanisms. Expanding the scope of model hijacking attacks, Si et al. (Si et al. 2023) introduced a novel attack named Ditto, which extends this threat to text generation and classification models.

While both Salem et al. and Si et al. have significantly advanced the understanding of model hijacking attacks, their work exhibits several limitations. First, the effectiveness of these attacks is highly contingent upon the adversary’s capability to generate stealthy modifications to the training data, which may be detectable by robust data monitoring and anomaly detection systems. Second, their approaches impose relatively stringent constraints on the hijacking dataset. For instance, the number of categories in the hijacking dataset must not exceed that of the original dataset. Third, they do not address emerging scenarios such as outsourcing and model marketplaces.

In our study, we introduce CAMH, a novel model hijacking attack designed for new scenarios. CAMH demonstrates exceptional attack performance even under constraints of multi-class hijacking and significant data discrepancies while avoiding the need for deceptive modifications to the training dataset.

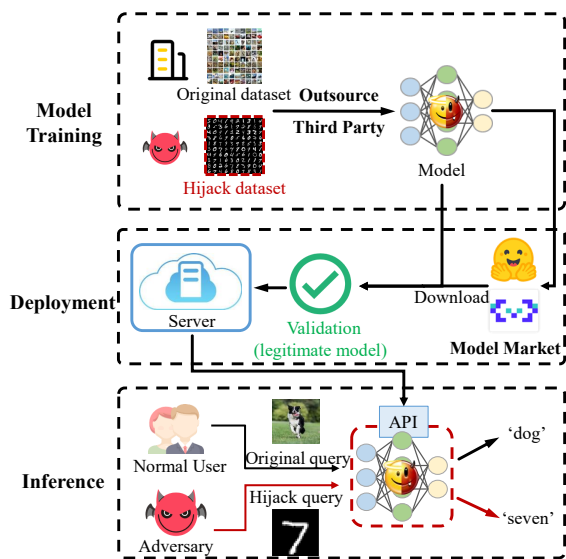


Figure 1: CAMH attacks explored in this paper.

### 3 CAMH

In this section, we present various techniques for model hijacking attacks. We first introduce the threat model and two attack scenarios. Then, we propose three specific CAMH techniques: Multi-Class Hijacking Task Design, Noise Optimization Technique, and Dual-loop Optimization.

#### 3.1 Threat Model and Attack Scenario

**Threat Model.** Model hijacking attacks can be widely applicable to any real-world scenario where model owners can independently train models, such as outsourcing, model marketplaces, federated learning (Bonawitz et al. 2019), etc. In the attack scenario proposed in this study, adversaries have complete control over the model training process, and obtain the logits output by the model. We assume adversaries possess a hijacked dataset, which they use to covertly embed the model with hijacking tasks. Whether through outsourced model training or adversaries independently deploying malicious models in the model marketplace, adversaries can access the training dataset, with the distinction being that the former is provided by the victim company and the latter is owned by the adversaries themselves. Additionally, attackers do not need to create disguised datasets as in previous work, nor do hijacked datasets need to resemble the visual features of the original dataset.

As shown in Figure 1, we show two attack scenarios, namely outsourcing and model market, which we will explain in detail below.

**Model Hijacking in Outsourcing Scenarios.** In the current field of machine learning, model companies often outsource model training tasks to external entities (Li et al. 2020), which may possess unique data resources or computational capabilities. While this collaborative model can bring technical advantages and cost-effectiveness, it also introduces potential security risks. The attack scenario dis-

cussed in this paper involves the outsourcing training process, where if the training party is malicious, they can leverage our model hijacking attack method to clandestinely implant an additional hijacking task in the model while completing the tasks assigned by the model company.

Specifically, the malicious training party would cleverly incorporate their own hijacking data alongside the data provided by the model company during the training process. These hijacking data are carefully crafted to have a negligible impact on the model’s performance for its original task, yet they are sufficient for the model to learn to execute a completely different hijacking task. Once the model has been trained and is publicly released by the model company via APIs (Tramèr et al. 2016), attackers can execute the hijacking task within the model through these API calls, leading it to perform actions that are illegal or unethical.

For example, attackers could hijack an image classification model to perform sensitive tasks such as facial recognition or sentiment analysis when receiving specific inputs, tasks that are illegal without explicit authorization. This not only leads to wastage of computational resources and maintenance costs for the model company but also exposes them to legal liability and reputational damage due to the model’s inappropriate behavior.

**Covert Hijacking in Model Marketplaces.** In the field of machine learning, model marketplaces such as Hugging Face (Clement et al. 2016) or the Model Scope (Alibaba 2022) have become crucial channels for sharing and distributing pre-trained models. These platforms offer users abundant resources, enabling them to rapidly deploy state-of-the-art machine learning models without needing to train from scratch. However, this convenience also brings new security risks (Yuan et al. 2022; Feng and Tramèr 2024).

In our envisioned attack scenario, attackers exploit these model marketplaces as distribution channels to propagate models with covert hijacking tasks. Attackers first train a model through standard procedures to perform a seemingly useful original task (e.g., image classification or language translation) while training with the hijacking dataset, the hijacking task is stealthily incorporated. Subsequently, attackers package this model with the covert hijacking task as one solely possessing original task functionality and upload it to model marketplaces for distribution.

During promotion, attackers deliberately conceal the hijacking nature of the model, emphasizing only its performance and advantages in the original task. Once other users or companies download and deploy this model, especially when they provide it with API access, attackers can utilize this API to remotely activate and execute the hidden hijacking task embedded within the model. This covert hijacking task could involve illegal or unethical activities, such as unauthorized data collection, sensitive information classification, or other malicious behaviors.

This form of attack not only poses risks to the model’s end-users but also has the potential to severely damage the reputation and trustworthiness of model marketplaces. Users may face legal liabilities and ethical reproach for employing seemingly legitimate but actually hijacked models.

### 3.2 Design of CAMH

As highlighted in Section 1, the realm of model hijacking confronts three principal challenges, which can be succinctly encapsulated as the mismatch in class numbers, divergence in data distributions, and the delicate balance of performance. In direct response to these intricacies, we have engineered a novel assault strategy aimed at accomplishing model hijacking. This section delineates the intricate design considerations that underpin our approach.

**Multi-Class Hijacking Task Design.** To address the limitation that the number of hijacking task categories cannot exceed the number of original task categories ( $C_1$ ), our CAMH approach includes an innovative Multi-Class Hijacking Task Design. This design introduces a Synchronized Optimization Layer (SOL) within the model architecture. Positioned after the output classification layer, the SOL is a fully connected layer that adeptly transforms the output logits from the original classification task into a space that aligns with the hijacking task’s output dimensions. By leveraging this layer, our method gracefully navigates the complexities arising from an imbalanced number of classes, enabling the seamless integration of a hijacking task into the model’s learning process. This layer is specifically designed to manage the data mapping for the hijacking task while leaving the data for the original task unaffected by its processing. This strategic design ensures that the performance of the original task remains stable and uncompromised.

Specifically, the mathematical expression of the SOL can be described as follows:

$$\begin{aligned} f_{\theta}(x) : x \in \mathbb{R}^{B \times C \times H \times W} &\longrightarrow y \in \mathbb{R}^{B \times C_1} \\ h_{\hat{\theta}}(y) : y \in \mathbb{R}^{B \times C_1} &\longrightarrow z \in \mathbb{R}^{B \times C_2} \end{aligned} \quad (1)$$

Here,  $C_1$  and  $C_2$  represent the class cardinalities of the original and hijacking datasets, respectively. The mapping function  $f_{\theta}$  represents the model’s internal mechanism for processing the input data  $x$ , which are the input features or images the model receives, whereas  $h$  delineates the intermediate mapping within a layer. The term  $B$  signifies the batch size, with  $C$ ,  $H$ , and  $W$  representing the channel count, height, and width of the image matrix, respectively. The output  $y$  is the model’s prediction for the original task, providing a set of class probabilities or scores based on  $x$ . Finally,  $z$  corresponds to the model’s output for the hijacking task after being processed by the SOL.

We define the parameters of the SOL as  $\hat{\theta}$ , and  $\delta$  represents noise that we will add to the hijacking sample to align its data distribution as closely as possible with the original sample, which we will introduce in detail next. Therefore,  $f_{\theta}$  and  $h_{\hat{\theta}}$  represent the model and layer. To this end, our objective can be formalized as:

$$\hat{\theta} = \arg \min_{\hat{\theta}} [\mathcal{L}(\mathcal{D}_H \oplus \delta, \mathcal{F})] \quad (2)$$

where  $\mathcal{F} = h_{\hat{\theta}} \circ f_{\theta}$ ,  $\mathcal{D}_H \oplus \delta = \{x_i + \delta | \forall x_i \in \mathcal{D}_H\}$ ,  $\mathcal{L}$  and  $\mathcal{D}_H$  denoting the loss function and hijacking datasets.

The model hijacking process leverages the SOL, a component intentionally excluded from the public release to

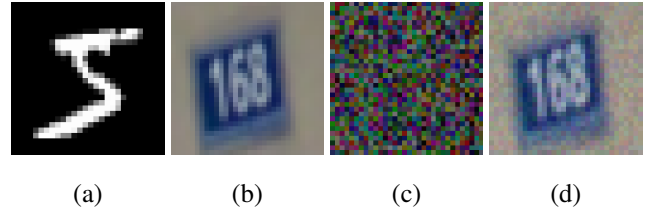


Figure 2: Optimized Noise Visualization. The figure illustrates the process of model hijacking where the original task of the model is based on the MNIST dataset, and the model is utilized for a hijacking task using the SVHN dataset. (a) an image from MNIST. (b) an image from SVHN. (c) the optimized noise. (d) an image from SVHN with added noise.

maintain the model’s legitimate functionality. Under normal circumstances, the model outputs  $C_1$  in response to user queries. However, an adversary equipped with the SOL and knowledge of the model’s logits can repurpose the model for a hijacking task.

**Noise Optimization Technique.** To address  $C_2$ , which is to ensure the effectiveness of model hijacking attacks under different data distributions, we propose a noise optimization technique, which meticulously calibrates the perturbations introduced into the training dataset. The primary goal of this technique is to adjust the data distribution of the hijacking dataset to be more aligned with that of the original dataset, thereby enhancing the model’s ability to generalize and perform well on both datasets. This technique is undergirded by an optimization algorithm that seeks to identify the optimal noise pattern  $\delta^*$  that minimally perturbs the model’s performance on the original task while instilling the capacity to execute the hijacking task. The optimization process of noise must ensure the performance of both the original task and the hijacking task. Thus, we have:

$$\delta^* = \arg \min_{\delta} [\mathcal{L}(\mathcal{D}_M, f_{\theta}) + \mathcal{L}(\mathcal{D}_H \oplus \delta, \mathcal{F})] \quad (3)$$

where  $\mathcal{D}_M$  denotes the original dataset. The formulation encapsulates the dual objectives of preserving the original task’s fidelity and instilling the hijacking task’s functionality.

To further elucidate the process of model hijacking, we present in Figure 2 a visual representation of the optimized noise that enables the model to perform the hijacking task effectively.

In the preceding sections and as evidenced by Figure 2d, there is no imperative to extensively obfuscate the physical appearance of the hijacking dataset. This is attributable to the fact that the dataset remains under the adversary’s control throughout its lifecycle, thereby rendering unnecessary any attempts to deceive the victim within the physical realm. In the inference phase, the adversary only needs to add the optimized custom noise to the input image when performing the hijacking task.

**Dual-loop Optimization.** Our proposed solution is dual-loop optimization to tackle  $C_3$ , balancing the performance between the original task and the hijacking task. Algorithm

1 in the Appendix shows the dual-loop optimization Training. Within the CAMH framework, this approach works by refining the model’s parameters in a synergistic manner, simultaneously enhancing performance for both the original task and the hijacking task. The outer loop focuses on optimizing the original task model parameters  $\theta$ , while the inner loop fine-tunes the parameters  $\hat{\theta}$  for the Synchronized Optimization Layer and the original task model parameters  $\theta$ , incorporating the optimized noise  $\delta^*$ . This dual-loop mechanism ensures that the model excels in its original function while effectively learning the hijacking task, all with minimal additional computational demand.

## 4 Experiment

In this section, we present our experimental results. We start by introducing our datasets and evaluation settings. Next, we evaluate our CAMH attacks. Finally, we study the impact of some of the hyperparameters in our model hijacking attack.

### 4.1 Datasets Description

We mainly use MNIST (LeCun 1998), SVHN (Netzer et al. 2011), GTSRB (Stallkamp et al. 2011), CIFAR10 (Krizhevsky, Hinton et al. 2009) and CIFARm datasets for experiments. The specific descriptions of MNIST, SVHN, GTSRB, and CIFAR10 are shown in the Appendix. CIFARm is a dataset we defined ourselves. CIFARm denotes the datasets derived from CIFAR100 (Krizhevsky, Hinton et al. 2009). These datasets are formed by randomly selecting  $m$  out of the 100 classes available in CIFAR100. The CIFAR100 dataset comprises 100 distinct classes, each class encompassing precisely 600 color images of 32×32 pixel resolution, amounting to a grand total of 60,000 images. By randomly choosing  $m$  classes from these 100, we created the CIFARm datasets where  $m$  can vary (e.g., 20, 40, 60, 80), simulating tasks of different classification scales. Each CIFARm dataset retains the original image dimensions and color channels of CIFAR100 while reducing the number of classes. This reduction in class count implies less information for models to handle during training and inference. This design allows us to control model complexity and training difficulty while assessing the model’s sensitivity to adversarial attacks across varying class sizes.

### 4.2 Model Structure and Training Parameters

To validate the generality and effectiveness of our CAMH method, we conducted experiments using the ResNet18 and ResNet34 (He et al. 2016) models, both of which are original models from Torchvision. For ResNet18, we employed 150 training epochs using the SGD optimizer with an initial learning rate of 0.1 and a batch size of 64. Considering the increased complexity and susceptibility to overfitting of the ResNet34 model, we introduced a dropout rate of 0.4 during training and extended the training epochs to 200, other settings remain the same as ResNet18. This configuration ensures the model achieves high accuracy on the original task while providing a stable foundation for subsequent hijacking attack experiments.

### 4.3 Evaluation Metrics

To evaluate the performance of our CAMH attack, we used two metrics, namely the Camouflage Ratio and the Exploitability Ratio.

**Camouflage Ratio (CR):** The Camouflage Ratio measures the performance of the hijacked model on the original task dataset  $\mathcal{D}_M$ . It is defined as the ratio of the accuracy of the model on  $\mathcal{D}_M$  after hijacking to the accuracy of the original model on the same  $\mathcal{D}_M$ :

$$CR = \frac{ACC_{h2o}}{ACC_{o2o}} \quad (4)$$

Here,  $ACC_{h2o}$  is the accuracy of the hijacked model on the original task dataset, and  $ACC_{o2o}$  is the accuracy of the benign model only trained on the original dataset.

The CR reflects how closely the performance of the hijacked model approaches that of the clean model on a specific original task dataset. As the CR approaches 1, it signifies that hijacked models possess stronger camouflage capabilities, thereby increasing the likelihood of successfully evading detection.

**Exploitability Ratio (ER) :** The Exploitability Ratio measures the performance of the hijacked model on the hijacking dataset  $\mathcal{D}_H$ . It is defined as the ratio of the accuracy of the hijacked model on  $\mathcal{D}_H$  to the accuracy of the benign model on the corresponding  $\mathcal{D}_M$ :

$$ER = \frac{ACC_{h2h}}{ACC_{o2h}} \quad (5)$$

Here,  $ACC_{h2h}$  is the accuracy of the hijacked model on the hijacking task dataset, and  $ACC_{o2h}$  is the accuracy of the benign model is only trained on hijacked dataset.

The ER evaluates how well the hijacked model maintains its performance when transferred to different datasets. A higher ER indicates greater efficiency in utilization when the hijacker employs the model.

These two metrics collectively form the framework for evaluating the model’s performance, enabling a comprehensive understanding of its performance and stability across different environments. In subsequent sections, we will present experimental results based on these metrics and conduct in-depth analyses of the model’s performance.

### 4.4 The Performance of CAMH

In Figure 3, we present the results demonstrating the CR and ER of various datasets after undergoing CAMH attacks. It is noteworthy that all datasets maintain a CR above around 98%, indicating that even after model hijacking, their performance on the original tasks remains at a high level. This high CR value increases the likelihood that victims will select our adversarial models from numerous models in the marketplace or that our models will pass the accuracy acceptance criteria of the first party more easily. In some cases, the CR exceeds 100%. This occurs because the inclusion of the hijacked dataset enhances the model’s generalization ability on the original task, especially when the data distributions of the two datasets are not significantly different.

Simultaneously, ER values indicate the efficiency of the models in performing hijacking tasks, with most datasets

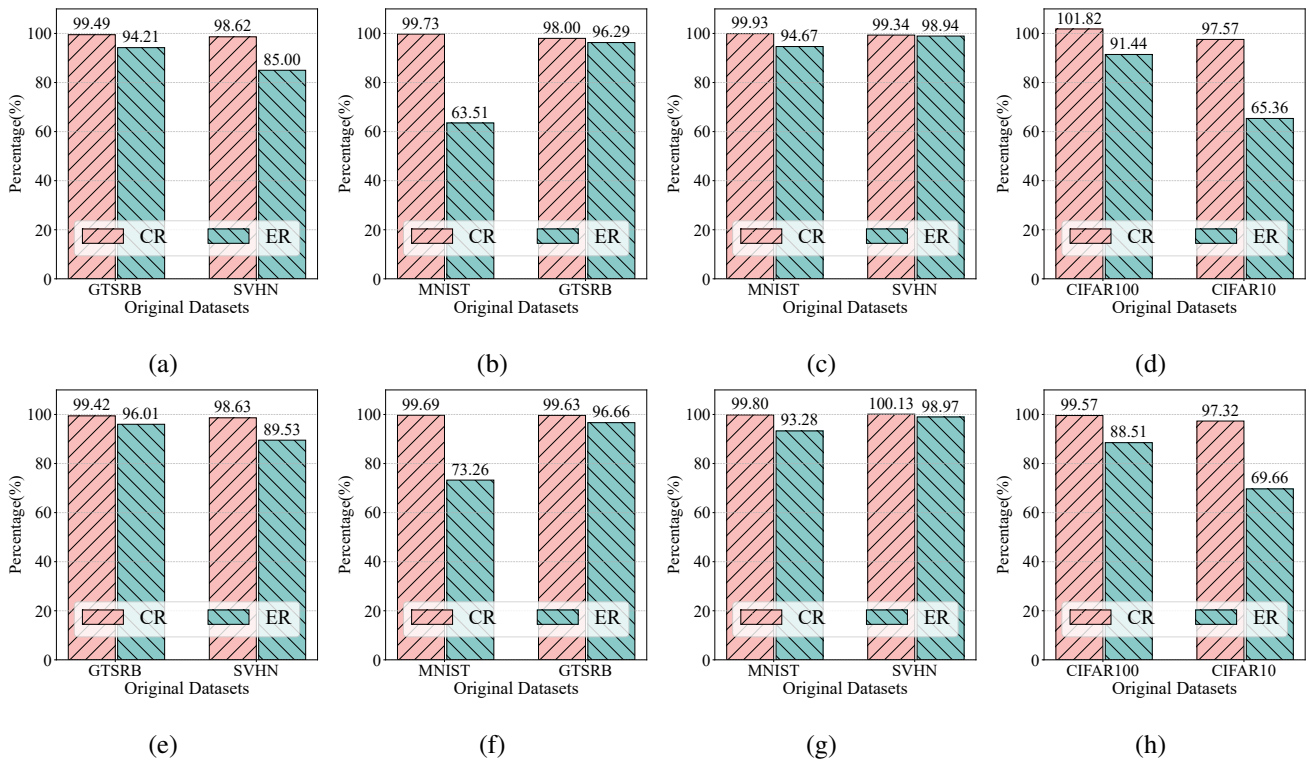


Figure 3: The results of our CAMH Attack. The original datasets are denoted on the x-axis, taking Figure 3 (a) as an example, when the original task is GTSRB and the hijacking task is SVHN, the model’s CR is 99.49%, and ER is 94.21%. When the datasets are exchanged, the CR is 98.62%, and the ER is 85%. (a)-(d) is for ResNet18, (e)-(h) is for ResNet34.

achieving an ER of over 85%. However, in cases where the original dataset is MNIST and the hijacking dataset is GTSRB, or where the original task is CIFAR100, and the hijacking task is CIFAR10, ER values show poorer performance but still exceed 60%. This may be due to the hijacking dataset having a larger number of categories than the original dataset, with significant differences in data distribution that exceed the capability of the hijacking method.

For instance, in Figure 3b, the original dataset has fewer categories (10) compared to the hijacking dataset (43), and there is substantial dissimilarity between MNIST and GTSRB. The combination of these factors results in ER values lower than anticipated. If only one condition is met, as shown in Figure 3a or 3c, the execution of hijacking tasks still falls within CAMH’s hijacking capability range.

Let’s focus on the second and fourth columns. We observe that more complex models exhibit stronger transferability, especially in challenging tasks. For the original task MNIST and the transfer task GTSRB, ResNet34 shows an increase of nearly 10% in ER compared to ResNet18. Similarly, for the original task CIFAR10 and the transfer task CIFAR100, ResNet34 demonstrates a 4% increase in error rate over ResNet18. This phenomenon may be attributed to the more complex models having a greater number of parameters and deeper network structures, allowing them to learn richer feature representations.

#### 4.5 The Impact of the Number of Categories

In this section, we investigate the impact of the number of categories on model hijacking capability. As illustrated in Figure 4, we examine the performance of the original task and the hijacking task across datasets with varying numbers of categories. While maintaining the baseline performance of the original task, we find that the model’s hijacking capability excels when the hijacking task faces fewer categories than the original task, achieving ER values exceeding 90% for both ResNet18 and ResNet34. Conversely, as the number of categories in the adversarial task exceeds that of the original task, ER values decrease, irrespective of whether ResNet18 or ResNet34 is employed. We observe that in such scenarios, ResNet34 outperforms ResNet18, suggesting that larger-scale models may handle more challenging hijacking tasks with greater ease.

#### 4.6 The Impact of Hijacking Data Volume

In this section, we investigate the impact of hijacking data volume on model hijacking tasks. Figure 5 vividly demonstrates that the scale of hijacking data significantly influences the ER of model hijacking tasks. Specifically, as the volume of data increases, both curves exhibit an upward trend. When the volume of hijacking data constitutes only 30% of the total, the model’s ER reaches a satisfactory high level. This phenomenon is particularly pronounced in the

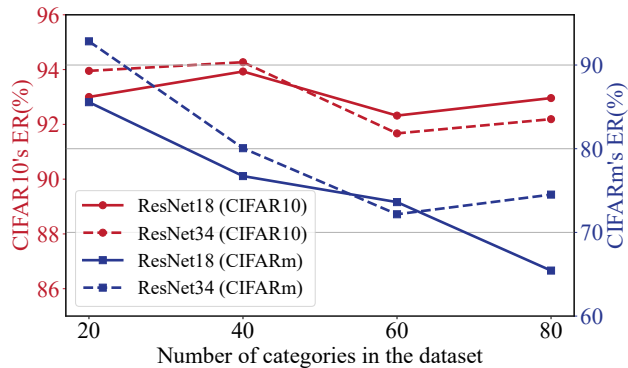


Figure 4: The impact of the number of categories on the ER of hijacking tasks. The red line illustrates the relationship between the ER of the hijacking task and the value of  $m$  when the original task is CIFARm, the hijacking task is CIFAR10, and the CR remains above 95%. The blue line represents the situation when the original and hijacking tasks are swapped. We compared ResNet18 and ResNet34.

SVHN and MNIST datasets: SVHN achieves an ER exceeding 85%, while MNIST approaches an ER over 98%.

This finding is of profound significance as it indicates that even with a relatively modest volume of hijacking data, our model hijacking method achieves efficient task execution. This efficacy is clearly delineated in Figure 5, where the red and blue lines denote the ER variations of the SVHN and MNIST datasets across varying percentages of hijacking data volumes, respectively.

Furthermore, we observe that as the volume of hijacking data continues to increase, the ER improves incrementally, albeit at a diminishing rate. This suggests the presence of an optimal data volume range beyond which the marginal benefit of increased data diminishes in enhancing hijacking task performance.

Our findings imply that even with constrained data volumes, adversaries may proficiently manipulate models to execute illicit tasks via meticulously model hijacking.

#### 4.7 Ablation Study

Model	Noise	SOL	CR	ER
ResNet18	×	×	99.89%	92.23%
	✓	×	99.95%	93.60%
	×	✓	99.95%	91.97%
	✓	✓	<b>99.93%</b>	<b>94.67%</b>
ResNet34	×	×	99.84%	88.37%
	✓	×	99.79%	91.55%
	×	✓	99.79%	91.95%
	✓	✓	<b>99.80%</b>	<b>93.28%</b>

Table 1: The results of the ablation study with MNIST as the original task and SVHN as the hijacking task.

In Section 3.3, we delineate our approach to optimizing a

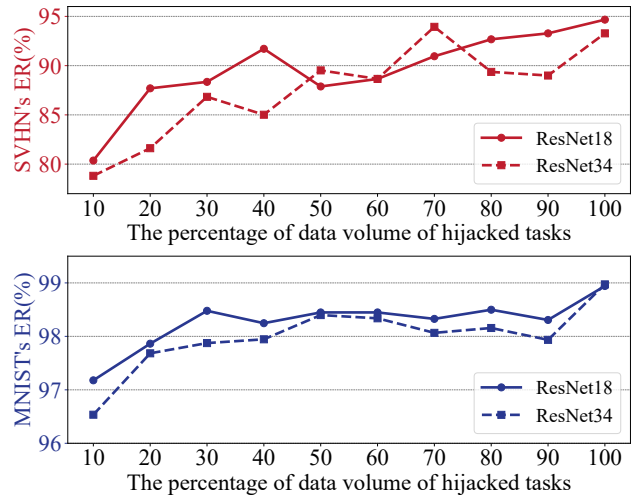


Figure 5: The impact of the percentage of data volume of hijacking tasks on the ER of hijacking tasks. The red line illustrates the relationship between the ER of the hijacking task and the value of  $m$  when the original task is MNIST, the hijacking task is SVHN, and the CR remains above 95%. The blue line represents the situation when the original and hijacking tasks are swapped. We compared ResNet18 and ResNet34.

noise and layer-wise projection mapping. To assess the impact of noise and SOL-based projections, we have undertaken ablation study.

To incorporate the layer into our analysis, we have selected MNIST as the baseline task and SVHN as the target for model hijacking. The findings detailed in Table 1 indicate that when the attacker employs only the optimized noise, the ER for the hijacking task escalates. When the synchronized layer is used exclusively, there is an improvement in performance. The hijacking performance of the model is notably enhanced when both strategies are combined, culminating in the highest ER. It is noteworthy that, due to the consistently high CR, our impact on the original task remains minimal across all scenarios, demonstrating strong camouflage capabilities.

## 5 Conclusion

In this paper, we introduce a Category-Agnostic Model Hijacking (CAMH) attack for the first time. The CAMH method effectively solves the key challenges in model hijacking through innovative multi-category hijacking task design, noise optimization technology, and double-loop optimization. Experimental results show that CAMH can maintain the high performance of the original task while successfully performing the hijacking task even when the number of categories does not match and the data distribution is significantly different. In the future, our research will focus on enhancing detection and defense mechanisms against such attacks and expanding the exploration of the application and impact of model hijacking across various scenarios.

## Acknowledgments

This work was partly supported by the National Key Research and Development Program of China under No. 2022YFB3102100, NSFC under No. U244120033, U24A20336, 62172243 and 62402425, and the Zhejiang Provincial Natural Science Foundation under No. LD24F020002.

## References

- Alibaba, A., DAMO. 2022. ModelScope. <https://community.modelscope.cn/>. [Accessed: 2024-07-30].
- Amazon. 2017. Amazon Sagemaker. <https://aws.amazon.com/sagemaker/>. [Accessed: 2024-07-30].
- Bonawitz, K.; Eichner, H.; Grieskamp, W.; Huba, D.; Ingerman, A.; Ivanov, V.; Kiddon, C.; Konečný, J.; Mazzocchi, S.; McMahan, B.; et al. 2019. Towards federated learning at scale: System design. In *Proceedings of machine learning and systems(MLSYS)*, volume 1, 374–388.
- Chen, X.; Salem, A.; Chen, D.; Backes, M.; Ma, S.; Shen, Q.; Wu, Z.; and Zhang, Y. 2021. Badnl: Backdoor attacks against nlp models with semantic-preserving improvements. In *Proceedings of the Annual Computer Security Applications Conference(ACSAC)*, 554–569.
- Cheng, S.; Liu, Y.; Ma, S.; and Zhang, X. 2021. Deep feature space trojan attack of neural networks by controlled detoxification. In *Proceedings of the AAAI Conference on Artificial Intelligence(AAAI)*, volume 35, 1148–1156.
- Clement, D.; Julien, C.; Thomas, W.; and Anthony, M. 2016. Hugging Face. <https://huggingface.co/>. [Accessed: 2024-07-30].
- Feng, S.; and Tramèr, F. 2024. Privacy Backdoors: Stealing Data with Corrupted Pretrained Models. *arXiv preprint arXiv:2404.00473*.
- Google. 2017. Google Cloud Machine Learning. <https://azure.microsoft.com/>. [Accessed: 2024-07-30].
- Gu, T.; Dolan-Gavitt, B.; and Garg, S. 2017. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*.
- Hayase, J.; and Oh, S. 2022. Few-shot backdoor attacks via neural tangent kernels. *arXiv preprint arXiv:2210.05929*.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the conference on computer vision and pattern recognition(CVPR)*, 770–778.
- Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems(NeurIPS)*, 33: 6840–6851.
- Jing, K., Yu. 2021. Model Zoo - Deep learning code and pretrained models for transfer learning, educational purposes, and more. <https://modelzoo.co/>. [Accessed: 2024-07-30].
- Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.
- LeCun, Y. 1998. The MNIST database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>.
- Li, M.; Chow, S. S.; Hu, S.; Yan, Y.; Shen, C.; and Wang, Q. 2020. Optimizing privacy-preserving outsourced convolutional neural network predictions. *IEEE Transactions on Dependable and Secure Computing(TDSC)*, 19(3): 1592–1604.
- Li, X.; Ma, C.; Wu, B.; He, Z.; and Yang, M.-H. 2019. Target-aware deep tracking. In *Proceedings of the conference on computer vision and pattern recognition(CVPR)*, 1369–1378.
- Li, Y.; Jiang, Y.; Li, Z.; and Xia, S.-T. 2022. Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems(TNNLS)*, 35(1): 5–22.
- Li, Y.; Li, Y.; Wu, B.; Li, L.; He, R.; and Lyu, S. 2021. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the international conference on computer vision(ICCV)*, 16463–16472.
- Liu, L.; Feng, G.; Beautemps, D.; and Zhang, X.-P. 2020a. Re-synchronization using the hand preceding model for multi-modal fusion in automatic continuous cued speech recognition. *IEEE Transactions on Multimedia(TMM)*, 23: 292–305.
- Liu, Y.; Ma, S.; Aafer, Y.; Lee, W.-C.; Zhai, J.; Wang, W.; and Zhang, X. 2017. Trojaning attack on neural networks.
- Liu, Y.; Ma, X.; Bailey, J.; and Lu, F. 2020b. Reflection backdoor: A natural backdoor attack on deep neural networks. In *Proceedings of the European Conference on computer vision(ECCV)*, 182–199. Springer.
- Microsoft. 2014. Microsoft Azure. <https://azure.microsoft.com/>. [Accessed: 2024-07-30].
- Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; Ng, A. Y.; et al. 2011. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, 4. Granada.
- Rakin, A. S.; He, Z.; and Fan, D. 2020. Tbt: Targeted neural network attack with bit trojan. In *Proceedings of the Conference on Computer Vision and Pattern Recognition(CVPR)*, 13198–13207.
- Saha, A.; Subramanya, A.; and Pirsiavash, H. 2020. Hidden trigger backdoor attacks. In *Proceedings of the AAAI conference on artificial intelligence(AAAI)*, volume 34, 11957–11965.
- Salem, A.; Backes, M.; and Zhang, Y. 2022. Get a Model! Model Hijacking Attack Against Machine Learning Models. In *Network and Distributed System Security Symposium(NDSS)*.
- Salem, A.; Wen, R.; Backes, M.; Ma, S.; and Zhang, Y. 2022. Dynamic backdoor attacks against machine learning models. In *IEEE European Symposium on Security and Privacy(EuroS&P)*, 703–718. IEEE.
- Shi, J.; Liu, Y.; Zhou, P.; and Sun, L. 2023. Badgpt: Exploring security vulnerabilities of chatgpt via backdoor attacks to instructgpt. In *Network and Distributed System Security Symposium(NDSS)*.
- Si, W. M.; Backes, M.; Zhang, Y.; and Salem, A. 2023. Two-in-One: A Model Hijacking Attack Against Text Generation Models. In *USENIX Security Symposium (USENIX Security)*, 2223–2240.

- Stallkamp, J.; Schlipsing, M.; Salmen, J.; and Igel, C. 2011. The German traffic sign recognition benchmark: a multi-class classification competition. In *The international joint conference on neural networks(IJCNN)*, 1453–1460. IEEE.
- Strubell, E.; Ganesh, A.; and McCallum, A. 2020. Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI conference on artificial intelligence(AAAI)*, volume 34, 13693–13696.
- Tramèr, F.; Zhang, F.; Juels, A.; Reiter, M. K.; and Ristenpart, T. 2016. Stealing machine learning models via prediction APIs. In *USENIX security symposium (USENIX Security)*, 601–618.
- Wang, H.; Sreenivasan, K.; Rajput, S.; Vishwakarma, H.; Agarwal, S.; Sohn, J.-y.; Lee, K.; and Papailiopoulos, D. 2020. Attack of the tails: Yes, you really can backdoor federated learning. *Advances in Neural Information Processing Systems(NeurIPS)*, 33: 16070–16084.
- Yao, Y.; Li, H.; Zheng, H.; and Zhao, B. Y. 2019. Latent backdoor attacks on deep neural networks. In *Proceedings of the conference on computer and communications security(ACM SIGSAC)*, 2041–2055.
- Yuan, Z.; Wu, F.; Long, Y.; Xiao, C.; and Li, B. 2022. SecretGen: Privacy Recovery on Pre-trained Models via Distribution Discrimination. In *European Conference on Computer Vision(ECCV)*, 139–155. Springer.
- Zhang, Y.; Xu, S.; Wu, B.; Shi, J.; Meng, W.; and Zhang, X. 2020. Unsupervised multi-view constrained convolutional network for accurate depth estimation. *IEEE Transactions on Image Processing(TIP)*, 29: 7019–7031.
- Zhao, S.; Ma, X.; Zheng, X.; Bailey, J.; Chen, J.; and Jiang, Y.-G. 2020. Clean-label backdoor attacks on video recognition models. In *Proceedings of the conference on computer vision and pattern recognition(CVPR)*, 14443–14452.