# Adversarial Knowledge Extraction via Steering Diffusion Models

Chi Hong[1(✉)], Jiyue Huang[1], Lydia Chen[1], and Robert Birke[2]

[1] Delft University of Technology, Delft, The Netherlands
{c.hong,j.huang-4,Y.Chen-10}@tudelft.nl
[2] University of Torino, Turin, Italy
robert.birke@unito.it

**Abstract.** Model stealing allows to extract the knowledge of a deployed target machine learning model by sending query images and training a substitute model via the inference results. However, accessing the same data used to train the target model is often impractical. Recent data-free model stealing methods overcome this limit using specifically crafted noise as queries. However, two major flaws limit their effectiveness. First, data-free methods suffer from low query efficiency, requiring high query budgets, such as over 20 million queries to train a substitute for CIFAR-10. Second, crafted queries lack any perceptual semantics. Hence, they can easily be filtered out from legitimate requests. To address these issues, we propose AEDM, a framework for **A**dversarial knowledge **E**xtraction via steering **D**iffusion **M**odels. AEDM leverages publicly available pre-trained diffusion models to craft adversarial query images, controlled through the latent variable fed into the diffusion models, to maximize the knowledge transferred to the substitute model. These queries not only resemble in all aspects legitimate requests of images with a semantic meaning, but also significantly lower the number of required queries. Results on three datasets demonstrate that, given a budget as limited as 1/200 queries of the baselines, the accuracy of our trained substitute model can outperform that of state-of-the-art data-free stealing methods.

**Keywords:** Knowledge extraction · adversarial learning · diffusion models

## 1 Introduction

A rich variety of online services such as Google OCR [7] and GPTZero [22] rely on deep learning models trained at high costs to provide their functionalities. Users interact with these services through APIs, which allow them to send queries to the model and receive the inference results. Most APIs provide a per-class probability distribution as inference results. While this open accessibility greatly enhances user experience, it also creates opportunities for malicious parties to steal knowledge from the deployed *target* models [16,26,27]. Adversaries extract knowledge by training a *substitute* model, which aims to replicate the mapping of the target model. Such substitute models can then be exploited for profit by


M. Mahmud et al. (Eds.): ICONIP 2024, CCIS 2293, pp. 336–350, 2025.
https://doi.org/10.1007/978-981-96-7005-5_23

**Table 1.** Number of queries to steal a CIFAR-10 classifier

| Method | DFME [23] | MAZE [16] | TandemGAN [12] | DS [1] | AEDM (ours) |
|---|---|---|---|---|---|
| Queries (x10K) | 800 | 900 | 2000 | 600 | **10** |

hosting inference services [17] or to launch further attacks, such as membership inference attacks [25] and adversarial attacks [3,8].

Ideally, adversaries use queries semantically similar to those used to train the target model. However, in many cases, adversaries face challenges in obtaining such real data, particularly from privacy- and business-sensitive domains like banks and medical institutions. Taking into account the limited accessibility to real training data, recent studies [1,12,16,23] propose data-free model stealing methods. Such methods use a generator designed to synthesise query examples out of random noise as inputs to query the target model, where real data is no longer necessary. Impressively, the performance of substitute models trained with such data-free methods has achieved levels similar to query by real data. However, data-free stealing has two significant drawbacks that limit its practical efficacy. First, it requires massive numbers of queries. For example, stealing a benchmark CIFAR-10 image ($32 \times 32$ pixels) classifier based on ResNet-34 [9] requires between 6 and 20 million synthetic queries (see Table 1). Querying the online service multiple times makes it easy for the target to detect the adversarial behaviour, and for the attacker to exceed the cost budget. Second, the key difference among diverse data-free model stealing baselines is how they generate query examples, but overall, all craft specifically designed noise with no semantic meaning. We report some example queries generated by data-free stealing baselines in Fig. 2. The lack of semantic meaning allows the target to easily single out adversarial queries from legitimate queries. For example, a simple defence that compares the entropy of the training data against the data from adversarial queries is already able to drop the defence pass rate below 21% for all baselines (see Table 5 for details).

To address both challenges we propose to enhance the query quality. Our observation study highlights that semantic information has a significant impact on both knowledge extraction and hiding adversarial queries from the target model. Leveraging the capabilities of emerging diffusion probabilistic models [4,6,11,13], which generate high-quality realistic images from latent noise, we introduce a novel framework called **A**dversarial knowledge **E**xtraction via steering **D**iffusion **M**odels (AEDM). This framework relies on a diffusion model pre-trained on public data to infuse the adversarial queries with semantic meaning so that they seem realistic and legitimate while being optimised to maximise knowledge extraction. The diffusion model serves as a semantic knowledge prior. During query generation, AEDM tunes the latent noise of the diffusion model so that the synthesised images maximise the entropy of the output of the substitute model. Then the tuned query is sent to the target model. To efficiently and effectively steal the target model, the substitute model minimizes the distance

**Fig. 1.** Example of query samples generated by our AEDM

between its predictions and those returned by the target model. Figure 1 shows examples of optimised queries when stealing a classifier trained on CIFAR10. One can note how the queries perceptually resemble images from the target model training dataset, though the diffusion model is pre-trained on a different dataset, i.e. ImageNet. This demonstrates the effectiveness of capturing knowledge from the target model.

This process is repeated until the performance of the substitute model resembles that of the target model. This approach enables the sampling of diverse and high-quality queries which maximise the learning capability of the substitute model and drastically reduces the number of required queries. Evaluation on three datasets has shown that using semantic-infused queries, AEDM is able to successfully steal the target model with 1/200–1/100 of the number of queries needed by state-of-the-art data-free baselines, achieving up to 65.88% higher stealing accuracy than baselines with more queries. At the same time, the stealthiness of the knowledge extraction is increased. Due to the closing of the semantic gap between adversarial and legitimate queries, AEDM overcomes simple defences effective against the other baselines.

In summary, the contributions of this paper are as follows. In Sect. 3, we propose a framework for knowledge extraction, AEDM that significantly reduces the number of queries required for successful stealing. The generation of high-quality queries is facilitated by a novel diffusion model prior. In Sect. 4, we conduct extensive comparisons between AEDM and state-of-the-art data-free model stealing approaches. Our empirical results demonstrate that AEDM can efficiently and effectively steal models using very few queries while achieving higher stealing accuracy. Additionally, AEDM is effective under federated extraction settings to further reduce the number of queries per attacker with only a slight accuracy loss. The existing defence against data-free attacks also fail to detect most of AEDM real-like queries with semantic meaning, arising the need for more advanced defences for such adversarial knowledge extraction methods.

## 2 Related Work

Model stealing aims at extracting the knowledge from a target model to a closely resembling substitute model, which is usually smaller in neural network size [2, 14, 18, 27]. Compared with white-box knowledge distillation [10, 26], where the layer-wise parameters of the target model are used to train the substitute model, adversarial model stealing can only leverage the inference information obtained from the target model through numerous queries [15, 21, 23]. Specifically, given

the same input, the substitute model aims to approximate the output of the target model by updating its network parameters, thus, expecting to copy the implicit mapping function of the target model. Early model stealing studies assume (partial) access to the training data used to query the target model [10]. Their application is limited when such training data is not available.

Data-free stealing is proven effective without the necessity of using real training data [12,16,18,27]. To accomplish this, the adversary continuously optimizes a generator to synthesize high-quality noise to query the target model. For output approximation, DFME [23] relies on the L1 norm loss when quantifying the output difference between the target and substitute model, whereas DaST [27] uses the cross entropy loss, and MAZE [16] applies the Kullback–Leibler divergence to measure the distance. Stealing by synthetic data benefits from diverse query examples [1,12]. To thoroughly investigate the input space, TandemGAN [12] brings a dual generator which explores the synthetic data space and then exploits high-quality examples to maximize the knowledge transfer. Differently, DS [1] explores a more diverse input space via two symmetrically trained substitute models. However, data-free model stealing typically requires a significant number of queries due to the substantial difference between the training inputs of the target and substitute models. In contrast AEDM leverages a public pre-trained distillation model to close this gap utterly reducing the number of queries to achieve successful stealing.

## 3 Methodology

In this section, we first present the fundamental concept and problem statement of classic data-free model stealing attacks. Then, we motivate the benefits of using adversarial queries which seem realistic to the target but are optimized to maximize knowledge extraction. Thus, we propose our AEDM framework, where we leverage a pre-trained diffusion model to generate semantic meaningful queries. Finally, the local training objective of the substitute model is further elaborated to closely mimic the target model.

### 3.1 Problem Statement

**Data-free Model Stealing.** Given a trained target $N$-class classification neural network $\mathcal{T}$, where its network architecture, training data and model parameters are unknown to adversaries. The attacker attempts to copy the knowledge of $\mathcal{T}$ by training a substitute model $\mathcal{S}$. It should be noted that under our setting with no access to the training data, the query sample is synthetic image data, denoted as $x \in \mathbb{R}^{H \times W \times C}$, where $H$, $W$, and $C$ are image height, width and the number of channels, respectively. Given the same input $x$, the substitute output $\mathcal{S}(x)$ is expected to be similar with the target output $\mathcal{T}(x)$. To achieve this, the target model $\mathcal{T}$ allows restricted access for querying, which means the attackers are able to obtain $\mathcal{T}(x)$ for a limited number of queries. Here the inference output $\mathcal{T}(x)$ is defined by probability vector $(p_1, p_2, ..., p_N)$ among different class labels

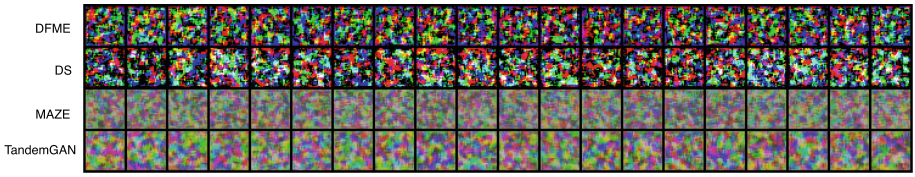


**Fig. 2.** Visualize baseline queries.

where $\mathcal{T}_j(x)$ denotes the $j$-th ($j = 1, ..., N$) element of the output vector $\mathcal{T}(x)$. Hence, data-free model stealing leverage the input $x$ and the inference $\mathcal{T}(x)$ pairs to train the substitute model $\mathcal{S}$. Then, the important problem is designing methods to craft synthetic query samples $x$.

### 3.2 Motivation Study

Current state-of-the-art data-free model stealing baselines seek to produce high-quality queries $x$ by optimizing a untrained generators, which generate queries from random latent variables. The generator parameters are tuned during the stealing process so as to effectively search the data space of $x$. Figure 2 illustrates the synthetic queries generated by baseline data-free methods of DFME, DS, MAZE, and TandemGAN. It is evident that they are all noisy images. According to the visualization, none of them exhibit any perceptual meaning.

We compare the model stealing performance of $\mathcal{S}$ using different types of query data: *i)* the same data used for training the target model, *ii)* Gaussian noise data sampled from $N(0, I)$ and *iii)* realistic images sourced from different public dataset, as shown in Table 2. Here, we employ the *CIFAR-10* dataset as an example and report the model classification accuracy of the substitute model over 10K queries. The accuracy of the target model $\mathcal{T}$ is 95.59%. The comparison in Table 2 reveals that the quality of queries significantly influences the training of substitute models. Random noise without any generator design targeting specific tasks fails in extracting the knowledge. Moreover, the substitute model's performance benefits from leveraging query data infused with semantic information. For instance, when querying with images from a different public datasets even with no semantic overlap, the accuracy is notably higher compared to using noise.

**Table 2.** Comparison of the (Acc)uracy (%) of the substitute models over 10000 queries of different type of query data.

| Dataset | Target | Substitute Acc. | | |
|---|---|---|---|---|
| | Acc. | CIFAR-10 | Noise | ImageNet |
| *CIFAR-10* | 96.59 | 59.56 | 10.00 | 18.81 |

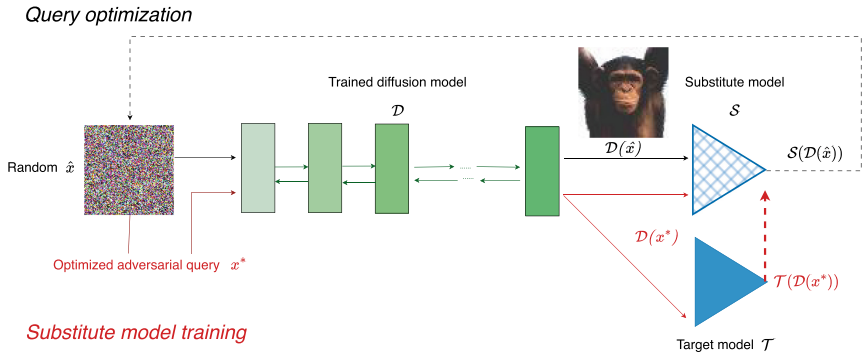


**Fig. 3.** AEDM framework on a single client: data-free model stealing process.

The above results motivated us to reconsider the power semantic information can have in knowledge extraction. Inspired by active learning, which proactively selects the subset of high quality examples to be labeled next by an expert from the pool of unlabeled data, we propose to train $\mathcal{S}$ by "selected" semantic-infused queries $x^*$ as follows.

### 3.3 AEDM Framework

We propose our model stealing algorithm based on Adversarial Knowledge Extraction via Steering Diffusion Models (AEDM), which can steal the knowledge from $N$-class image classifier $\mathcal{T}$ and train an accurate substitute model $\mathcal{S}$ with low query budget. The model parameters for $\mathcal{S}$ are represented as $\theta_s$. To optimize $\theta_s$ so that $\mathcal{S} \sim \mathcal{T}$, we first generate realistic high-quality queries infused with semantic information, which leverages a pre-trained diffusion models as prior. Then, such diverse queries are used to minimize the inference difference between the target model and the substitute model. Algorithm 1 summarizes the attack steps.

**Architecture.** The architecture of AEDM, is shown in Fig. 3. We first generate random Gaussian noise $\hat{x}$ that is of the same dimension of the realistic synthetic images $\mathcal{D}(\hat{x})$. The Gaussian noise will be used to sample a realistic image containing semantic information form a pre-trained diffusion model to query the current (or initialized, for the first iteration) $\mathcal{S}$ and get a high-quality query $x^*$ which is the final fine-tuned $\hat{x}$. Given $x^*$, AEDM feds $x^*$ into $\mathcal{T}$ and $\mathcal{S}$ at the same time to get outputs for updating the neural network parameters of $\mathcal{S}$. Note that this is an iterative process. Each query will use the substitute model updated based on to the previous query. The algorithm procedure alternates query generation and substitute model optimization.

**Query Generation.** Our objective is to generate queries that exhibit diversity, exploring a wide data space, while also containing semantic information. To achieve this, we apply a pre-trained diffusion model to provide semantic information in the produced queries while design a query optimization method to search the input data space of the diffusion model.

**Algorithm 1** AEDM

**Require:** A trained target model $\mathcal{T}$, a pre-trained diffusion model $\mathcal{D}$
**Input:** $n_r$,$n_d$, $\mu_d$ and $\mu_s$
Initializing the substitute parameters $\theta_s$
**Function** `AdversarialExtraction`($\theta_s$, $\mathcal{D}$)**:**
  Initialize $\mathcal{S}$ by $\theta_s$
  **for** $t = 1, ..., n_r$ **do**
1   $\hat{x} \sim \mathcal{N}(0, I)$
    **for** $i = 1, ..., n_d$ **do** `// Query generation`
      Get the synthetic output $\mathcal{D}(\hat{x})$
      Get the substitute output $\mathcal{S}(\mathcal{D}(\hat{x}))$
      $\mathcal{L}_{\mathcal{D}} = \sum_{j=1}^{N} p(\mathcal{S}_j(\mathcal{D}(\hat{x}))) \log p(\mathcal{S}_j(\mathcal{D}(\hat{x})))$
      $\hat{x} \leftarrow \hat{x} - \mu_d \nabla_{\hat{x}} \mathcal{L}_{\mathcal{D}}$
    Let $x^*$ be the optimized $\hat{x}$
    Get $\mathcal{T}(\mathcal{D}(x^*))$ and $\mathcal{S}(\mathcal{D}(x^*))$
    $\mathcal{L}_{\mathcal{S}}(\mathcal{D}(x^*)) = Distance(\mathcal{T}(\mathcal{D}(x^*)), \mathcal{S}(\mathcal{D}(x^*)))$, compute $\nabla_{\theta_s} \mathcal{L}_{\mathcal{S}}(\mathcal{D}(x^*))$
    $\theta_s \leftarrow \theta_s - \mu_s \nabla_{\theta_s} \mathcal{L}_{\mathcal{S}}(\mathcal{D}(x^*))$ `// Training the substitute model` $\mathcal{S}$
Return $\theta_s$
**Result:** The trained $\theta_s$

Diffusion models are a class of latent variable generative models designed to synthesize high-quality data. A well-trained diffusion model possesses the capability to generate realistic and diverse data following a training procedure that includes both a forward and reverse process. During the forward process, Gaussian noise is progressively added to the original training data across multiple steps. Each step introduces a controlled level of structured complexity. Conversely, the reverse denoising process is responsible of learning to identify and remove the specific noise patterns introduced at each step, thus restoring the original data. Consequently, the input and output dimensions remain the same across all steps, and a well-trained diffusion model can effectively sample data resembling real-world data with semantic information from Gaussian noise.

Given a pre-trained diffusion model $\mathcal{D}(\cdot)$ learned from a public dataset and random Gaussian noise input $\hat{x} \in \mathbb{R}^{H \times W \times C}$, an intuitive method to obtain semantic-enhance queries with diversity is:

$$x^* = \arg\max_{\hat{x}}(-\sum_{j=1}^{N} p(\mathcal{S}_j(\mathcal{D}(\hat{x}))) \log p(\mathcal{S}_j(\mathcal{D}(\hat{x})))), \tag{1}$$

where $x^* \in \mathbb{R}^{H \times W \times C}$ is the optimized input for producing semantic-enhanced queries $\mathcal{D}(x^*)$, obtained by maximizing the entropy of the output from the substitute model. In Eq. 1, we optimize $\hat{x}$ while $\mathcal{D}$ is a static pre-trained diffusion model on public datasets. Remarkably, it is not necessary that $\mathcal{D}(\cdot)$ generates very similar images as the training data of $\mathcal{T}$. This makes our method fit for general applications. In the evaluation, we will present the results of different datasets for the target model and diffusion model to demonstrate the versatility.

A query sample $\mathcal{D}(x^*)$ is crafted to maximize the entropy of the substitute output. The high entropy on the produced query sample means that it is a

**Algorithm 2** Federated adversarial extraction

**Input:** $n$, $K$
Initializing the substitute parameters $\theta_{s,0}$
**for** $t = 1, ..., n$ **do** `// n rounds of Federated training`
  **for** $k = 1, ..., K$ **do**
    $\theta^k_{s,t} =$ `AdversarialExtraction` $(\theta_{s,t-1}, \mathcal{D})$
  $\theta_{s,t} \leftarrow \frac{1}{K} \sum_{k=1}^{K} \theta^k_{s,t}$
**Result:** The global $\theta_{s,n}$

informative sample for $\mathcal{S}$. The substitute $\mathcal{S}$ rarely encountered similar samples in previous training. Thus, it is reasonable that we use such a sample, $\mathcal{D}(x^*)$, to query the target model $\mathcal{T}$ and get the inference result for training the substitute model.

**Substitute Model Optimization.** Since $\mathcal{S}$ is a substitute of $\mathcal{T}$, their outputs on the same given input are expected to be as similar as possible. In our method, $\mathcal{S}$ imitates the outputs of $\mathcal{T}$ through minimizing their output distance on the same input $\mathcal{D}(x^*)$ obtained from the previous step. The loss function of $\mathcal{S}$ over $\mathcal{D}(x^*)$ is:

$$\mathcal{L}_{\mathcal{S}}(\mathcal{D}(x^*)) = Distance(\mathcal{T}(\mathcal{D}(x^*)), \mathcal{S}(\mathcal{D}(x^*)))$$

In knowledge extraction, the commonly used distance metrics are cross entropy and $l_1$ norm. Experimentally, we find that using $l_1$ norm can make $\mathcal{S}$ converge with less iteration. Thus, we choose $l_1$ norm in our implementation. Then, we have the following loss

$$\mathcal{L}_{\mathcal{S}}(\mathcal{D}(x^*)) = \|\mathcal{T}(\mathcal{D}(x^*)) - \mathcal{S}(\mathcal{D}(x^*))\|_1. \tag{2}$$

Accordingly, the optimization objective of training the substitute model is:

$$\arg\min_{\theta_s} \|\mathcal{T}(\mathcal{D}(x^*)) - \mathcal{S}(\mathcal{D}(x^*))\|_1 \tag{3}$$

After training $\mathcal{S}$ by many rounds of queries, the mapping of $\mathcal{T}$ from a given image to the classification output can be captured by $\mathcal{S}$. Thus, the knowledge of the target model $\mathcal{T}$ is extracted by the algorithm via building the substitute $\mathcal{S}$.

### 3.4 Federated Extraction

To further reduce the number of queries from a single adversarial party and lower the possibility of detection by the target model, AEDM can also be applied in a federated context for model stealing. In our federated extraction approach, multiple adversarial parties, i.e., clients or attackers, collude to collaboratively train over $n$ global rounds the substitute model without directly sharing queries. Consequently, the number of queries for each adversary decreases to $1/K$ of the original, where $K$ represents the number of clients involved.

In particular, a single adversarial server coordinates multiple client entities engaged in stealing. In each global training round $t$, every client employs the

stealing algorithm shown in Algorithm 1 to train their local substitute model $\theta_s^k$, based on the publicly available diffusion model $\mathcal{D}$. The parameters of these local substitute models at the end of each round $t$ are then aggregated through averaging:

$$\theta_{s,t} = \frac{1}{K} \sum_{k=1}^{K} \theta_{s,t}^k. \tag{4}$$

The aggregated substitute model parameters will be distributed to every clients for next round of training. The federated adversarial extraction process by multiple attackers is shown in Algorithm 2.

## 4 Evaluation

In this section, we conduct a thorough evaluation of the performance of our adversarial knowledge extraction via steering diffusion models, measured by the accuracy of the substitute model and the number of queries. We start with introducing our experimental setups and then provide a comprehensive comparison between our AEDM and state-of-the-art baseline stealing methods. Furthermore, we expand the scope of our evaluation by demonstrating its performance under Federated knowledge extraction settings, where multiple adversarial parties collaboratively launch the stealing attack. We also apply a commonly-adopted defence mechanism against knowledge extraction to validate the effectiveness of our attack. In addition, we investigate the sensitivity of our framework across different neural network architectures of substitute models, providing insights into real-world applications of stealing.

### 4.1 Experimental Setup

**Dataset and Model Structure:** Our proposed method is evaluated on three 3-channel image classification datasets: CIFAR-10 [19], CIFAR-100 [19], and SVHN [20]. It is important to note that different network structures are used for the target and substitute models since attackers lack prior knowledge of the target structure. Specifically, for the three data sets, we use the commonly adopted ResNet-34 architecture for the target model $\mathcal{T}$, while we use VGG-16 for the substitute model $\mathcal{S}$ in the experiments.

**Pre-trained Diffusion Models:** In the context of AEDM, it is not required for the adversary to have in-distribution similar data as the training set used by the target model. Instead, access to publicly available pre-trained diffusion models is sufficient and practical. Therefore, for all three data sets, we use diffusion models trained on ImageNet [5].

**Federated Adversarial Setting:** We evaluate Federated collaborative stealing with 5, 10 and 20 collusive attackers. The total number of queries remains the same as for a single attacker, with each attacker using an equal split, i.e. $1/K$, of the total number of queries. The pre-trained diffusion models held by each

**Table 3.** Comparison of the $\mathcal{S}$ubstitute model Acc(uracy) under MAZE, DFME, DS, TandemGAN and AEDM on *SVHN*, *CIFAR-10* and *CIFAR-100*. The network architectures is ResNet-34 for all target models and VGG-16 for all substitute models.

| Dataset | $\mathcal{T}$arget | Baselines | | | | | Ours | |
|---|---|---|---|---|---|---|---|---|
| | Acc. | Querie | DFME | MAZE | DS | TandemGAN | Querie | AEDM |
| *SVHN* | 95.18 | 2M | 91.50 | 89.84 | 89.68 | 93.45 | **20K** | **93.47** |
| *CIFAR-10* | 96.59 | 20M | 60.18 | 55.33 | 58.84 | 77.55 | **0.1M** | **88.45** |
| *CIFAR-100* | 78.71 | 20M | 25.91 | 14.12 | 22.00 | 37.17 | **0.2M** | **61.66** |

client are trained by ImageNet, which is different from our evaluation datasets to validate the versatility of the approach.

**Evaluation Metrics:** Our aim in adversarial knowledge extraction is to attain high classification **accuracy** on the substitute model with a minimal **number of queries**. When encountering defences, we also evaluate the **defence pass rate** (DPR $\in [0, 1]$) of adversarial queries, which is defined by the rate of undetected attacker queries by the target model.

**Baseline Methods:** We compare our adversarial knowledge extraction with four state-of-the-art baseline methods under the same settings: DFME [23], MAZE [16], DS [1] and TandemGAN [12].

### 4.2 Adversarial Knowledge Extraction Performance

We evaluate AEDM against the state-of-the-art baselines based on the accuracy of substitute models and the number of queries required to achieve such accuracy. The results are detailed in Table 3. The accuracy of the target ResNet-34 models trained on three different datasets is provided as reference for comparison. For each baseline method, the substitute model is trained by VGG-16 due to limited knowledge of the target network by attackers. Since AEDM is designed to notably minimize the required query number, we also present the number of queries for each dataset accordingly.

As shown in Table 3, baseline data-free extraction methods, which employ tuned synthetic noisy data to query the target model, typically need a large number of queries ranging from 2 million (notated as 2M in the table) to 20M. This aligns with our expectations, as illustrated by the motivation results in Table 2, and indicates the inefficiency of using synthetic noisy data without semantic meanings. Although such noise is well-designed and generated to explore diverse random data spaces, its resemblance to the actual training data of the target model is limited. Consequently, certain important features containing semantic information (such as smooth pixel-wise edges and typical luminance contrast) from the target model's training images are difficult to capture with noisy data. In contrast, leveraging the semantic information provided by the diffusion model, alongside our optimized latent variables, our AEDM requires merely 20K to

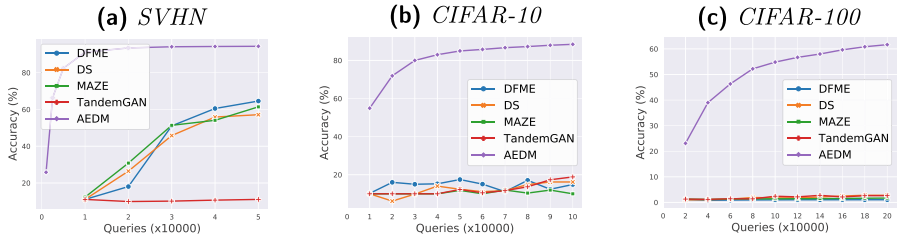


**Fig. 4.** Substitute model accuracy of AEDM over the increasing number of queries used on *SVHN*, *CIFAR-10* and *CIFAR-100* datasets.

0.2M queries on the same target model to achieve even higher accuracy than the baselines. We visualized queries generated by our AEDM in Fig. 1. Note that the diffusion model is pre-trained by public data, i.e. ImageNet [5], and we optimize the latent variable to generate query data which maximises the knowledge extraction from the target model.

Zooming into the impact across different datasets, we can clearly observe that baseline methods employing synthetic noisy data perform effectively on the *SVHN* dataset, which is the relatively simple task of classifying street house numbers. The target model for this dataset is originally trained on monotonous data with shallow features, that makes it easily extractable by substitute models. However, when confronted with a more complex task such as the 100-class classification of *CIAFR-100*, synthetic noisy queries struggle to train high accuracy substitute models, even with query numbers reaching tens of millions. However, our AEDM maintains good performance with only 1/100 number of the queries, underlining the essential need of using semantically informative queries.

To compare by the number of queries explicitly, we maintain a consistent number of queries and track the accuracy of the substitute models as the queries increase in Fig 4. Specifically, we set track the accuracy up to 50K, 0.1M, and 0.2M queries for *SVHN*, *CIFAR-10*, and *CIFAR-100*, respectively. Figure 4 illustrates the superiority of AEDM in knowledge extraction through adversarial queries, demonstrating both effectiveness and efficiency. For simple tasks like *SVHN*, AEDM converges with approximately 30K queries, whereas for *CIFAR-10* and *CIFAR-100*, AEDM continues to improve, though it already significantly outperforms the baselines. This observation is in line with our findings in Tab 3 that more complex tasks require a greater number of queries to thoroughly explore and extract knowledge from the target model.

It is also interesting to notice that the performance of the baseline methods varies depending on the task at hand. The primary distinction among these baseline methods lies in how they generate the noisy query data to mimic the training data space. For instance, DFME outperforms other baselines on *SVHN* but exhibits significant fluctuations on *CIFAR-10* and falls behind TandemGAN on *CIFAR-100*. This can be explained by the fact that single-generator data-free frameworks lack diversity in the query data space, which aligns with the findings

**Table 4.** Comparison of the (Acc)uracy of the substitute models over different number of clients (attackers) in Federated learning.

| Dataset | Centralized | Federated Model Stealing | | |
|---|---|---|---|---|
| | Acc. | 5 Clients | 10 Clients | 20 Clients |
| *SVHN* (20K Queries) | 93.47 | 91.11 | 90.76 | 91.49 |
| *CIFAR-10* (0.1M Queries) | 88.45 | 83.35 | 85.56 | 84.43 |

of TandemGAM [12]. Thanks to the exploration and exploitation process facilitated by dual generators, TandemGAN performs well and consistently, especially on more complex classification tasks, while suffering from inefficiencies, particularly in simple tasks.

### 4.3 Extraction Under Federated Setting

Given the fact that extremely large number of queries by a single adversary will cause easy detection by the target model, her we evaluate our AEDM stealing method under a federated settings where multiple federated attackers collusively launch the attack. The results on *SVHN* with 20K queries in total and *CIFAR-10* with 0.1M queries in total are summarized in Table 4. Note that the number of queries is equally split over all participating attackers (clients) and every attacker holds a diffusion model pre-trained by ImageNet. From Table 4, we see that on both datasets stealing under federated setting reaches similar but slightly lower substitute model accuracy than stealing by a single adversary. This result validates that distributing the stealing process to multiple clients can significantly reduce the number of per-client queries while maintaining similar knowledge extraction effectiveness. One interesting note from Table 4 is the substitute model accuracy over different number clients. The accuracy does not monotonically increasing or decreasing with more clients but the same total number of queries. Yet, among 5 to 20, 20-clients achieves the highest substitute model accuracy on *SVHN* while 10-clients works better on *CIFAR-10*. This finding indicates that more query partitions for stealing does not necessarily decrease the accuracy and the best number of clients for collaborative stealing can be specified on different task and data.

**Table 5.** Comparison of the defence pass rate (DPR) with baseline extraction methods on *Cifar-10* with the presence of the entropy-based defence.

| Methods | DFME | MAZE | DS | TandemGAN | AEDM |
|---|---|---|---|---|---|
| DPR | 0.060 | 0.021 | 0.210 | 0.164 | **0.853** |

### 4.4 Effectiveness Against Defense

To validate the robustness of AEDM, we conduct the algorithm on a setting with defense. The target model may employ defense mechanisms to detect abnormal queries. In this part we analyse our extraction effectiveness in the presence of stealing defenses. One lightweight but effective approach is to detect the quality of received queries, in which low-quality queries are determined as outliers and possible attackers [24]. The context of image quality in this paper is the entropy of the query. We present the defense pass rate of each knowledge extraction method in Table 4. From the results, it is obvious that the baseline data-free stealing methods which generate noise query images, are easily identified as outliers and stopped, e.g., the highest DPR across all baselines is 21% by DS followed by TandemGAN 16%. Thanks to the diffusion model infusing semantics into the queries, more than 85% of the queries of AEDM bypass the defense.

As our AEDM generates adversarial knowledge extraction queries similar to real legitimate queries, current simple defense mechanisms are no longer effective against our attack. To enhance the defence one possible way left for future work is to leverage the image similarity between the query data and the training data, i.e., only allow in-distribution data for queries. However, this comes at a trade-off with additional computational overhead at the target model.

**Table 6.** Comparison of the (Acc)uracy of the substitute models over different network (Arc)hitecture for AEDM.

| Dataset | Target | Substitute Arc. | |
|---|---|---|---|
| | Acc. | VGG-16 | ResNet-18 |
| *CIFAR-10* | 96.59 | 88.45 | **90.77** |
| *CIFAR-100* | 78.71 | 61.66 | **65.19** |

### 4.5 Sensitivity over Different Substitute Architecture

Given that the adversary is assumed to be unaware of the target model architecture, the substitute model can be selected from various neural network architectures. Therefore, in this section, we investigate the impact of choosing different neural network architectures on *CIFAR-10* and *CIFAR-100*. The findings are presented in Table 6. Here, the target model is ResNet-34, and we compare the knowledge extraction effectiveness of substitute models using VGG-16 and ResNet-18, with the same number of queries, i.e., 0.1M and 0.2M for two datasets respectively. According to the results, ResNet-18 demonstrates superior extraction accuracy, likely due to the fact that it belongs to the same neural network family as the target model. This outcome aligns with our expectations that similar neural network architectures exhibit better ability in knowledge transfer/extraction, and is in line with the findings from [12]. However, in our paper, we employ VGG-16 for other experiments to show better versatility without prior knowledge of the target model, where VGG-16 achieves only slightly lower accuracy.

## 5 Conclusion

Existing data-free model stealing attacks benefit from broader applications due to not requiring real data, but suffer from extremely large number of queries to achieve high substitute model accuracy. In order to reduce the number of queries, this paper proposes a novel adversarial knowledge extraction via steering diffusion models AEDM. Different from generator-based data-free approaches, AEDM generates synthetic data by diffusion model pretrained on publicly available data to generate semantically meaningful queries for the target model. To ensure diverse and realistic data queries, AEDM iteratively optimizes the diffusion model input latent variable to explore the large image space. We empirically demonstrated the effectiveness and efficiency of stealing by AEDM, in terms of higher substitute model accuracy with as little as 1/100 to 1/200 of the number of queries required by state-of-the-art baseline methods. Our extensive experiments show that AEDM also works against data-free query defence, as well as under federated setting, indicating the possibility of further reducing the per-client query number by collaborative stealing.

## References

1. Beetham, J., Kardan, N., Mian, A.S., Shah, M.: Dual student networks for data-free model stealing. In: ICLR 2023. OpenReview.net (2023)
2. Chandrasekaran, V., Chaudhuri, K., Giacomelli, I., Jha, S., Yan, S.: Exploring connections between active learning and model extraction. In: USENIX Security 2020, pp. 1309–1326. USENIX Association (2020)
3. Chen, F., Shen, Z.: A data-free substitute model training method for textual adversarial attacks. In: International Conference on Neural Information Processing. pp. 295–307. Springer (2023). https://doi.org/10.1007/978-981-99-8181-6_23
4. Deja, K., Trzcinski, T., Tomczak, J.M.: Learning data representations with joint diffusion models. In: Koutra, D., Plant, C., Rodriguez, M.G., Baralis, E., Bonchi, F. (eds.) ECML PKDD 2023, Turin, Italy, vol. 14170, pp. 543–559. Springer (2023). https://doi.org/10.1007/978-3-031-43415-0_32
5. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: (CVPR 2009), 20-25 June 2009, Miami, Florida, USA, pp. 248–255. IEEE Computer Society (2009)
6. Gilhuber, S., Busch, J., Rotthues, D., Frey, C.M.M., Seidl, T.: Diffusal: coupling active learning with graph diffusion for label-efficient node classification. In: ECML PKDD 2023 (2023)
7. Google, C.T.: Google cloud vision api for ocr (2016). https://cloud.google.com/vision/docs/ocr
8. Guo, Z., Han, K., Ge, Y., Li, Y., Ji, W.: Attribution of adversarial attacks via multi-task learning. In: International Conference on Neural Information Processing, pp. 81–94. Springer (2023). https://doi.org/10.1007/978-981-99-8082-6_7
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR 2016
10. Hinton, G.E., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. CoRR abs/ arXiv: 1503.02531 (2015)

11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, 6-12 December 2020, virtual (2020)
12. Hong, C., Huang, J., Birke, R., Chen, L.Y.: Exploring and exploiting data-free model stealing. In: ECML PKDD 2023, Turin, Italy. Springer (2023). https://doi.org/10.1007/978-3-031-43424-2_2
13. Huang, B., Yu, W., Xie, R., Xiao, J., Huang, J.: Two-stage denoising diffusion model for source localization in graph inverse problems. In: ECML PKDD 2023, Turin, Italy. Springer (2023). https://doi.org/10.1007/978-3-031-43418-1_20
14. Jagielski, M., Carlini, N., Berthelot, D., Kurakin, A., Papernot, N.: High accuracy and high fidelity extraction of neural networks. In: Capkun, S., Roesner, F. (eds.) 29th USENIX Security Symposium, USENIX Security 2020, pp. 1345–1362. USENIX Association (2020)
15. Juuti, M., Szyller, S., Marchal, S., Asokan, N.: PRADA: protecting against DNN model stealing attacks. In: IEEE European Symposium on Security and Privacy, EuroS&P 2019, pp. 512–527. IEEE (2019)
16. Kariyappa, S., Prakash, A., Qureshi, M.K.: MAZE: data-free model stealing attack using zeroth-order gradient estimation. In: CVPR 2021 (2021)
17. Krishna, K., Tomar, G.S., Parikh, A.P., Papernot, N., Iyyer, M.: Thieves on sesame street! model extraction of bert-based apis. In: 8th International Conference on Learning Representations, ICLR 2020
18. Krishna, K., Tomar, G.S., Parikh, A.P., Papernot, N., Iyyer, M.: Thieves on sesame street! model extraction of bert-based apis. In: 8th International Conference on Learning Representations, ICLR 2020. OpenReview.net (2020)
19. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
20. Netzer, Y., et al.: Reading digits in natural images with unsupervised feature learning. In: NIPS workshop on deep learning and unsupervised feature learning, Granada, Spain, vol. 2011, p. 7 (2011)
21. Orekondy, T., Schiele, B., Fritz, M.: Knockoff nets: stealing functionality of black-box models. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, pp. 4954–4963. Computer Vision Foundation / IEEE (2019)
22. Team, G.: Gptzero api (2023). https://gptzero.stoplight.io
23. Truong, J., Maini, P., Walls, R.J., Papernot, N.: Data-free model extraction. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021. pp. 4771–4780. Computer Vision Foundation / IEEE (2021)
24. Tsai, D., Lee, Y., Matsuyama, E.: Information entropy measure for evaluation of image quality. J. Digit. Imaging **21**(3), 338–347 (2008)
25. Yeom, S., Giacomelli, I., Fredrikson, M., Jha, S.: Privacy risk in machine learning: Analyzing the connection to overfitting. In: 2018 IEEE 31st Computer Security Foundations Symposium (CSF), pp. 268–282. IEEE (2018)
26. Yin, H., et al.: Dreaming to distill: data-free knowledge transfer via deepinversion. In: CVPR 2020, Seattle, WA, USA (2020)
27. Zhou, M., Wu, J., Liu, Y., Liu, S., Zhu, C.: Dast: data-free substitute training for adversarial attacks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, pp. 231–240. Computer Vision Foundation / IEEE (2020)