MODEL STEALING ATTACKS AGAINST VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models have flourished these years and are regarded as promising solutions to vision-language tasks. However, training vision-language models always requires enormous effort, making the models valuable intellectual properties (IPs). In this paper, we pioneer to propose the first model stealing attack against the vision-language models, the goal of which is to steal the functionality of the target models. Specifically, we target fine-tuned CLIP models with black-box access. We query the model to extract model information through either the text-to-image retrieval or the image-to-text retrieval API and then leverage the information to train a local copy of the target model. Experiments show the effectiveness of the model stealing attacks. We validate that our attacks are query-efficient, API-agnostic, data-agnostic, and architecture-agnostic, which broaden the attack scenarios. As a counterpart, we examine a defense based on the idea of out-of-distribution detection, which is impotent without strong assumptions. Our research pressures the unprotected release and prevalence of powerful vision-language models, and appeals to the community that their IP protections, if not the least, cannot be less.

1 INTRODUCTION

With the prospering growth of multimedia data from social networks, resolving vision-language tasks such as image-text retrieval (Yan & Mikolajczyk, 2015) and visual question answering (Antol et al., 2015) has attracted massive attention in recent years (Suhr et al., 2019; Nichol et al., 2021; Ramesh et al., 2022). To meet this rapidly growing demand, a considerable number of vision-language models have been proposed and achieved significant progress (Radford et al., 2021; Li et al., 2021; 2022). However, training a well-generalized model is time- and energy-consuming. The enormous amount of data, sophisticated model designs, and huge computational resource consumption make the vision-language models themselves valuable intellectual properties for the model owners.

Previous works uncover that remotely-deployed machine learning models are vulnerable to model stealing attacks via prediction APIs, where attackers with only black-box access can steal the functionality of target models (Tramèr et al., 2016; Chandrasekaran et al., 2020; Jagielski et al., 2020). Such attacks have been demonstrated to be a practical threat to the intellectual property of different types of models (e.g., discriminative models (Truong et al., 2021) and generative models (Hu & Pang, 2021)) and different types of data (e.g., images (Orekondy et al., 2019), texts (Krishna et al., 2020), and graphs (Shen et al., 2022)) in real-world scenarios. However, these attacks remain unexplored in the vision-language tasks.

In this paper, we pioneer to investigate the efficacy of the model stealing attacks against the vision-language models. Specifically, we aim to steal the functionality of the fine-tuned CLIP models (Radford et al., 2021), which can learn a visual-language embedding space and align the representations between given image-text pairs. In this case, attackers can access the target model via either image-to-text retrieval or text-to-image retrieval API. They can query with either images or texts, and the outputs are the other data modality.

Our experiments demonstrate the effectiveness of the model stealing attacks against the fine-tuned CLIP models. Moreover, such attacks have several advantageous properties, we summarize them in the following: 1) Our attacks are query-efficient. Specifically, our attacks only have about 3.15% text Recall@1 performance deterioration on the Flickr30K dataset and 2.69% on the MSCOCO dataset using queries with 10% of the original fine-tuning dataset size. 2) Our attacks are API-agnostic.

Attacking through either image-to-text retrieval or text-to-image retrieval API leads to similar efficacy, indicating that these two APIs leak about the same amount of information. We reason this is because both the images and texts are mapped from their domains to the same embedding space. 3) Our attacks are data-agnostic, as the attack performance only has a slight decrease when the attacker leverages an auxiliary dataset from a different distribution. 4) Our attacks are also architecture-agnostic. Concretely, experimental results show that the attack performance is closely related to the surrogate model's architecture, i.e., more powerful architecture leads to better performance, regardless of the architecture of the target model. We also perform a fine-grain analysis to conclude that high-agreement queries are favorable to enable the model stealing attacks.

We thoroughly study the model stealing attacks against vision-language models in this paper, and show that model stealing is a real-world threat to intellectual property. As a counterpart, we examine a defense mechanism (Hendrycks & Gimpel, 2017) by leveraging the idea of out-of-distribution detection; we find that the defense is impotent without strong assumptions. From a measurement view, our attack shows model stealing attacks can be enabled with several advantageous properties (e.g., query-efficient, API-agnostic), indicating the severity of the vulnerability. We hope our work can appeal to the community to pay necessary attention to protecting the model intellectual property, such as proposing stronger defense mechanisms.

2 Related Work

Vision-language representation learning. Vision-language tasks target to associating image-text pairs with the same semantic meanings, which essentially function in several real-world applications, including but not limited to visual question answering (Antol et al., 2015), natural language visual reasoning (Suhr et al., 2019), visual dialog (Das et al., 2019), and text-driven image generation (Ramesh et al., 2021; 2022) and editing (Nichol et al., 2021). Among the existing techniques, vision-language representation learning methods (Jia et al., 2021; Mu et al., 2021; Li et al., 2021; 2022) have shown their ascendancy and have been regarded as promising solutions to the tasks. To better understand their working mechanisms, we take contrastive language-image pre-training (CLIP) (Radford et al., 2021) as a representative. CLIP model takes as input an image-text pair $(\boldsymbol{i}_k, \boldsymbol{t}_k)$ where image $\boldsymbol{i}_k \in I$ comes from the image space and text $\boldsymbol{t}_k \in T$ is from the text space. CLIP projects $\boldsymbol{i}_k$ and $\boldsymbol{t}_k$ into a common latent space E through two learnable embedding functions $h : I \to E$ and $g : T \to E$, i.e., $\big(h(\boldsymbol{i}_k), g(\boldsymbol{t}_k)\big) \in E \times E$. To obtain better representation ability, CLIP optimizes both h and g such that the embeddings from the same pair have high cosine similarity while embeddings from different pairs have low similarity.

Model stealing attacks. Model stealing attacks aim to steal the functionality of the target model with black-box access. Attackers first collect samples from the same distribution as the training dataset and query the target model for their responses. These query-response pairs compose a surrogate dataset to train the surrogate model. Tramèr et al. (2016) take the first step towards attacking the deep neural classifiers; after that, researchers put effort into making the attack more practical (Orekondy et al., 2019; Juuti et al., 2019; Chandrasekaran et al., 2020; Jagielski et al., 2020; Truong et al., 2021), i.e., relaxing the assumptions on the surrogate dataset.

Although model stealing attacks have been widely studied in recent years, most works focus on images (Tramèr et al., 2016; Orekondy et al., 2019; Juuti et al., 2019; Chandrasekaran et al., 2020; Jagielski et al., 2020; Truong et al., 2021), texts (Krishna et al., 2020), or graphs (DeFazio & Ramesh, 2019; Wu et al., 2020; Shen et al., 2022). For these tasks, the roles of query-response pairs are predefined; for example, when stealing an image classifier, we have to use images to query the model and obtain labels as output, but the other direction does not hold, i.e., we are unable to query the model using labels. However, this is not the case when stealing vision-language models. In our attacks, the queries can be either images or texts, while the responses are the other data modality. All known attacks focus on discriminative models (Jagielski et al., 2020; Truong et al., 2021) or generative models (Hu & Pang, 2021; Szyller et al., 2021) as target models, which are different with our target models. Therefore, it is infeasible to apply existing attacks against vision-language models.

With the prosperity of vision-language representation learning popularizing applications in various domains, its vulnerability to security and privacy issues turns out increasingly critical yet remains largely unexplored. To our best knowledge, the only known attack against vision-language represen-

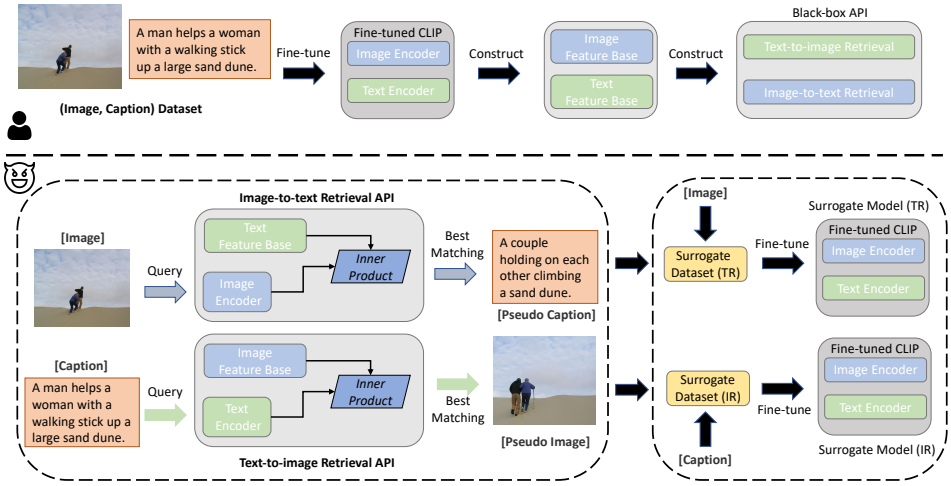


Figure 1: Overview of the model stealing attack via image-text retrieval tasks. In the upper part, the victim fine-tunes a well-generalized target model and generates embeddings for all image-text pairs he has to construct the image/text feature base and enable the image-to-text/text-to-image retrieval API. In the below part, the attacker first queries either the image-to-text retrieval or the text-to-image retrieval API and then leverages its retrieval outputs to fine-tune the surrogate CLIP models.

tation learning (Carlini & Terzis, 2021) leverages poisoning techniques to compromise the integrity of the model. In our paper, we instead focus on the model stealing attacks.

3 Vision-Language Model Stealing

3.1 Problem Statement

There are two parties involved in the model stealing attack, i.e., the *attacker* and *victim*. As illustrated in Figure 1, the victim is the identity who owns a well-generalized target model fine-tuned from a pretrained CLIP model (Radford et al., 2021), and wants to earn profits by providing machine learning services (MLaaS). They release the model and respond to users' queries in a black-box API manner; that is, the attacker has no information about the model structure and parameters; when the attackers submit queries, the only information they get from the server is the retrieval results. To construct the APIs, the victim generates embeddings for all image-text pairs he has to construct the image feature base and text feature base. The image-to-text retrieval API ($\mathcal{TR}$) receives an image as the query. Different from traditional settings, the CLIP-based API does not return a specific label. After generating the image embedding for the query, the $\mathcal{TR}$ performs an inner product with text feature base and outputs texts whose embeddings are most similar to the query image embedding. In the same manner, the text-to-image retrieval API ($\mathcal{TR}$) takes a text as input and outputs the corresponding images in the image feature base whose embeddings maximize the inner product with the query text embedding. In this paper, we consider the most challenging setting: given a text/image query, these two APIs only return the best matching image/text.

The attacker aims to reconstruct a surrogate model (the local copy of the target model) with black-box access to the target model. Besides the black-box access to the target model, we also assume the attacker has an auxiliary dataset $\mathcal{D}_{\mathcal{A}}$, which can be used to query the target model. This auxiliary dataset may come from the same distribution as the target dataset, or can be different. We discuss the influence of the auxiliary dataset in Section 4.3. Formally, we define the model stealing attack as follows:

$$\mathcal{A} : \mathcal{M}_{\mathcal{T}}, \mathcal{D}_{\mathcal{A}} \rightarrow \mathcal{M}_{\mathcal{S}}$$

where $\mathcal{M}_{\mathcal{T}}$ denotes the target model and $\mathcal{M}_{\mathcal{S}}$ denotes the surrogate model. Following the convention (Jagielski et al., 2020), we use accuracy and agreement to quantitatively define the attack performance, where accuracy measures the surrogate model's utility, and agreement reflects the fidelity toward the target model. We defer the concrete definition to Section 4.1. In summary, the goal of our attack is to construct a high-quality surrogate model with *high* accuracy on the target task and *high* agreement with the target model.

3.2 METHODOLOGY

The overview of our attack is also depicted in Figure 1. To initiate the attack, the attacker mainly leverages two types of APIs provided by the victim, i.e., the image-to-text retrieval API ($\mathcal{TR}$) and the text-to-image retrieval API ($\mathcal{IR}$). As we can either query $\mathcal{TR}$ to get corresponding captions for images or query $\mathcal{IR}$ to get corresponding images for text, it is feasible for the attacker to perform the attack from either the image or text side. Since the key steps are the same for both approaches, we only illustrate the workflow of stealing the model through $\mathcal{TR}$ for simplicity.

We assume that the attacker has access to an auxiliary dataset containing images without captions. The attacker queries the $\mathcal{TR}$ using images $\{\boldsymbol{i}_k\}_{k=1}^n$, and gets the best matching text ($\{\boldsymbol{t}'_k = \mathcal{TR}(\boldsymbol{i}_k)\}_{k=1}^n$) for each image. These ($\{\boldsymbol{t'}_k = \mathcal{TR}(\boldsymbol{i}_k)\}_{k=1}^n$) can be regarded as pseudo-captions, acting as the role of pseudo-labels in the traditional model stealing process. We compose a surrogate dataset $\mathcal{D}_{\mathcal{S}}^{\mathcal{TR}} = \{\boldsymbol{i_k}, \boldsymbol{t'_k}\}_{k=1}^n$ using these image-text pairs, and leverage this surrogate dataset to train the surrogate model, like training a normal CLIP model.

4 EXPERIMENTS

In this section, we first introduce the experimental setup and then demonstrate the effectiveness of our model stealing attack with advantageous properties, following with analysis and defense. Experiments are performed on an NVIDIA DGX-A100 server. We enclose the code in the supplemental material for reproduction.

4.1 SETUP

Datasets. We demonstrate the efficacy of the model stealing attacks on two benchmark datasets: MSCOCO (Chen et al. (2015)) and Flickr30K (Young et al. (2014)). These two datasets consist of images of everyday events in a natural context and are all harvested from the Flickr website. Each image is paired with five reference captions annotated by Amazon Mechanical Turk (AMT) workers. We split the datasets using the widely accepted Karpathy split Karpathy & Fei-Fei (2017). Specifically, for the Flickr30K dataset, we have a 29K training set and a 1K test set. For the MSCOCO dataset, we have a 113K training set and a 5K test set.

Implementation details. As training CLIP models require a large amount of data, following previous work (Krishna et al., 2020), we consider the fine-tuning scenario in this paper, i.e., the target model is fine-tuned on a base model, which is released by OpenAI (Radford et al., 2021). We randomly split the training set equally into two disjoint datasets, $\mathcal{D}_{\mathcal{A}}$ and $\mathcal{D}_{\mathcal{F}}$, and use $\mathcal{D}_{\mathcal{F}}$ to fine-tune the base model to obtain our target model. The other half, $\mathcal{D}_{\mathcal{A}}$, is used to launch the attacks. We consider two attack scenarios, i.e., the attacker has access to the $\mathcal{TR}$ or $\mathcal{IR}$ API. The attacker could leverage unlabeled images/text to query the API and get corresponding pseudo-labels based on different scenarios. Both the target models and surrogate models are evaluated on the same test sets D_T.

We use Vision Transformer vision model (ViT-B/32) (Dosovitskiy et al., 2021) and Transformer language model (Vaswani et al., 2017) as the backbone of the CLIP model. The target model with its backbone model is fine-tuned for 10 epochs using AdamW optimizer. The optimizer is initialized with a learning rate 1×10^{-7} with a cosine scheduler, and the weight decay is set to 0.02. The surrogate model follows the same training procedure.

Evaluation metrics. We mainly leverage two metrics for evaluation: accuracy and agreement, where accuracy reflects the utility of the surrogate model and agreement denotes the fidelity. Specifically, as CLIP models are always modified to perform image/text retrieval tasks, it is common to use Recall@K as the accuracy metric. Therefore, we adopt text-to-image Recall@K (abbreviated as R@K) and image-to-text Recall@K in this paper. These two recalls are calculated by counting

Table 1: Image-text retrieval results of the target models on the Flickr30K and MSCOCCO datasets (zero-shot (Radford et al., 2021) and fine-tuned settings).

Target Dataset	Settings	Image-to-text Retrieval R@1	R@5	R@10	Text-to-image Retrieval R@1	R@5	R@10
Flickr30K (1K test set)	Zero-shot	59.7	86.3	91.6	59.6	84.2	89.8
	Fine-tuned	71.2	91.5	95.1	71.9	90.5	95.2
MSCOCO (5K test set)	Zero-shot	34.1	59.6	70.1	30.4	55.5	66.4
	Fine-tuned	44.0	71.4	80.8	43.4	69.4	79.6

Table 2: A comparison of $\mathcal{M}_S^{\mathcal{IR}}$ and $\mathcal{M}_S^{\mathcal{TR}}$ in terms of the image-text retrieval Recall@K and agreement on the test set. Unless specified, all of the attacks use the same number of queries as the original fine-tuning dataset $\mathcal{D}_{\mathcal{F}}$.

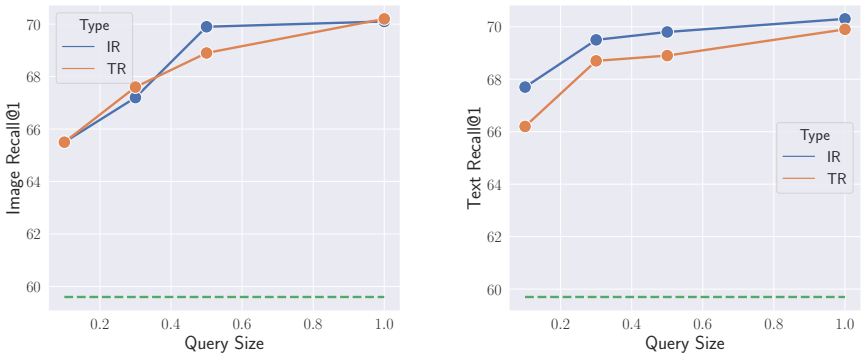
Surrogate Dataset	Surrogate Type	Image-to-text Retrieval R@1	R@5	R@10	Agr.	Text-to-image Retrieval R@1	R@5	R@10	Agr.
Flickr30K (1K test set)	IR	70.3	90.7	94.4	88.2	70.1	89.0	94.7	89.1
	TR	69.9	90.7	94.4	87.9	70.2	89.9	95.0	88.1
MSCOCO (5K test set)	IR	41.9	68.8	78.8	73.7	40.4	66.9	77.2	74.1
	TR	41.8	68.5	79.2	73.9	41.3	67.4	77.7	74.6

the fraction of times the matched texts/images appear in the top-K retrieved results. Agreement measures the fraction of data where the target and surrogate models generate the same retrieval results. In this paper, we define the agreement metric at the Recall@1 level – counting when the best matching retrieval results for the target and surrogate model are the same. As we follow the Karpathy split Karpathy & Fei-Fei (2017), the results on the MSCOCO test set that has 5K image-text pairs are naturally lower because Recall@K and the agreement are functions of the test set size.

4.2 RESULTS

We first exhibit the image-text retrieval results of the target models on the Flickr30K and MSCOCO datasets in Table 1. We use "zero-shot" to denote the case where the victim uses the base model, and "fine-tuned" indicates the victim fine-tunes the base model to adapt their target tasks. Both settings are evaluated on the corresponding test sets. Compared to the zero-shot setting, the fine-tuned CLIP model gains improvements in both image-to-text retrieval and text-to-image retrieval tasks, especially by a large margin at the R@1 level. Take the text-to-image retrieval task as an example: the fine-tuned version achieves an improvement of 12.3% R@1 on the Flickr30K dataset and 13.0% R@1 on the MSCOCO dataset. This illustrates the necessity of fine-tuning when the victim aims to apply CLIP to their own tasks, which also explains why we focus on stealing fine-tuned models instead of base models. In the following experiments, we treat the zero-shot retrieval results as the baseline. We report the main results of our attacks in Table 2. In this part, we consider the case where the attacker has access to the dataset sampled from the same distribution as the target task, and the number of images/texts is the same as the target dataset $\mathcal{D}_{\mathcal{F}}$. We investigate the influence of the query budget and dataset distribution in Section 4.3.

As forementioned, there are two APIs which may be exploited to launch the attacks. The attacker can construct the surrogate dataset from either the text-to-image retrieval API or the image-to-text retrieval API. The results show that our attack has encouraging performance, and the surrogate model has negligible utility deterioration compared to the target models. For example, both surrogate models drop less than 1.3% R@K w.r.t the image-to-text retrieval and 1.8% difference w.r.t the text-to-image retrieval R@K on the Flickr30K dataset. We also find that the model stealing attacks are **API-agnostic**, as we achieve comparable attack performance via these two APIs, indicating that information leaked by images is close to that leaked by texts. We attribute this to the fact that both the images and texts are mapped from their domains to a shared latent space. On the other hand, the performance on different datasets varies.



(a) Text-to-image retrieval (b) Image-to-text retrieval

Figure 2: Image-text retrieval Recall@1 of two types of surrogate models with varying query budgets on the Flickr30K dataset. More results can be seen in Figure 5 of Appendix A.

Table 3: Image-text retrieval on the Flickr30K dataset for surrogate models $\mathcal{M}_S^{\mathcal{TR}}$ and $\mathcal{M}_S^{\mathcal{IR}}$ with mismatched architectures/models between the model owner and the attacker. More results on the MSCOCO dataset can be seen in Table 7 of Appendix A.

Target Architecture	Surrogate Architecture	Surrogate Type	Image-to-text Retrieval R@1	R@5	R@10	Text-to-image Retrieval R@1	R@5	R@10
ResNet-50	ResNet-50	IR	64.4	88.0	93.4	61.0	86.1	91.9
		TR	64.3	88.2	93.4	61.0	85.8	91.9
ResNet-50	ViT-B/32	IR	68.5	90.4	94.5	69.6	89.6	94.1
		TR	69.6	89.8	94.1	70.4	89.0	93.2
ViT-B/32	ResNet-50	IR	64.7	88.8	93.6	62.5	87.5	92.4
		TR	66.8	89.6	93.9	64.5	89.1	93.5
ViT-B/32	ViT-B/32	IR	70.3	90.7	94.4	70.1	89.0	94.7
		TR	69.9	90.7	94.4	70.2	89.9	95.0
CLIP (ViT-B/32)	BLIP	IR	81.1	94.6	97.0	82.0	95.3	97.9
		TR	81.7	95.3	96.7	81.9	95.0	97.4

We find that the attack on the MSCOCO dataset is less effective than the attack on the Flickr30K dataset. We reason this is because the target model trained on the MSCOCO dataset is less effective in terms of the R@K metric; thus, the quality of the constructed surrogate dataset is worse. Moreover, as the feature bases used for these two tasks have different magnitudes, the baseline recalls for both tasks are different. The agreement evaluation also evinces the effectiveness of our attacks. We further confirm this attack can be generalized to other vision-language models (e.g., BLIP (Li et al., 2022)) with the same conclusion, concrete results can be found in Table 8.

4.3 Advantageous Properties

The experimental results in Section 4.2 illustrate the effectiveness of our attack with the **API-agnostic** property. However, we have made a series of attack assumptions; for example, the attacker has access to the dataset sampled from the same distribution of the target dataset. In this section, we demonstrate the existence of other advantageous properties of our attacks by relaxing the attack assumptions.

Query-efficient. We start by investigating the influence of query budgets on attack performance. The varying query budgets are described as different fractions of the original fine-tuning dataset size. Figure 2 reports the retrieval performance (Recall@1) of different types of surrogate models fine-tuned and evaluated on the Flickr30K dataset. As we can observe, the model stealing attacks with small queries are still successful (e.g., only 10% $|\mathcal{D}_{\mathcal{F}}|$ queries gain an improvement of at least 6%). Although more queries benefit more for model stealing, performance gains diminish. More results on the MSCOCO dataset are in Figure 5 of Appendix A and same conclusions can be drawn.

Table 4: Image-text retrieval performance on two test sets. The surrogate models are fine-tuned either on the same distribution dataset or on the CC3M dataset.

Test Dataset	Surrogate Type	Fine-tune Dataset	Image-to-text Retrieval R@1	R@5	R@10	Text-to-image Retrieval R@1	R@5	R@10
Flickr30K (1K test set)	IR	Flickr30K	70.3	90.7	94.4	70.1	89.0	94.7
		CC3M	67.7	89.8	94.0	65.4	87.6	93.0
	TR	Flickr30K	69.9	90.7	94.4	70.2	89.9	95.0
		CC3M	67.5	89.6	93.9	67.8	89.3	94.1
MSCOCO (5K test set)	IR	MSCOCO	41.9	68.8	78.8	40.4	66.9	77.2
		CC3M	40.7	67.5	76.8	38.5	64.0	74.4
	TR	MSCOCO	41.8	68.5	79.2	41.3	67.4	77.7
		CC3M	40.2	66.7	77.0	40.9	66.2	76.8

Architecture-agnostic. In the previous experiments, we assumed both the target and surrogate models fine-tune the pretrained CLIP model that has a ViT-B/32 image encoder and a language Transformer as the text encoder. However, it is likely that the attacker has no knowledge about the target architecture. Therefore, we investigate the variants of different image backbones of surrogate models. Besides the Vision Transformer, CLIP provides other image backbones, e.g., ResNet-50 and ResNet-101 (He et al., 2016). Here, we leverage ResNet-50 as the image encoder and measure the retrieval results on the Flickr30K test set when the attacker and the model owner use mismatched model architectures. As shown in Table 3, given a fixed target architecture, both the text-to-image retrieval and image-to-text retrieval Recall@K are always higher when the attacker leverages ViT-B/32 as the surrogate architecture, even when the target model is initialized with ResNet-50. Additionally, when we fix the surrogate architecture, the retrieval Recall@K is always higher with a more powerful target architecture because the attacker can construct a high-quality surrogate dataset through it. We further investigate if the non-CLIP pretrained model can still be considered as the surrogate model. In specific, we leverage BLIP, a more recent state-of-the-art work outperforming CLIP by a large margin in a variety of vision-language tasks. We can observe the improved performance of the BLIP surrogate model, compared to the original surrogate setting, especially by a large margin on image-to-text Recall@1 (+11.3% on average) and text-to-image Recall@1 (+11.8% on average). Overall, the results suggest that adversaries can maximize the performance of surrogate models by fine-tuning more powerful pretrained architectures/models, regardless of the target architecture. Hence, we conclude that our attacks are architecture-agnostic. More results on the MSCOCO dataset in Table 7 of Appendix A and we can draw the same conclusion.

Data-agnostic. So far we leverage the auxiliary dataset that comes from the same distribution of the original fine-tuning dataset to query the target model. Here, we relax the assumption by leveraging an auxiliary dataset with a different distribution. Specifically, we use the Conceptual Captions dataset (Sharma et al., 2018) (abbreviated as CC3M), which has 3 million image-caption pairs scraped from the Internet. We randomly sample an identical number of queries as the original fine-tuning dataset $\mathcal{D}_{\mathcal{F}}$ to construct a different distribution auxiliary dataset. As we can observe in Table 4, both surrogate models trained on the CC3M dataset work similarly well on two different test sets. They achieve similar performance on the MSCOCO test set by a marginal difference on image-to-text Recall@1 (-1.4% on average) and text-to-image Recall@1 (-1.2% on average). Although the surrogate models perform worse on the Flickr30K test set, they still outperform the baseline (i.e., zero-shot retrieval) by over 7.9% R@1 in the image-to-text task and over 7.0% R@1 in the text-to-image task. We can also observe that the $\mathcal{M}_S^{\mathcal{TR}}$ slightly outperforms $\mathcal{M}_S^{\mathcal{IR}}$ in the text-to-image retrieval task when both of them are fine-tuned on the subset of CC3M dataset. We later show that text queries from the different distribution dataset (i.e., CC3M dataset) have higher agreements, which helps the surrogate model $\mathcal{M}_S^{\mathcal{TR}}$ achieve better retrieval performance.

Besides using the auxiliary dataset that comes from a different distribution, we consider a more challenging scenario where only random inputs are leveraged to launch the attack. For the text-to-image retrieval, we use randomly generated captions as input. Specifically, we first construct a vocabulary using the in-domain dataset. For example, if the target model is fine-tuned on the Flickr30k subset, we use all captions from the disjoint Flickr30K subset with an identical size to construct the vocabulary. We use all the uni-gram tokens that appear in the captions to form the

Table 5: Image-text retrieval performance on two test sets. The $\mathcal{D}_{\mathcal{S}}^{\mathcal{IR}}$ is composed of random captions and the best matching images via the $\mathcal{IR}$ API. The $\mathcal{D}_{\mathcal{S}}^{\mathcal{TR}}$ is composed of random noise images and the best matching captions via the $\mathcal{TR}$ API.

Test Dataset	Surrogate Type	Image-to-text Retrieval			Text-to-image Retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10
Flickr30K (1K test set)	IR	67.9	90.3	93.6	65.9	87.3	93.3
	TR	59.9	85.5	91.3	90.3	84.4	89.9
MSCOCO (5K test set)	IR	41.6	68.5	79.2	41.3	67.4	77.7
	TR	33.5	59.1	69.8	30.0	55.0	65.8

vocabulary. Following the previous work Krishna et al. (2020), we generate nonsensical input via uniformly randomly sampling tokens from the vocabulary up to the chosen length. The chosen length is the most frequently occurring length in the caption sets. For the image-to-text retrieval, we directly generate random noise images as input. In Table 5, the results show that we can still achieve competitive performance, leveraging random captions as the input of $\mathcal{IR}$ API. When evaluating on the MSCOCO test set, the $\mathcal{M}_{\mathcal{S}}^{\mathcal{IR}}$ fine-tuned on random captions even outperforms the $\mathcal{M}_{\mathcal{S}}^{\mathcal{IR}}$ fine-tuned on the different distribution dataset. Meanwhile, the $\mathcal{M}_{\mathcal{S}}^{\mathcal{TR}}$ fine-tuned on the random noise image fails, as random noises provide almost no information for the model optimization.

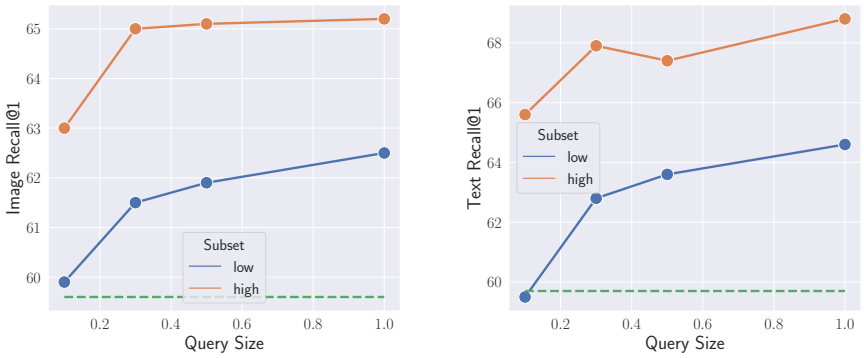
In conclusion, we show our attacks have four advantageous properties: query-efficient, API-agnostic, architecture-agnostic, and data-agnostic. These favorable properties make the model stealing attacks a practical threat against the vision-language models.

4.4 Analysis

Which type of queries are beneficial to our attacks. Previous work (Krishna et al., 2020) shows that high-agreement queries are better for BERT-based NLP model extraction. Here, we investigate if the same conclusion can be drawn from the fine-tuned CLIP models. We first train five target models on two benchmark datasets, respectively. The only varying variable in their training processes is the random seed. The retrieval performance of these target models is similar. Then, we measure the agreement among the outputs returned by these models for different types of queries, i.e., images and texts. We sample a subset of 10 times the size of the original fine-tuning dataset from CC3M (i.e., 10x $|\mathcal{D}_{\mathcal{F}}|$) to query these target models and the results of agreement are shown in Figure 4 (Appendix A). The agreement values, ranging from 1 to 5, represent the number of models with the same outputs, and a value of 1 indicates that the outputs of these five models are different. As we can observe, the overall agreements for image-to-text retrieval are much higher than for text-to-image retrieval on both sets of target models. For example, the highest agreement value accounts for close to 40% of the image-to-text retrieval agreements, yet only about 10% of text-to-image retrieval agreements on two sets of target models. This explains why $\mathcal{M}_{\mathcal{S}}^{\mathcal{TR}}$ outperforms $\mathcal{M}_{\mathcal{S}}^{\mathcal{IR}}$ when both surrogate models are fine-tuned on the subsets of CC3M dataset in Table 4. To further investigate if the higher-agreement queries are more beneficial for model stealing attacks, we sort all text queries for text-to-image retrieval mentioned above by their agreements and select the highest and lowest agreement subsets with an identical size of the original fine-tuning dataset $\mathcal{D}_{\mathcal{F}}$ to construct two types of surrogate datasets $\mathcal{D}_{\mathcal{S}}^{\mathcal{IR}}$ and $\mathcal{D}_{\mathcal{S}}^{\mathcal{TR}}$. Figure 3 shows the image/text Recall@1 results evaluated on the Flickr30K test set. The surrogate models are trained on the lowest and highest subsets whose sizes are varying fractions of the $|\mathcal{D}_{\mathcal{F}}|$. We can see improvements of both image and text Recall@1 when constructing surrogate models using high-agreement subsets, constantly outperforming low-agreement subsets of identical sizes by over 3.1% Recall@1 on text-to-image retrieval and 4.2% on image-to-text retrieval. This validates that the agreement among different target models benefits the construction of surrogate datasets. More results on the MSCOCO dataset can be seen in Figure 6 of Appendix A and we can come to the same conclusion.

4.5 Defense

Having demonstrated that the fine-tuned CLIP model is vulnerable to the model stealing attacks, we now concentrate on the defense mechanism. We apply **out-of-distribution detection** (Hendrycks & Gimpel, 2017; Liang et al., 2018; Lee et al., 2018), which has been widely used to safely deploy



(a) Text-to-image retrieval (b) Image-to-text retrieval

Figure 3: Image-text retrieval results (Recall@1) on the Flickr30K test set. The surrogate models are fine-tuned on the highest and lowest subsets of CC3M datasets with different fractions of the original fine-tuning dataset size $|\mathcal{D}_{\mathcal{F}}|$. Subsets are selected according to the agreement between the retrieval results of different runs of the target model. More results can be seen in Figure 6 of Appendix A.

ML models in real-world scenarios via detecting anomalous queries. Specifically, we treat the out-of-distribution detection as a binary classification problem. We label examples from the original fine-tuning dataset $\mathcal{D}_{\mathcal{F}}$ as in-distribution and randomly draw the same amount of data as $\mathcal{D}_{\mathcal{F}}$ from the CC3M dataset to mark as out-of-distribution. As in the construction of the surrogate models, the out-of-distribution classifier O can be constructed using either images or texts. We leverage the image embeddings extracted from the target model as input features to train the image classifier $\mathcal{O}_{\mathcal{I}}$. Meanwhile, we train the text classifier $\mathcal{O}_{\mathcal{T}}$ using the text embeddings generated by the target model as input features. As shown in Table 6, both the $\mathcal{O}_{\mathcal{I}}$ and $\mathcal{O}_{\mathcal{T}}$ work well when examining on either images or texts that are sampled from a different distribution than the fine-tuning data distribution. However, this defense only works well with malicious queries we pre-defined and we cannot define all types of malicious queries.

5 Conclusion

We investigate the model stealing attacks against fine-tuned CLIP models via image-text retrieval APIs. We demonstrate that our attacks enabled by these two APIs can be a practical threat in real-world scenarios. Specifically, the results show that the model stealing attacks against vision-language models are data-efficient, API-agnostic, architecture-agnostic, and data-agnostic. We finally show that high-agreement queries benefit the high-quality of surrogate datasets. Unfortunately, existing defenses such as out-of-distribution detection are impotent against our attacks without strong assumptions. Although our attacks can work well in extensive scenarios, they still suffer from the limitation that requires the attackers to put effort into collecting images or text queries from the Internet. One future direction can be reducing the cost of constructing the auxiliary dataset, e.g., building nonsensical texts generators. Although such attacks can be implemented in a wide range of realistic scenarios with little technical skill requiring, we believe the profit of releasing our attacks exceeds the potential harms, as we facilitate the development of solid defense mechanisms. Our research is thus a call to action, which pushes the community to impede the unprotected release and prevalence of powerful vision-language models and puts more focus on their IP protections.

Table 6: Out-of-Distribution Detection

Dataset	Type	Accuracy	AUC Score
Flickr30K	Image	96.6	99.4
	Text	96.5	98.8
MSCOCO	Image	95.2	98.9
	Text	95.1	98.9

REFERENCES

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 2425–2433. IEEE, 2015.

Nicholas Carlini and Andreas Terzis. Poisoning and Backdooring Contrastive Learning. *CoRR abs/2106.09667*, 2021.

Varun Chandrasekaran, Kamalika Chaudhuri, Irene Giacomelli, Somesh Jha, and Songbai Yan. Exploring Connections Between Active Learning and Model Extraction. In *USENIX Security Symposium (USENIX Security)*, pp. 1309–1326. USENIX, 2020.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO Captions: Data Collection and Evaluation Server. *CoRR abs/1504.00325*, 2015.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Stefan Lee, José M. F. Moura, Devi Parikh, and Dhruv Batra. Visual Dialog. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

David DeFazio and Arti Ramesh. Adversarial Model Extraction on Graph Neural Networks. *CoRR abs/1912.07721*, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. IEEE, 2016.

Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Hailong Hu and Jun Pang. Stealing Machine Learning Models: Attacks and Countermeasures for Generative Adversarial Networks. In *Annual Computer Security Applications Conference (ACSAC)*, pp. 1–16. ACM, 2021.

Matthew Jagielski, Nicholas Carlini, David Berthelot, Alex Kurakin, and Nicolas Papernot. High Accuracy and High Fidelity Extraction of Neural Networks. In *USENIX Security Symposium (USENIX Security)*, pp. 1345–1362. USENIX, 2020.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *International Conference on Machine Learning (ICML)*, pp. 4904–4916. PMLR, 2021.

Mika Juuti, Sebastian Szyller, Samuel Marchal, and N. Asokan. PRADA: Protecting Against DNN Model Stealing Attacks. In *IEEE European Symposium on Security and Privacy (Euro S&P)*, pp. 512–527. IEEE, 2019.

Andrej Karpathy and Li Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

Kalpesh Krishna, Gaurav Singh Tomar, Ankur P. Parikh, Nicolas Papernot, and Mohit Iyyer. Thieves on Sesame Street! Model Extraction of BERT-based APIs. In *International Conference on Learning Representations (ICLR)*, 2020.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 7167–7177. NeurIPS, 2018.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven Chu-Hong Hoi. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 9694–9705. NeurIPS, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. *CoRR abs/2201.12086*, 2022.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Norman Mu, Alexander Kirillov, David A. Wagner, and Saining Xie. SLIP: Self-supervision meets Language-Image Pre-training. *CoRR abs/2112.12750*, 2021.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. *CoRR abs/2112.10741*, 2021.

Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Knockoff Nets: Stealing Functionality of Black-Box Models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4954–4963. IEEE, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763. PMLR, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-Shot Text-to-Image Generation. In *International Conference on Machine Learning (ICML)*, pp. 8821–8831. JMLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. *CoRR abs/2204.06125*, 2022.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2556–2565. ACL, 2018.

Yun Shen, Xinlei He, Yufei Han, and Yang Zhang. Model Stealing Attacks Against Inductive Graph Neural Networks. In *IEEE Symposium on Security and Privacy (S&P)*. IEEE, 2022.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A Corpus for Reasoning about Natural Language Grounded in Photographs. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 6418–6428. ACL, 2019.

Sebastian Szyller, Vasisht Duddu, Tommi Gröndahl, and N. Asokan. Good Artists Copy, Great Artists Steal: Model Extraction Attacks Against Image Translation Generative Adversarial Networks. *CoRR abs/2104.12623*, 2021.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. Stealing Machine Learning Models via Prediction APIs. In *USENIX Security Symposium (USENIX Security)*, pp. 601–618. USENIX, 2016.

Jean-Baptiste Truong, Pratyush Maini, Robert J. Walls, and Nicolas Papernot. Data-Free Model Extraction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4771–4780. IEEE, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 5998–6008. NIPS, 2017.

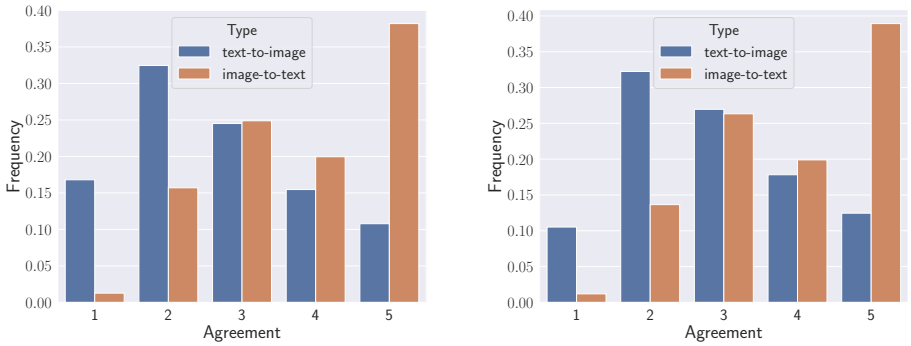
Bang Wu, Xiangwen Yang, Shirui Pan, and Xingliang Yuan. Model Extraction Attacks on Graph Neural Networks: Taxonomy and Realization. *CoRR abs/2010.12751*, 2020.

Fei Yan and Krystian Mikolajczyk. Deep Correlation for Matching Images and Text. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3441–3450. IEEE, 2015.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2014.

A APPENDIX

Our code and data are available at `https://anonymous.4open.science/r/vl_model_steal-1D9F`.



(a) Flickr30K (b) MSCOCO

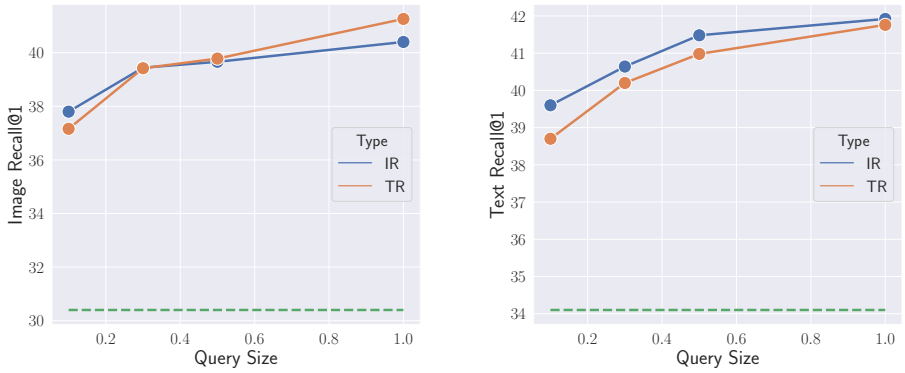
Figure 4: Histogram of retrieval agreements among five target models with varying random seeds fine-tuned on two benchmark datasets. The query sets come from the CC3M dataset. For both settings, the image-to-text retrieval in general obtain higher agreements than the text-to-image retrieval.

Table 7: Image-text retrieval on the MSCOCO dataset for surrogate models $\mathcal{M}_S^{\mathcal{TR}}$ and $\mathcal{M}_S^{\mathcal{IR}}$ with mismatched architectures/models between the model owner and the attacker.

Target Architecture	Surrogate Architecture	Type	Image-to-text Retrieval			Text-to-image Retrieval		
			R@1	R@5	R@10	R@1	R@5	R@10
ResNet50	ResNet50	IR	38.3	64.1	74.2	34.9	60.3	71.7
		TR	38.0	63.6	74.2	34.3	60.9	71.5
ResNet50	ViT-B/32	IR	42.4	68.8	78.3	40.5	67.1	77.4
		TR	41.4	68.2	78.4	40.5	67.2	77.6
ViT-B/32	ResNet50	IR	39.7	65.7	75.9	35.2	61.5	72.9
		TR	39.3	65.4	75.7	35.3	62.3	72.7
ViT-B/32	ViT-B/32	IR	41.9	68.8	78.8	40.4	66.9	77.2
		TR	41.8	68.5	79.2	41.3	67.4	77.7
CLIP (ViT-B/32)	BLIP	IR	55.6	80.6	87.4	56.9	80.7	87.4
		TR	55.4	81.0	88.1	57.0	80.7	87.8

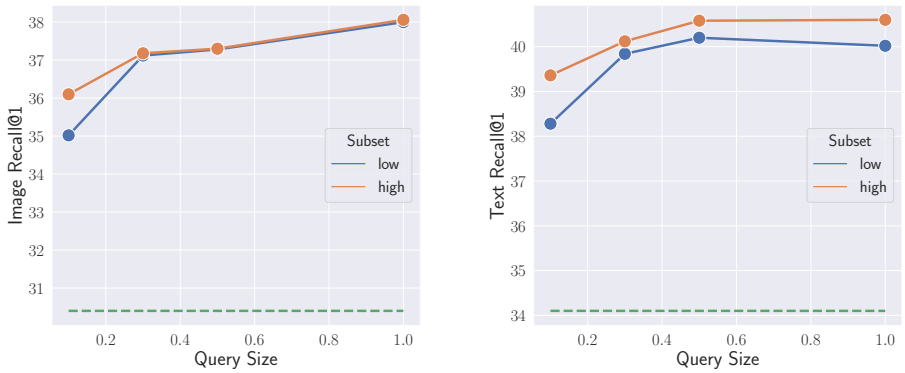
Table 8: Image-text retrieval results evaluated on the Flickr30K and MSCOCO dataset. Both the surrogate models and target models are BLIP.

Target Dataset	Type	Settings	Image-to-text Retrieval			Text-to-image Retrieval		
			R@1	R@5	R@10	R@1	R@5	R@10
Flickr30K (1K test set)	Target	Zero-shot	60.5	83.8	89.3	77.6	93.1	96.2
	Target	Fine-tuned	85.0	96.3	98.0	85.5	96.1	97.9
	Surrogate (IR)	Fine-tuned	83.2	95.8	97.7	84.8	96.1	97.5
	Surrogate (TR)	Fine-tuned	83.0	96.8	98.1	84.6	95.9	97.7
MSCOCO (5K test set)	Target	Zero-shot	51.5	76.5	84.6	55.9	79.5	86.8
	Target	Fine-tuned	61.7	84.9	91.4	62.0	84.1	90.3
	Surrogate (IR)	Fine-tuned	60.6	83.0	89.3	60.8	83.1	88.9
	Surrogate (TR)	Fine-tuned	60.5	83.6	89.7	61.0	83.6	89.6



(a) Text-to-image retrieval (b) Image-to-text retrieval

Figure 5: Image-text retrieval Recall@1 of two types of surrogate models with varying query budgets on the MSCOCO dataset.



(a) Text-to-image retrieval (b) Image-to-text retrieval

Figure 6: Image-text retrieval results at the Recall@1 level on the MSCOCO test set. The surrogate models are fine-tuned on the highest and lowest subsets of CC3M datasets with different fractions of the original fine-tuning dataset size $|\mathcal{D}_{\mathcal{F}}|$. Subsets are selected according to the agreement between the retrieval results of different runs of the target model.