ViP: A Differentially Private Foundation Model for Computer Vision

Yaodong Yu [1 2] **Maziar Sanjabi** [2] **Yi Ma** [1] **Kamalika Chaudhuri** [2] **Chuan Guo** [2]

Abstract

Artificial intelligence (AI) has seen a tremendous surge in capabilities thanks to the use of *foundation models* trained on internet-scale data. On the flip side, the uncurated nature of internet-scale data also poses significant privacy and legal risks, as they often contain personal information or copyrighted material that should not be trained on without permission. In this work, we propose as a mitigation measure a recipe to train foundation vision models via self-supervised learning with differential privacy (DP) guarantee. We identify masked autoencoders as a suitable learning algorithm that aligns well with DP-SGD, and train *ViP*—a **Vi**sion transformer with differential **P**rivacy—under a strict privacy budget of $\epsilon = 8$ on the LAION400M dataset. We evaluate the quality of representation learned by ViP using standard downstream vision tasks; in particular, ViP achieves a (non-private) linear probing accuracy of 55.7% on ImageNet, comparable to that of end-to-end trained AlexNet (trained and evaluated on ImageNet). Our result suggests that scaling to internet-scale data can be practical for private learning. Code and DP pre-trained models are available at `https://github.com/facebookresearch/ViP-MAE`.

1. Introduction

Foundation models (*e.g.*, GPT-3, SimCLR, CLIP, *etc.* (Brown et al., 2020; Chen et al., 2020a; Radford et al., 2021)) pre-trained on vast amounts of diverse unlabeled data through self-supervised learning (SSL) have emerged as an important building block for artificial intelligence (AI) systems (Bommasani et al., 2021). These foundation models enable downstream applications via fine-tuning, prompting, or training a simpler model on top of the learned representations to perform more specialized tasks, and have performed tremendously well on challenging benchmarks in both language and vision domains (Brown et al., 2020; Radford et al., 2021; Touvron et al., 2023).

Despite the widespread deployment of foundation models, there are significant privacy and legal risks of training these models on uncurated data that often contain personal information or copyrighted material. Although the training data for these models are considered *public* in most cases, some of the data may be sensitive (Tramèr et al., 2022); additionally, there are certain privacy and copyright laws that apply to model training even on such *public* data (Henderson et al., 2023). In addition, studies have shown that generative foundation models such as GPT-3 can sometimes regurgitate memorized information about individuals and licensed content from its training data when prompted to do so (Carlini et al., 2021). More recently, Meehan et al. (2023) showed that non-generative vision SSL models can also be probed to reveal sensitive information about individual samples in its training data when given partial information.

Given these risks, there is an urgent need to train foundation models that can adhere to relevant privacy and copyright laws. To this end, differential privacy (DP; Dwork et al. (2006)) seeks to limit the influence of individual training data points on the trained model, and hence has the potential to mitigate both privacy and copyright risks for sensitive information that is confined to a single or a few training examples (Henderson et al., 2023). For any model that can be trained using gradient-based optimization, DP-SGD (Song et al., 2013; Abadi et al., 2016) can be applied to ensure that the trained model satisfies the rigorous definition of DP. However, there are still significant technical challenges in DP-SGD training of large-scale foundation vision models:

1. Differentially private representation learning in general is a difficult problem. Tramer & Boneh (2020) showed that even handcrafted features can outperform feature learned by state-of-the-art DP-trained models, and attaining high-utility learned representations requires significantly more training data—much more than what is provided in typical supervised/curated datasets.

2. Combining self-supervised learning (SSL) with internet-scale *uncurated* datasets may seem like a natural ap-

[1]UC Berkeley [2]Meta. Correspondence to: Yaodong Yu <yyu@eecs.berkeley.edu>, Chuan Guo <chuanguo@meta.com>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

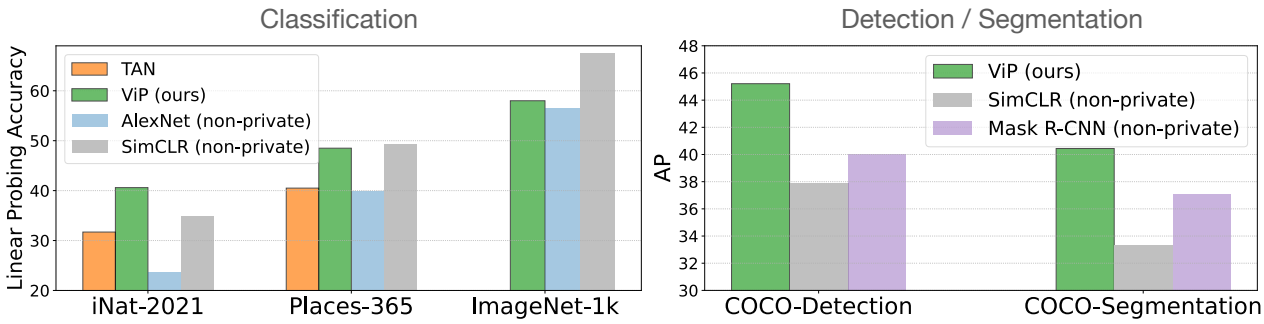


Figure 1. (**left**) Linear probing accuracies of TAN (Sander et al., 2022) (state-of-the-art DP training method), AlexNet (Krizhevsky et al., 2017), SimCLR (Chen et al., 2020a) and ViP—our DP-trained model with $\epsilon = 8$. ViP-Large can achieve similar transfer learning result as SimCLR on iNat-2021 and Places-365, and achieves similar accuracy on ImageNet as end-to-end trained AlexNet. (**right**) Average precision (AP) evaluations of SimCLR (Chen et al., 2020a), Mask R-CNN (He et al., 2017) and ViP-Base on MS-COCO. Our DP-trained model outperforms both SimCLR and Mask R-CNN.

proach to gain access to the large amount of data needed for DP training. However, most vision SSL training algorithms are based on *contrastive learning*, where the objective function depends on multiple samples in an entangled manner. This makes it difficult to perform the per-sample gradient computation needed in DP-SGD.

3. SSL training requires a much larger number of training epochs compared to supervised learning, which sharply increases the DP parameter ϵ, leading to meaningless privacy guarantees.

In this paper, we describe a successful recipe for training differentially private large-scale vision foundation models via SSL. Firstly, we identify masked autoencoder (MAE; He et al. (2022)) as a promising SSL training algorithm that is amenable to DP-SGD. MAE uses an instance-separable loss function and does not require batch normalization, and hence per-sample gradients can be easily computed. We also show that it is tolerant to the large amount of Gaussian noise added in DP-SGD. Next, we demonstrate that MAE can effectively leverage synthetic datasets containing only programmatically-generated synthesized textures (Baradad et al., 2022) to warm-start the DP training process, significantly reducing the number of training epochs required to reach a high-utility model. The combination of these two ingredients forms a powerful DP training recipe for obtaining high-utility differentially private foundation vision models.

We implement this training recipe on the LAION400M dataset (Schuhmann et al., 2021). We show that the resulting model, which we call *ViP* (**Vi**sion transformer with differential **P**rivacy), learns highly useful and transferable representations—*rivaling that of representation learned by SimCLR on ImageNet*—while providing a strong DP guarantee with $\epsilon = 8$. In Figure 1, we compare ViP with other private and non-private models in terms of downstream linear probing accuracy and fine-tuning accuracy for different image datasets:

- For iNat-2021 and Places-365 classification, ViP outperforms both TAN (Sander et al., 2022)—the previous SOTA for DP supervised training—and AlexNet (Krizhevsky et al., 2017), while matching or exceeding the performance of SimCLR pre-trained on ImageNet.
- On ImageNet, the linear probing accuracy of ViP matches that of end-to-end trained AlexNet[1].
- On MS-COCO detection and segmentation, ViP outperforms both SimCLR pre-trained on ImageNet and Mask R-CNN.

Our experiments demonstrate that by scaling DP-SGD training to vast amounts of unlabeled data and using synthetic data to warm-start the model, we can attain high-utility foundation vision models under stringent privacy guarantees. Consequently, we hope that future work can continue to build on our successful recipe and further push the performance boundary of large-scale DP training.

2. Background

Differential privacy (Dwork et al., 2014) is a mathematical framework for formal reasoning about information leakage through a private mechanism. A learning algorithm $\mathcal{A}$ is said to be (ϵ, δ)-*differentially private* (denoted (ϵ, δ)-DP) if for all training datasets $\mathcal{D}, \mathcal{D}'$ that differ[2] in a single training sample, we have:

$$P(\mathcal{A}(\mathcal{D}) \in S) \leq e^{\epsilon} P(\mathcal{A}(\mathcal{D}') \in S) + \delta \qquad (1)$$

for all outcome sets S. More generally, equation 1 can be expressed as a statistical divergence $D(\mathcal{A}(\mathcal{D})||\mathcal{A}(\mathcal{D}'))$ between the distribution of models trained on $\mathcal{D}$ vs. $\mathcal{D}'$, with (ϵ, δ)-DP corresponding to the "hockey-stick" divergence (Sharma & Warsi, 2013). Another useful variant is

[1]The model is sourced from the PyTorch website and is end-to-end trained with supervised learning.

[2]We adopt the removal notion of adjacency, *i.e.*, $\mathcal{D}' = \mathcal{D} \cup \mathbf{z}$ for some $\mathbf{z}$ and vice versa.

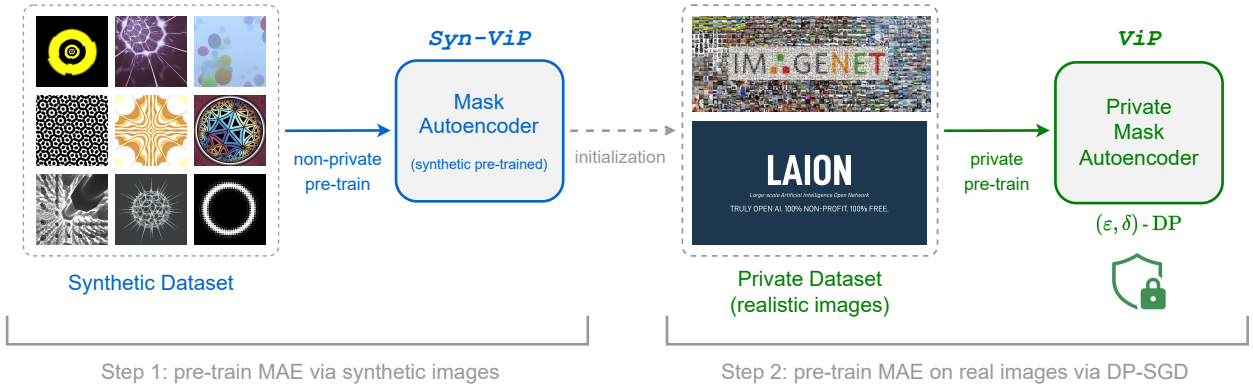


Figure 2. **How to pre-train differentially private transformers (*ViP*) with synthetic data?** In Step 1, we first pre-train a MAE model on synthetic images with standard optimizers (*e.g.*, SGD, AdamW). We denote this model by *(Syn)-ViP*. In Step 2, we use the MAE model pre-trained on synthetic images as initialization, and then apply differential private optimizers (*e.g.*, DP-SGD, DP-AdamW) to train a *ViP* model that satisfies (ϵ, δ)-DP.

Rényi differential privacy (RDP; (Mironov, 2017)), which uses the Rényi divergence D_α (Rényi, 1961): $\mathcal{A}$ is said to be (α, ϵ)-RDP if $D_\alpha(\mathcal{A}(\mathcal{D})||\mathcal{A}(\mathcal{D}')) \leq \epsilon$. Moreover, RDP can be converted to DP via the following (Balle et al., 2020): if $\mathcal{A}$ is $(\alpha, \epsilon_\alpha)$-RDP then it is also (ϵ, δ)-DP with

$$\epsilon = \epsilon_\alpha + \log\left(\frac{\alpha - 1}{\alpha}\right) - \frac{\log\delta + \log\alpha}{\alpha - 1}. \tag{2}$$

DP-SGD training. Abadi et al. (2016) showed that stochastic gradient descent (SGD)—the quintessential learning algorithm—can be made differentially private by perturbing the per-iteration gradient with Gaussian noise. The modified SGD update with gradient perturbation (often referred to as *DP-SGD*) is given by $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \widetilde{\boldsymbol{g}}_t$, and

$$\widetilde{\boldsymbol{g}}_t = \frac{1}{B}\left(\sum_{\mathbf{x}\in\mathcal{B}_t} \mathsf{clip}_C(\nabla_{\boldsymbol{\theta}}\ell(\mathbf{x}; \boldsymbol{\theta}_t) + \mathcal{N}(0, \sigma^2 C^2 \boldsymbol{I})\right), \tag{3}$$

where η_t is the learning rate, $\mathcal{B}_t$ is the sampled batch, B is the average batch size, $\sigma > 0$ is the noise multiplier, and clip_C is the operation that clips the per-sample gradient norm to at most $C > 0$. It can be shown that this update procedure is $(\alpha, \epsilon_\alpha)$-RDP for some computable ϵ_α (Mironov et al., 2019). The end-to-end learning algorithm by running T iterations of SGD is thus $(\alpha, T\epsilon_\alpha)$-RDP via composition (Mironov, 2017), and a conversion to (ϵ, δ)-DP can be obtained using equation 2. Such privatization mechanism—per-sample clipping and injecting noise—can be easily integrated with other first-order optimization algorithms such as Adam (Kingma & Ba, 2014) and AdamW (Loshchilov & Hutter, 2017).

Self-supervised learning (SSL) has emerged as a prominent approach for scaling up the training of machine learning models to large-scale unlabeled datasets. Restricting our attention to the vision domain, SSL pre-trained models generalize effectively across a wide range of transfer learning downstream tasks such as classification, instance segmentation and object detection (Chen et al., 2020b; Bommasani et al., 2021), especially under the scenario of limited downstream training data. Vision SSL methods can be broadly categorized as either *joint embedding-based learning* (JE) (Chen et al., 2020a; He et al., 2020; Grill et al., 2020; Zbontar et al., 2021; Chen & He, 2021) or *reconstruction-based learning* (REC) (Bao et al., 2021; Xie et al., 2022; He et al., 2022). JE-based approaches design objective functions so that all views (or image augmentations) of the same sample have similar embeddings, while views of different samples have different embeddings. As a result, most JE-based approaches *require* a batch containing multiple samples in order to define the objective function. On the other hand, REC-based approaches aim to optimize models to reconstruct image inputs in the pixel space based on partially masked inputs, which promotes the model to learn compressed representations that can generalize well.

Related Work

Scaling up DP training. Recently, an expanding body of literature has emerged on scaling DP training to large-scale datasets and models in both NLP and vision domains. In NLP, a series of works (Yu et al., 2021; Li et al., 2022a) showed that by combining public pre-training and scaling up the training batch size, it is possible to fine-tune the pre-trained language model to achieve reasonable downstream performance. In computer vision, Kurakin et al. (2022) first attempted to scale DP training of convolutional neural networks (ResNets) to ImageNet. De et al. (2022) further improved the performance of Kurakin et al. (2022) with a Normalizer-Free ResNet architecture and an improved training recipe. More recently, Sander et al. (2022) proposed a more efficient hyperparameter tuning method for DP training that led to state-of-the-art performance on ImageNet. It is worth noting that all these works on DP-trained computer vision models focus on training supervised models.

DP pre-training. There are existing work on exploring the possibility of learning effective representations with large-scale differentially private pre-training. In particular, Anil et al. (2021) studied how to DP pre-train BERT models and Ponomareva et al. (2022) focused on pre-training T5 models (Raffel et al., 2020) with DP. The main difference between this work and Anil et al. (2021); Ponomareva et al. (2022) is that ours focuses on pre-training vision foundation models, whereas Anil et al. (2021); Ponomareva et al. (2022) focused on pre-training language models. To the best of our knowledge, this is the first work to explore how to pre-train large-scale vision foundation models. Furthermore, previous works (Anil et al., 2021; Ponomareva et al., 2022) lack comprehensive studies on the learned representations of the DP pre-trained models, such as linear probing and few-shot learning. In contrast, we conduct extensive experiments on understanding the quality of the learned representations in this paper. Recent work by Ganesh et al. (2023) considered a simplified toy setting and demonstrated that public pre-training is necessary for training DP models in this specific context. However, this does not rule out the possibility of effective DP representation learning in large-scale real-world scenarios, as shown in this work.

3. Recipe for Training DP Foundation Vision Models

In this work, we identify a successful recipe for training differentially private foundation vision models. Training DP foundation models, or in general any deep learning model with a large number of parameters, poses a significant challenge due to the large amount of injected noise—$\mathcal{N}(0, \sigma^2 C^2 \boldsymbol{I})$ in equation 3. Indeed, current state-of-the-art differentially private deep learning models even underperform linear models with handcrafted features when ϵ is small (De et al., 2022; Tramer & Boneh, 2020). We propose two effective techniques that reduce the magnitude of noise injected during training while attaining strong (ϵ, δ)-DP guarantees: **1.** Scaling up the number of training samples via self-supervised learning with masked autoencoder; and **2.** Facilitating faster training by warm-starting the model with weights pre-trained on synthetic samples.

3.1. Differential Private SSL with Mask Autoencoder

Most existing works on differentially private training (De et al., 2022; Sander et al., 2022; Bu et al., 2022) focus on supervised learning, which inherently restricts the quantity of training samples that can be utilized. In contrast, self-supervised learning approaches unlock the use of (albeit uncurated) internet-scale training data that can be on the order of billions of samples, which can potentially satisfy the amount of data needed for DP training of high-utility models (Tramer & Boneh, 2020).

On the other hand, most existing SSL training approaches do not align with requirements in DP-SGD training. For example, SimCLR (Chen et al., 2020a) requires a mini-batch of samples in order to compute the contrastive loss; BYOL (Grill et al., 2020) computes per-sample loss but it utilizes batch normalization (BN) (Ioffe & Szegedy, 2015) in the model architecture, resulting in each loss depending on a mini-batch of training samples.[3] Therefore, it is challenging to perform the per-sample gradient clipping as described in equation 3. Among various types of SSL methods, we identify reconstruction-base learning with masked autoencoders (MAE) (He et al., 2022) as one of the most suitable SSL approaches for training DP foundation vision models. The training objective $L_{\text{MAE}}(\boldsymbol{\theta})$ is defined as:

$$L_{\text{MAE}}(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n} \underbrace{\ell_{\text{MSE}}(g \circ \psi(\text{mask}(\mathbf{x}_i); \boldsymbol{\theta}), \mathbf{x}_i)}_{\ell(\mathbf{x}_i; \boldsymbol{\theta})}, \quad (4)$$

where n is the number of training samples, $\mathbf{x}_i \in \mathbb{R}^{C\times H\times W}$ is the input of the i-th training image (C-number of channels, H-height, W-width), mask$(\cdot)$ is a function that mask out a fraction of the image, $\psi : \mathbb{R}^{C\times H\times W} \to \mathbb{R}^d$ is the encoder and $g : \mathbb{R}^d \to \mathbb{R}^{C\times H\times W}$ is the decoder. We use $\boldsymbol{\theta}$ to denote the trainable parameters of the ψ and g, and use ℓ_{MSE} to denote the mean squared error (MSE) loss defined on the pixel space, *i.e.*, $\ell_{\text{MSE}}(\mathbf{x}_1, \mathbf{x}_2) = \|\mathbf{x}_1 - \mathbf{x}_2\|_F^2$. Similar to He et al. (2022), we apply vision transformers (Dosovitskiy et al., 2020) to instantiate the encoder and decoder maps. As shown in equation 4, the training objective can be decomposed into n individual losses, and each individual loss $\ell(\mathbf{x}_i; \boldsymbol{\theta})$ only depends on the i-th training sample $\mathbf{x}_i$ and does not require the label of $\mathbf{x}_i$. Therefore, we can compute per-sample gradient $\nabla_{\boldsymbol{\theta}}\ell(\mathbf{x}_i; \boldsymbol{\theta})$ and perform per-sample gradient clipping without modifying the MAE training.

By leveraging the self-supervised MAE training paradigm, we can now significantly scale up the training data size for DP SSL pre-training. Dataset scaling can effectively reduce the magnitude of noise in DP-SGD while maintaining the same (ϵ, δ_n)-DP guarantee, where $\delta_n = 1/2n$. As shown in Figure 3a, we investigate the impact of injected noise in ViP training by keeping all training hyperparameters the same except for the number of training samples[4]. With more training samples, the magnitude of the injected noise σ becomes smaller. We find that when the noise magnitude is large, the training loss cannot be further optimized after certain

[3] Subsequent work by Richemond et al. (2020) demonstrated that BN can be substituted with group normalization by carefully modifying the model architecture. However, we have observed that the design of exponential moving averaged online network in BYOL can result in dynamic instability during training, which poses challenges in the context of DP training.

[4] We maintain the same batch size across various data size settings while modifying the noise multiplier σ. Consequently, as the data size increases, the corresponding σ values decrease.

number of training steps. In contrast, smaller magnitude of noise (as a result of larger training dataset) facilitates faster optimization of the training loss in comparison to larger noise scenarios. Importantly, the optimization trajectory is stable despite the presence of noise, allowing the MAE model to learn useful features.

3.2. Synthetic Pre-training Enables Faster DP Training for ViP

Non-private training of SSL models often require a significant number of training epochs, much larger than what is required in supervised learning (Chen et al., 2020a; He et al., 2022; Balestriero et al., 2023). This creates an additional challenge for DP training since the number of training iterations T directly impacts the privacy guarantee. Indeed, as mentioned in Section 2, DP-SGD with T iterations is $(\alpha, T\epsilon_\alpha)$-RDP. Consequently, naively applying DP-SGD to MAE training results in an unfavorable privacy-utility trade-off.

Fortunately, He et al. (2019) demonstrated that using pre-trained initialization enables much faster model convergence compared to random initialization. However, in light of our discussion in Section 1, it is critical that the pre-training data does not contain any private information, even if the data is deemed "public". One promising alternative is pre-training on programmatically-generated synthetic images (Kataoka et al., 2020; Baradad et al., 2022), which was shown to achieve competitive downstream performance compared to pre-training on natural images. Doing so allows the MAE to learn spatial structure in the transformer modules (Jelassi et al., 2022) without expending any privacy budget for the natural image data. More importantly, synthetic pre-training does not carry any privacy risk, and legal risk is limited to obtaining proper license for the synthetic image generation code.

Thus, to accelerate ViP training, we pre-train the model on synthetic images generated using the Shaders21k tool developed in Baradad et al. (2022). Figure 2 shows samples of synthetic images generated by the tool. In Figure 3b, we compare the ViP training with and without synthetic pre-trained initialization. Notably, training ViP with synthetic pre-trained weights converges significantly faster than those with random initialized weights. Increasing the synthetic pre-training from 20 to 900 epochs further improves convergence for ViP training. Interestingly, as shown in Table 1, MAE trained on the synthetic dataset already outperforms existing state-of-the-art DP-trained models (De et al., 2022; Sander et al., 2022) under our transfer learning evaluation, which shows that DP training on datasets even as large as ImageNet does not learn sufficiently expressive features (see Table 1). A concurrent work (Tang et al., 2023) also studied how to leverage the programmatically-generated synthesized textures (Baradad et al., 2022) for improving the privacy-utility trade-off of DP-SGD trained models. However, Tang et al. (2023) mainly studied the small-scale supervised learning setting, whereas our work focuses on differentially private self-supervised learning at scale and building foundation models with DP.

3.3. Our Proposed Approach

We now summarize our approach for DP foundation vision model training (also see Figure 2). It is worth mentioning that our proposed approach offers flexibility in the selection of both SSL training methods and synthetic datasets. For example, developing better synthetic datasets or more effective SSL learning method can further push the performance of the final DP foundation model.

DP-MAE with Synthetic Pre-training

- **Step 1:** *Synthetic pre-training for initialization.* Pre-train mask autoencoder on the synthetic dataset with non-private optimizers.
- **Step 2:** *DP training with synthetic initialization.* Apply the synthetic pre-trained model as initialization and train mask autoencoder on a large-scale natural image dataset (*e.g.*, LAION400M) with DP-SGD. The DP guarantee then applies to the natural image dataset.

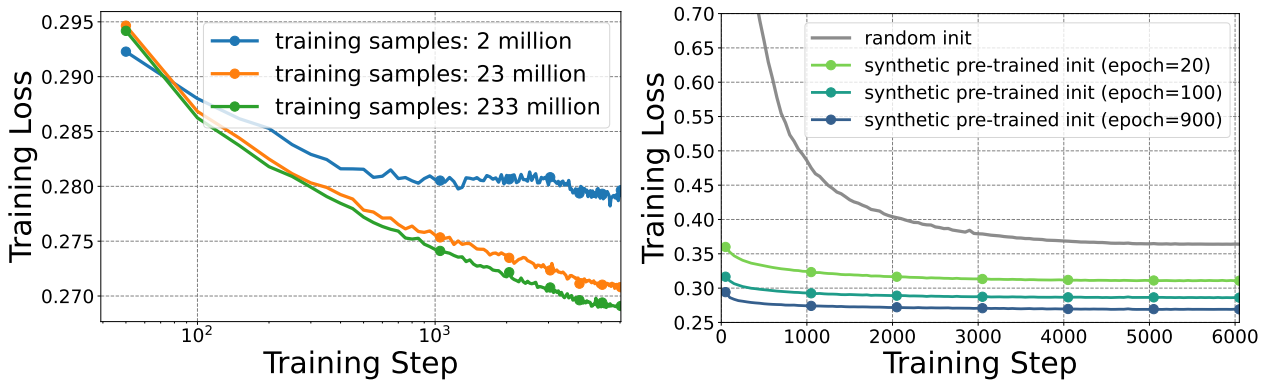
4. Evaluation

We evaluate the effectiveness of our training recipe by applying it to the LAION400M dataset to train our private foundation vision model: **ViP**. We consider various downstream tasks in order to demonstrate the quality and transferability of its learned representation. Furthermore, we compare ViP to previous state-of-the-art DP-trained models as well as widely adopted non-privately trained models, and find that ViP significantly improves SOTA for DP training on downstream transfer tasks (Section 4.2) and even outperforms non-private models on several challenging datasets. In addition to assessing the performance of ViP on non-private downstream tasks, in Section B.3, we also evaluate the ViP model via DP fine-tuning on ImageNet-1K, which shows a notable improvement of 10%+ absolute top-1 accuracy compared to previous SOTA (Sander et al., 2022). For additional experimental results on ViP, see Appendix B.

4.1. Evaluation Setup

Our implementation uses PyTorch, along with the functorch package (Horace He, 2021) for computation of per-sample gradients and the opacus package (Yousefpour et al., 2021) for privacy accounting. See Appendix A for additional implementation details.

Datasets. We use 1.05 million samples generated using the



(a) Role of dataset size.

(b) Effectiveness of synthetic pre-training.

Figure 3. (a). We vary the number of training samples n with the (ϵ, δ_n)-DP guarantee ($\delta_n = 1/2n$), and compare the training losses of MAE-DP. By scaling up the training dataset size, we can consistently improve the ViP training under the same ϵ-DP budget. (b). Compared to ViP training from random initialization, we can significantly speed up the ViP training by leveraging the synthetic pre-trained MAE model as initialization.

Table 1. Linear probing evaluation on downstream classification. We compare *ViP* with both private pre-training (DP-NFNet and TAN) and non-private pre-training (AlexNet and SimCLR) baselines, as well as the synthetically pre-trained MAE model: *(Syn)-ViP*. *ViP* consistently outperforms all private baselines, and has similar transfer learning performance as non-private SimCLR pre-trained on ImageNet-1K. (‡All models except for *(Syn)-ViP* and *ViP* are pre-trained on ImageNet-1K, giving them an unfair advantage for the linear probing evaluation on ImageNet-1K.)

Model	DP?	SSL?	Pre-train dataset	# pre-train samples	ImageNet-1K‡	Places-365	Places-205	iNat-2021
DP-NFNet	✓	✗	ImageNet-1k	∼1 million	45.3%	40.1%	39.2%	28.2%
TAN	✓	✗	ImageNet-1k	∼1 million	49.0%	40.5%	38.2%	31.7%
AlexNet	✗	✗	ImageNet-1k	∼1 million	56.5%	39.8%	35.1%	23.7%
SimCLR	✗	✓	ImageNet-1k	∼1 million	67.5%	46.8%	49.3%	34.8%
(Syn)-ViP	✓	✓	Shaders21k	∼1 million	49.8%	43.2%	45.8%	32.4%
ViP-LAION	✓	✓	LAION	∼233 million	**55.7%**	**46.1%**	**48.5%**	**38.1%**
ViP-ImageNet	✓	✓	ImageNet-1k	∼1 million	52.6%	44.3%	46.5%	34.2%

Shader21k (Baradad et al., 2022) tool as our synthetic pre-training dataset, and the LAION400M (Schuhmann et al., 2021) as our private pre-training dataset for the ViP model[5]. We evaluate ViP and baseline models via *non-private* linear probing and fine-tuning on the following downstream classification datasets: ImageNet-1K (Deng et al., 2009), Places-365 and Places-205 (Zhou et al., 2014), iNaturalist-2021 (Van Horn et al., 2021), CIFAR-100 (Krizhevsky et al., 2009), Caltech101 (Fei-Fei et al., 2006), and Aircraft (Maji et al., 2013). The input images are resized and center-cropped to 224×224 resolution. We also evaluate using MS-COCO instance segmentation and object detection (Lin et al., 2014), and semantic segmentation with the ADE20K dataset (Zhou et al., 2019) (in Appendix B.1).

Model architecture. Following He et al. (2022), we use vision transformer (ViT) (Dosovitskiy et al., 2020) to instantiate the masked autoencoder models. The default MAE-encoder has 12 transformer blocks and width 768, and the default MAE-decoder has 4 transformer blocks and width 512. We denote this MAE model as MAE-base. We also consider MAE models with different model sizes, including MAE-Nano, MAE-Tiny, MAE-Small and MAE-Large in Section 4.3. We present the ViP-Large (DP-trained MAE-Large) results in Figure 1. The complete numerical values can be found in Table 9. Apart from the data shown in Figure 1, the remainder of the experiments in this paper primarily focus on ViP-Base unless otherwise noted.

Optimization and hyperparameters for (DP-)MAE training. We use AdamW (Loshchilov & Hutter, 2017) for training MAE – both for synthetic pre-training and differentially private MAE pre-training. When evaluating pre-trained models in downstream tasks, we apply LARS (You et al., 2017) for linear probing and AdamW for fine-tuning. For MAE training, we set the masking ratio to 75%. In terms of DP training, we set $\epsilon = 8.0$ and $\delta = 1/2n$ by default

[5]The pre-training dataset we applied in this work is a curated and deduplicated version of LAION based on the curation techniques (Abbas et al., 2023; Xu et al., 2023a). We use LAION233M to denote this subsampled version of LAION400M.

Table 2. Fine-tuning evaluation on few-shot downstream classification. ViP consistently outperforms both TAN (private) and AlexNet (non-private), as well as (Syn)-ViP by a large margin. Performance does fall short compared to non-private SimCLR pre-trained on ImageNet-1K despite having access to more than 100× more data, suggesting that there is much room for improvement for private learning.

Model	Aircraft			Caltech-101			CIFAR-100		
	10-shot	20-shot	30-shot	5-shot	10-shot	30-shot	5-shot	10-shot	30-shot
AlexNet	23.27%	34.47%	41.35%	64.70%	73.57%	81.40%	29.74%	36.31%	49.28%
SimCLR	38.79%	56.90%	64.90%	81.70%	89.11%	94.51%	49.93%	60.18%	71.84%
TAN	22.84%	37.93%	46.01%	49.32%	66.42%	77.87%	21.28%	27.78%	42.35%
(Syn)-ViP	21.79%	46.85%	58.45%	60.51%	76.21%	88.48%	27.62%	38.96%	55.84%
ViP	**31.62%**	**53.05%**	**64.26%**	**68.05%**	**79.03%**	**88.90%**	**30.73%**	**40.95%**	**57.52%**

for training (ϵ, δ)-DP model. We set the clipping parameter $C = 0.1$, sampling ratio $q = 98304/n$, noise parameter $\sigma = 0.48$, and the total number of iterations $T = 6100$. More details about the training hyperparamters can be found in Appendix A.2. We apply the Opacus package (Yousefpour et al., 2021) for privacy accounting in all training runs. Figure 6 in Appendix A.2 visualizes the RDP curves across training iterations.

Existing methods for comparison. We compare with existing state-of-the-art DP-trained models: DP-NFNet (De et al., 2022) and TAN (Sander et al., 2022)), both of which are trained differentially privately on ImageNet-1K using supervised learning. In addition, we present the results of several widely used *non-private* models that are pre-trained on ImageNet-1K including AlexNet (Krizhevsky et al., 2017) (supervised learning-based) and SimCLR (Chen et al., 2020a) (SSL-based) for reference. To measure the effectiveness of DP pre-training compared to synthetic pre-training, we also evaluate the model pre-trained on synthetically generated Shader21k data, denoted **(Syn)-ViP**. We also compare ViP with the non-private MAE model pre-trained on the same datasets and summarize the results in Table 6 (Appendix B.4).

4.2. Transfer Learning Evaluation

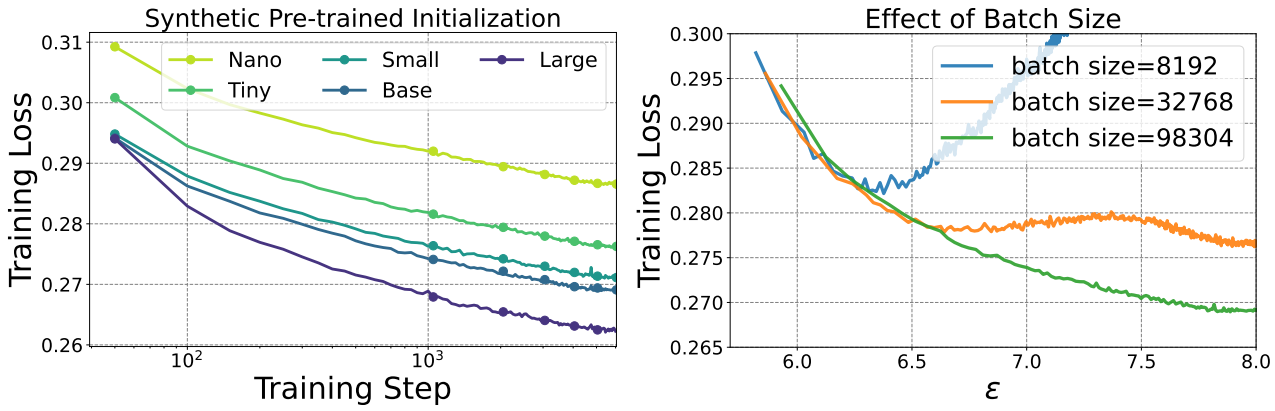
To show that ViP learns high-quality representations from its training data, we evaluate its transfer learning performance on a suite of image classification tasks using both linear probing and few-shot fine-tuning. For linear probing, we use all the training samples in the downstream task training set to learn the linear classifier, while freezing all layers except for the final linear layer. For few-shot fine-tuning, we randomly select K training samples from each class and fine-tune the entire model. It is worth noting that both linear probing and fine-tuning evaluations are done using *non-private* training; our pre-trained ViP model only satisfies (ϵ, δ)-DP on the LAION233M dataset.

Linear probing. Table 1 shows the linear probing results on four large-scale image classification datasets: ImageNet-1K, Places-365/205 and iNat-2021. The most suitable baselines in this setting are DP-NFNet and TAN, both of which are DP-trained on ImageNet-1K with $\epsilon = 8$ and represent previous state-of-the-art in large-scale DP pre-training. First of all, we find that MAE pre-training only on synthetic images (*i.e.*, (Syn)-ViP) is already comparable or even outperforms SOTA DP pre-trained models. After differentially privately pre-training on LAION233M, ViP effectively improves the performance of (Syn)-ViP on all datasets by a large margin.

Importantly, ViP even outperforms *non-private* SimCLR pre-trained on ImageNet-1K on all datasets (except ImageNet-1k itself because SimCLR does not need to transfer), and achieves similar performance as end-to-end non-privately trained AlexNet. To the best of our knowledge, this is the first time a DP-trained model can achieve similar performance on vision benchmark datasets as that of a mainstream (albeit older) model, which demonstrates the potential of our training recipe.

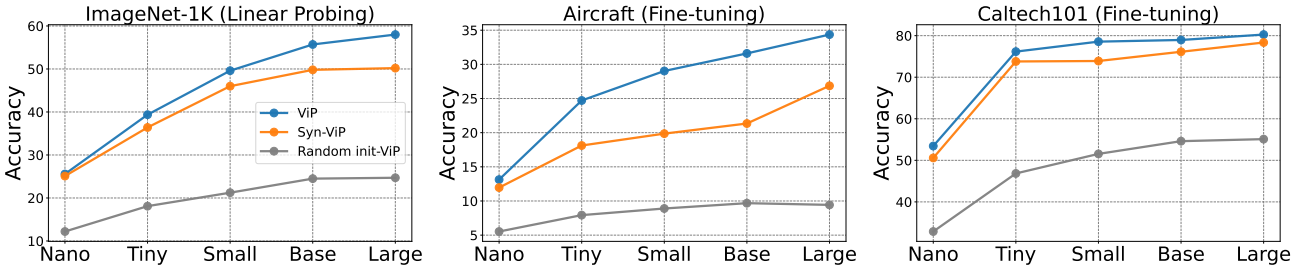
Few-shot fine-tuning. Table 2 shows the few-shot fine-tuning results on Aircraft, Caltech-101 and CIFAR-100. Similar to the linear probing result, (Syn)-ViP already outperforms TAN—the previous SOTA DP-trained model—across all evaluation settings except for 10-shot classification on Aircraft. Next, we find that ViP can largely improve upon (Syn)-ViP when the number of samples per class is small, attaining SOTA performance in all evaluation settings. ViP also achieves better performance than non-privately pre-trained AlexNet by a large margin, but falls short against non-private SimCLR despite having access to more than 100× training data. Thus, our result can be viewed as both a positive and a negative result, showing that there is still a long way to go for private learning before matching the performance of mainstream vision models across the board.

COCO object detection and segmentation. We fine-tune the pre-trained *(Syn)-ViP* and *ViP* on COCO with the `Detectron2` package (Wu et al., 2019). We apply the pre-trained *(Syn)-ViP-Base* and *ViP-Base* as the ViT initializations for the detection and segmentation tasks, and apply the default hyperparameter config in `Detectron2`



(a) Scaling up model. (b) Scaling up batch size.

Figure 4. (**Left**) Effect of scaling up model size on MAE training loss. Larger models attain lower training loss despite the larger magnitude of noise added during DP-SGD. (**Right**) Effect of batch size on MAE training loss while fixing ϵ. A large batch size is necessary for convergence.



(a) ImageNet-1K (linear probing) (b) Aircraft (fine-tuning) (c) Caltech (fine-tuning)

Figure 5. Effect of scaling up model size on downstream performance. ViP with synthetic pre-training (blue line) benefits substantially from larger model size. In comparison, ViP with random initialization (gray line) does not benefit as much from model scaling, as the difference in performance between MAE-Large and MAE-Nano is considerably smaller.

for ViTDet-Base. As shown in Figure 1 and Table 4, we find that the pre-trained ViPmodel outperforms several existing non-private models, such as SimCLR and Mask R-CNN, and performs slightly worse than the non-private MAE on these tasks.

4.3. Scaling Properties

We now study scaling properties of our training recipe, including scaling up (1) the model size, (2) the training set size, and (3) the previously known successful recipe of scaling up batch size.

Scaling up model size. DP-SGD training is generally unfavorable to large models because the noise magnitude increases with model size. Interestingly, we show that model performance in fact improves by scaling up model size using our training recipe. Specifically, we change the MAE-encoder size while fixing the MAE-decoder size, resulting in five different model sizes from MAE-Nano to MAE-Large; Table 3 in Appendix A.1) gives architecture details including number of parameters. All models are trained to satisfy the same (ϵ, δ)-DP guarantee with $\epsilon = 8$.

Figure 4a plots the training curve for the different-sized models. At the beginning of DP training, due to synthetic pre-training, a larger MAE model can learn more expressive features and hence the MAE training loss on LAION233M decreases as model size increases. Intriguingly, the training losses of MAE-Small/Base/Large are similar at the beginning, but larger ViT models achieve faster convergence *despite the large amount of DP noise*. Although similar observations on larger models converge faster have also been described in the context of non-private learning (Li et al., 2020), the fact that we observe the same phenomenon in Figure 4a suggests that model scaling can be effective even for *private* learning under our training recipe.

Figure 5 shows the effect of model scaling on downstream linear probing and fine-tuning performance. In particular, the effective reduction in training loss shown in Figure 4a indeed translates to better downstream performance, with larger ViP model consistently achieving better accuracy without modifications to the training process. Moreover, comparing ViP with synthetic pre-training (blue line) vs. random initialization (gray line) shows that *synthetic pre-training is crucial for unlocking this scaling behavior*: the difference in performance between MAE-Large and MAE-Nano is much smaller when the model is randomly initialized.

Scaling up dataset size. Next, we investigate the effect of scaling up the number of training samples in ViP training. We vary the training dataset size from 2M to 23M to 233M while choosing the magnitude of injected noise σ so that models trained on different dataset sizes satisfy (ϵ, δ_n)-DP guarantee with $\epsilon = 8$ and $\delta_n = 1/2n$, where n is the number of training samples. Table 10 shows downstream evaluation results. The first row corresponds to the synthetically pre-trained ViP model and rows 2-4 correspond to DP-trained ViP models with different dataset sizes. As expected, a larger pre-training dataset size results in a higher-utility ViP model. For example, scaling from 2M to 233M gives 3.1% linear probing accuracy gain on ImageNet-1K (from 52.6% to 55.7%). Given that the collection of large labeled datasets is very costly in practice, these results highlight the significance of self-supervised learning in DP training.

Scaling up batch size. Scaling up the training batch size is a known effective way to achieve strong performance in DP supervised learning (Li et al., 2022a). We analyze the effect of batch size in training ViP models and show that the same observation holds for DP self-supervised learning. We consider three different batch size $B \in \{8192, 32768, 98304\}$, and keep the computational budget—number of per-sample gradient computation—the same for all batch sizes. We then select the noise σ such that models trained with different batch size satisfy the same (ϵ, δ)-DP. As shown in Figure 4b, we find that larger batch size leads to better stability in the training process as well as faster convergence under the same computational budget. Rows 5-7 in Table 10 demonstrate that larger batch size also translates to a substantial improvement in ViP's transfer learning performance.

5. Discussion and Future Work

We developed a recipe for DP self-supervised learning of foundation vision models, and showed that the resulting model—ViP—can achieve downstream performance matching or exceeding that of mainstream non-private models such as SimCLR (with ImageNet-1K pre-training). Our work shows the potential of scaling DP training to internet-scale unlabeled datasets and presents several opportunities for future work.

1. Our recipe adapted MAE to DP-SGD training with minimal modifications. One limitation of this recipe is that it does not support other popular SSL methods such as SimCLR (Chen et al., 2020a), BYOL (Grill et al., 2020) and DINO (Caron et al., 2021; Oquab et al., 2023). It may be possible to design more specialized SSL training algorithms that conform to the requirements of DP-SGD and are more effective at learning useful representations.
2. Multi-modal SSL is generally more effective than single-modality pre-training due to the additional supervision from cross-modal alignment (Mu et al., 2022). However, existing multi-modal SSL methods are mostly based on contrastive learning (*e.g.*, CLIP (Radford et al., 2021), SLIP (Mu et al., 2022) and FLIP (Li et al., 2022b)) and do not admit per-sample gradient computation. Recent work (Huang et al., 2023) investigated how to fine-tune CLIP on vision-language tasks with DP guarantee. Additional work may be needed to adapt these methods to DP-SGD training.

Discussion on DP pre-training vs. DP fine-tuning. One popular paradigm for training large-scale DP deep learning models is: (1) non-private pre-training on *public data*; and (2) further fine-tuning on sensitive data with DP. This paradigm has achieved great success in a wide range of applications for specialized tasks (Yu et al., 2021; Li et al., 2022a). However, one major weakness of this "public pre-training then DP fine-tuning" paradigm is that it relies heavily on having high-utility pre-trained foundation models that are typically trained on potentially privacy-sensitive "*public data*" (Thomas et al., 2020; Yang et al., 2022). Recent work (Tramèr et al., 2022) questioned whether the use of large, web-scraped datasets should be considered privacy-preserving. Meanwhile, Thomas et al. (2020) demonstrates that one can successfully infer membership of pre-training examples when given access to fine-tuned models.

Given these limitations, it is valuable to consider the alternative paradigm of training foundation models that can privacy-preserving with respect to the pre-training data. Our work seeks to address this weakness by leveraging unlabeled private datasets to train foundation models that satisfy DP. On the other hand, one limitation of DP pre-training is the requirement for a large amount of data samples in the pre-training dataset. If the pre-training dataset is not sufficiently large, the noise added during the DP-SGD process would significantly degrade the performance of the DP pre-trained models. We believe that both paradigms—DP fine-tuning and DP pre-training—are valuable and complement each other for future research and development.

Discussion on sample-level DP vs. user-level DP. In this work, our privacy unit is at the sample level, which is the standard approach adopted in most works on differentially private machine learning. While sample-level DP is the most dominant notion, it is also important to consider user-level DP in scenarios where data can be clearly attributed to a particular user, for example, on-device data (Xu et al., 2023b) is a prime example of such a scenario.

Acknowledgements

This research was supported as a BAIR Open Research Common Project with Meta. We thank Xinlei Chen for helpful discussions on masked autoencoders and Xudong Wang for helpful discussions on MS-COCO detection and segmentation.

Impact Statement

Our research aims to develop scalable privacy-preserving methods for training large foundation models. The goal of this work is to explore the possibilities and limits of differentially private pre-training at scale. Through comprehensive analysis and experiments, we demonstrate the practical feasibility of applying differential privacy mechanisms to mitigate real-world privacy risks in AI.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Abbas, A., Tirumala, K., Simig, D., Ganguli, S., and Morcos, A. S. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*, 2023.

Anil, R., Ghazi, B., Gupta, V., Kumar, R., and Manurangsi, P. Large-scale differentially private bert. *arXiv preprint arXiv:2108.01624*, 2021.

Balestriero, R., Ibrahim, M., Sobal, V., Morcos, A., Shekhar, S., Goldstein, T., Bordes, F., Bardes, A., Mialon, G., Tian, Y., et al. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*, 2023.

Balle, B., Barthe, G., Gaboardi, M., Hsu, J., and Sato, T. Hypothesis testing interpretations and renyi differential privacy. In *International Conference on Artificial Intelligence and Statistics*, pp. 2496–2506. PMLR, 2020.

Bao, H., Dong, L., Piao, S., and Wei, F. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Baradad, M., Chen, R., Wulff, J., Wang, T., Feris, R., Torralba, A., and Isola, P. Procedural image programs for representation learning. *Advances in Neural Information Processing Systems*, 35:6450–6462, 2022.

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

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Bu, Z., Mao, J., and Xu, S. Scalable and efficient training of large convolutional neural networks with differential privacy. *arXiv preprint arXiv:2205.10683*, 2022.

Bu, Z., Wang, Y.-X., Zha, S., and Karypis, G. Differentially private optimization on large model at small cost. In *International Conference on Machine Learning*, pp. 3192–3218. PMLR, 2023.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T. B., Song, D., Erlingsson, U., et al. Extracting training data from large language models. In *USENIX Security Symposium*, volume 6, 2021.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020a.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. E. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020b.

Chen, X. and He, K. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, 2021.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

De, S., Berrada, L., Hayes, J., Smith, S. L., and Balle, B. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*, 2022.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In

Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3, pp. 265–284. Springer, 2006.

Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014.

Fei-Fei, L., Fergus, R., and Perona, P. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.

Ganesh, A., Haghifam, M., Nasr, M., Oh, S., Steinke, T., Thakkar, O., Thakurta, A. G., and Wang, L. Why is public pretraining necessary for private model training? In *International Conference on Machine Learning*, pp. 10611–10627. PMLR, 2023.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

He, K., Girshick, R., and Dollár, P. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4918–4927, 2019.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Henderson, P., Li, X., Jurafsky, D., Hashimoto, T., Lemley, M. A., and Liang, P. Foundation models and fair use. *arXiv preprint arXiv:2303.15715*, 2023.

Horace He, R. Z. functorch: Jax-like composable function transforms for pytorch. https://github.com/pytorch/functorch, 2021.

Huang, A., Liu, P., Nakada, R., Zhang, L., and Zhang, W. Safeguarding data in multimodal ai: A differentially private approach to clip training. *arXiv preprint arXiv:2306.08173*, 2023.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. pmlr, 2015.

Jelassi, S., Sander, M., and Li, Y. Vision transformers provably learn spatial structure. *Advances in Neural Information Processing Systems*, 35:37822–37836, 2022.

Kataoka, H., Okayasu, K., Matsumoto, A., Yamagata, E., Yamada, R., Inoue, N., Nakamura, A., and Satoh, Y. Pre-training without natural images. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Kurakin, A., Song, S., Chien, S., Geambasu, R., Terzis, A., and Thakurta, A. Toward training at imagenet scale with differential privacy. *arXiv preprint arXiv:2201.12328*, 2022.

Li, X., Tramer, F., Liang, P., and Hashimoto, T. Large language models can be strong differentially private learners. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=bVuP3ltATMz.

Li, Y., Fan, H., Hu, R., Feichtenhofer, C., and He, K. Scaling language-image pre-training via masking. *arXiv preprint arXiv:2212.00794*, 2022b.

Li, Z., Wallace, E., Shen, S., Lin, K., Keutzer, K., Klein, D., and Gonzalez, J. Train big, then compress: Rethinking model size for efficient training and inference of transformers. In *International Conference on machine learning*, pp. 5958–5968. PMLR, 2020.

Lin, G., Milan, A., Shen, C., and Reid, I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1925–1934, 2017.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Maji, S., Rahtu, E., Kannala, J., Blaschko, M., and Vedaldi, A. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Meehan, C., Bordes, F., Vincent, P., Chaudhuri, K., and Guo, C. Do ssl models have déjà vu? a case of unintended memorization in self-supervised learning. *arXiv e-prints*, pp. arXiv–2304, 2023.

Mironov, I. Renyi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pp. 263–275. IEEE, 2017.

Mironov, I., Talwar, K., and Zhang, L. R\'enyi differential privacy of the sampled gaussian mechanism. *arXiv preprint arXiv:1908.10530*, 2019.

Mu, N., Kirillov, A., Wagner, D., and Xie, S. Slip: Self-supervision meets language-image pre-training. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pp. 529–544. Springer, 2022.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. 2017.

Ponomareva, N., Bastings, J., and Vassilvitskii, S. Training text-to-text transformers with privacy guarantees. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2182–2193, 2022.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21 (140):1–67, 2020.

Rényi, A. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, pp. 547–562. University of California Press, 1961.

Richemond, P. H., Grill, J.-B., Altché, F., Tallec, C., Strub, F., Brock, A., Smith, S., De, S., Pascanu, R., Piot, B., et al. Byol works even without batch statistics. *arXiv preprint arXiv:2010.10241*, 2020.

Sander, T., Stock, P., and Sablayrolles, A. Tan without a burn: Scaling laws of dp-sgd. *arXiv preprint arXiv:2210.03403*, 2022.

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294, 2022.

Sharma, N. and Warsi, N. A. Fundamental bound on the reliability of quantum information transmission. *Physical review letters*, 110(8):080501, 2013.

Song, S., Chaudhuri, K., and Sarwate, A. D. Stochastic gradient descent with differentially private updates. In *2013 IEEE global conference on signal and information processing*, pp. 245–248. IEEE, 2013.

Tang, X., Panda, A., Sehwag, V., and Mittal, P. Differentially private image classification by learning priors from random processes. *Advances in Neural Information Processing Systems*, 36, 2023.

Thomas, A., Adelani, D. I., Davody, A., Mogadala, A., and Klakow, D. Investigating the impact of pre-trained word embeddings on memorization in neural networks. In *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings 23*, pp. 273–281. Springer, 2020.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Tramer, F. and Boneh, D. Differentially private learning needs better features (or much more data). *arXiv preprint arXiv:2011.11660*, 2020.

Tramèr, F., Kamath, G., and Carlini, N. Considerations for differentially private learning with large-scale public pretraining. *arXiv preprint arXiv:2212.06470*, 2022.

Van Horn, G., Cole, E., Beery, S., Wilber, K., Belongie, S., and Mac Aodha, O. Benchmarking representation

learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12884–12893, 2021.

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., and Girshick, R. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

Xiao, T., Liu, Y., Zhou, B., Jiang, Y., and Sun, J. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 418–434, 2018.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., and Hu, H. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9653–9663, 2022.

Xu, H., Xie, S., Huang, P.-Y., Yu, L., Howes, R., Ghosh, G., Zettlemoyer, L., and Feichtenhofer, C. Cit: Curation in training for effective vision-language data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15180–15189, 2023a.

Xu, Z., Collins, M., Wang, Y., Panait, L., Oh, S., Augenstein, S., Liu, T., Schroff, F., and McMahan, H. B. Learning to generate image embeddings with user-level differential privacy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7969–7980, 2023b.

Yang, K., Yau, J. H., Fei-Fei, L., Deng, J., and Russakovsky, O. A study of face obfuscation in imagenet. In *International Conference on Machine Learning*, pp. 25313–25330. PMLR, 2022.

You, Y., Gitman, I., and Ginsburg, B. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

Yousefpour, A., Shilov, I., Sablayrolles, A., Testuggine, D., Prasad, K., Malek, M., Nguyen, J., Ghosh, S., Bharadwaj, A., Zhao, J., et al. Opacus: User-friendly differential privacy library in pytorch. *arXiv preprint arXiv:2109.12298*, 2021.

Yu, D., Naik, S., Backurs, A., Gopi, S., Inan, H. A., Kamath, G., Kulkarni, J., Lee, Y. T., Manoel, A., Wutschitz, L., et al. Differentially private fine-tuning of language models. *arXiv preprint arXiv:2110.06500*, 2021.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pp. 12310–12320. PMLR, 2021.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A. Learning deep features for scene recognition using places database. *Advances in neural information processing systems*, 27, 2014.

Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., and Torralba, A. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019.

Appendix

A. Implementation and Evaluation Details

In this section, we provide implementation details for training and evaluating *(Syn)-ViP*, *ViP*, as well as other existing methods.

A.1. Details for MAE model

In Table 3, we provide details for backbones of MAE model with different model sizes. Both MAE-Large and MAE-Base encoders are constructed following the identical setup described in He et al. (2022).

Table 3. Details of MAE backbone variants used in *ViP*.

ViP model	MAE Backbone	Encoder depth	Encoder width	Decoder depth	Decoder width	# parameters
ViP-Nano	MAE-Nano	12	192	4	512	18.6M
ViP-Tiny	MAE-Tiny	12	384	4	512	34.8M
ViP-Small	MAE-Small	12	576	4	512	61.6M
ViP-Base	MAE-Base	12	768	4	512	99.0M
ViP-Large	MAE-Large	24	1024	4	512	233.3M

A.2. Details for ViP Pre-training

For *(Syn)-ViP* pre-training, we follow the training setup outlined in (He et al., 2022): we apply the training parameters specified in Table 8 of He et al. (2022) and pre-train pre-train *(Syn)-ViP* on the S21k dataset developed in Baradad et al. (2022), which comprises of 1,300,000 training samples, for a total of 1,000 epochs. Our *(Syn)-ViP* pre-training applies the self-supervised MAE training methodology and does not use the label information available in the S21k dataset.

We now present details for differentially private *ViP* pre-training. As mentioned in Section 3, we first initialize the model weights with *(Syn)-ViP* pre-trained on S21k dataset. Then we apply DP-AdamW[6]. See the table below for training hyperparameters.

Model	lr (η)	warmup iterations	wd (λ)	(β_1, β_2)	epsilon (ϵ^{adamw})	lr decay
ViP-Base	$3.84 \cdot 10^{-4}$	1,000	0.005	(0.9, 0.95)	10^{-8}	cosine

For masking in the MAE training, we follow the random masking strategy and masking ratio of 75% in He et al. (2022) for both *(Syn)-ViP* pre-training and *ViP* pre-training. The process of executing each iteration of DP-AdamW for training the *ViP-Base* model takes approximately 25 seconds when utilizing 48 A100 (40GB) GPUs. Each epoch of the *(Syn)-ViP-Base* model's training process takes roughly 90 seconds to complete with 48 A100 (40GB) GPUs.

Comparative analysis with non-private training in terms of computational costs and memory. To provide a more detailed comparison between differentially private training and non-private training, we consider the case with NVIDIA V100 32GB GPUs. Here we take the default MAE model (MAE-base) we studied in the submission as an example: (1). For non-private training, consider the batch size of 4096 (the batch size used in the MAE paper (He et al., 2022)), we distributed the batch across 32 GPUs with a per-GPU batch size of 128. Each iteration of non-private MAE training takes around 0.32 seconds. The default training iterations for MAE training is 124,800. (2). For differentially private training, we used a batch size of 81,920 distributed across 128 GPUs. Here, because of the memory overhead of using `functorch.vmap` for computing the per-sample gradients, we used a per-GPU batch size of 12 with gradient accumulation. Each iteration of private ViP training takes around 10 seconds. We mainly considered the number of DP iterations as 6,000, which is much less than the one used in MAE training. This is mainly because we applied the pre-trained Syn-ViP model that has been

[6]A variant of the standard DP-SGD — we first compute the noisy clipped stochastic gradient described in equation 3, then apply one step update of AdamW (Loshchilov & Hutter, 2017) using the estimated gradient.

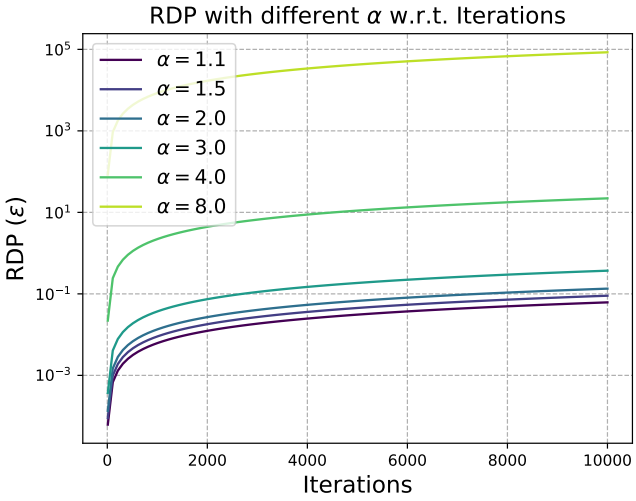


Figure 6. RDP curve across training iterations for the sampling rate $q = 98304/n$.

already trained on programmatically-generated synthesized textures (Baradad et al., 2022), which significantly speeds-up the training.

Overall, the ViP DP training is more costly than non-private MAE training, the main reasons are: (a). Larger batch size is needed to reduce the effective noise in DP training; (b). Per-sample gradient computation is less efficient than standard data-parallel computations in PyTorch. Although (b) can be alleviated by newer techniques for per-sample gradient computation such as Bu et al. (2023), we believe (a) will remain as the primary computational overhead for DP training going forward.

A.3. Details for Downstream Classification Task

Linear probing. We follow the training setup in He et al. (2022): we apply BatchNorm (Ioffe & Szegedy, 2015) before the last linear layer, and use the LARS (You et al., 2017) optimizer. We choose the base learning rate `blr` $\in \{0.1, 0.05, 0.01\}$, batch size $B = 16,384$, weight decay $\lambda = 0.0$. We set warmup epoch as 10, and total training epoch as 90. We use the `RandomResizedCrop` and `RandomHorizontalFlip` augmentations.

Few-shot fine-tuning. For vision transformer based architectures, we apply the AdamW optimizer with learning rate of lr $\in \{3 \cdot 10^{-3}, 3 \cdot 10^{-4}, 3 \cdot 10^{-5}\}$ and set weight decay as 0.05. For convolutional neural networks (AlexNet, ResNet used in SimCLR), we apply the SGD optimizer because it consistently outperforms AdamW. We select learning rate lr $\in \{1 \cdot 10^{-2}, 1 \cdot 10^{-3}, 1 \cdot 10^{-4}\}$, while setting the momentum as 0.9 and the weight decay as 0.0. For all models we apply the cosine learning rate decay, and use 10 warm-up epochs and fine-tine with 200 total epochs. We apply AutoAugment (Cubuk et al., 2018) for data augmentation.

A.4. Details for Downstream Segmentation and Detection Tasks

COCO object detection and segmentation. We fine-tune the pre-trained *(Syn)-ViP* and *ViP* on COCO with the `Detectron2` package (Wu et al., 2019). We apply the pre-trained *(Syn)-ViP-Base* and *ViP-Base* as the ViT initializations for the detection and segmentation tasks, and apply the default hyperparameter config in `Detectron2` for ViTDet-Base.

ADE20K semantic segmentation. We follow the setup described in He et al. (2022) on evaluating pre-trained MAE models for semantic segmentation. We apply the UPerNet (Xiao et al., 2018) and perform fine-tuning for 100 epochs with a batch size of 16.

A.5. Details for Differentially Private Fine-tuning on ImageNet

We use the pre-trained encoders of *(Syn)-ViP* and *ViP* and apply DP-AdamW for DP end-to-end fine-tuning. The details for parameters in DP-AdamW can found in the following table.

Model	sampling ratio q	noise σ	iterations T	lr	wd
ViP-Base / (Syn)-ViP-Base	$262,144/n$	5.6	1,500	$1.02 \cdot 10^{-3}$	0.005

We use 50 iterations for learning rate warm-up, and then keep the learning rate constant afterwards. For selecting parameters not presented in the aforementioned table, we adopt the default configuration of AdamW in `PyTorch` (Paszke et al., 2017). The fine-tuned model satisfies $(8, 8 \cdot 10^{-7})$-DP on the ImageNet-1K dataset in addition to the LAION233M dataset.

A.6. Details for Figure 1

For the linear probing results, we present the performance of the ViP-Large model, with the summarized results shown in the last row of Table 3. Regarding the detection and segmentation results, we utilize the ViP-Base model as the ViT backbone, and the corresponding outcomes can be found in Table 4.

B. Additional Experimental Results

In this section, we provide additional experimental results on evaluating *(Syn)-ViP*, *ViP*, as well as other existing methods.

B.1. Segmentation and Detection Evaluations of (Syn)-ViP/ViP

We summarize the results for object detection and segmentation in Table 4. Training details can be found in Appendix A.4.

Table 4. Evaluation of our DP models (*(Syn)-ViP*, *ViP*) as well as existing non-private baselines on COCO object detection/segmentation and ADE20K semantic segmentation.

Model	DP?	COCO		ADE20K
		AP^{box}	AP^{mask}	mIoU
SimCLR (Chen et al., 2020a)	✗	37.9	33.3	-
Mask R-CNN (He et al., 2017)	✗	40.0	37.1	-
RefineNet (Lin et al., 2017)	✗	-	-	40.7
MAE (He et al., 2022)	✗	50.3	44.9	48.1
(Syn)-ViP	✓	45.0	40.1	38.8
ViP	✓	45.2	40.4	40.1

B.2. Additional Experiments on ViP Pre-training

In Figure 7, we plot the training loss v.s. number of training steps for *ViP* training *without (Syn)-ViP initialization*. Compared to the results in Figure 4a, when pre-training from scracth with DP-AdamW, larger models do not converge faster than smaller ones. These results further demonstrate the effectiveness of synthetic pre-training for unlocking DP-SGD training of larger vision models.

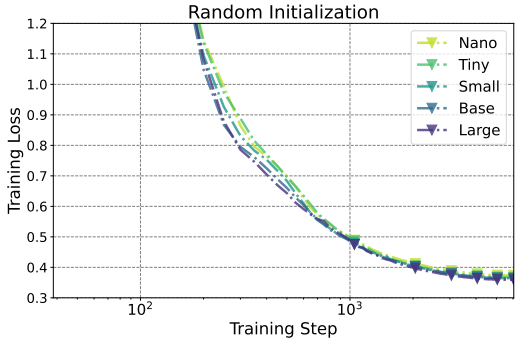


Figure 7. Training loss of different model sizes. (with random initialization).

B.3. DP Fine-tuning ViP on ImageNet-1K

Thus far, our main emphasis has been on evaluating DP pre-trained ViP through *non-private* linear probing or fine-tuning on downstream tasks. For certain use cases, the downstream task training set may be privacy-sensitive as well and DP fine-tuning is required. We simulate such a scenario by fine-tuning the privately pre-trained ViP model[7] on ImageNet-1K with DP-SGD. As a result, the fine-tuned model satisfies $(8, 8 \cdot 10^{-7})$-DP on the ImageNet-1K dataset in addition to the LAION233M dataset. We compare against prior works on training DP ImageNet models without pre-training (Kurakin et al., 2022; De et al., 2022; Sander et al., 2022); results are summarized in Table 5.

By utilizing our pre-trained *ViP* as an initialization, we observe an improvement in top-1 accuracy of more than 10% compared to the previous SOTA (Sander et al., 2022), demonstrating the efficacy of our DP pre-training recipe.

[7] ViP-Base pre-trained on LAION233 shown in the last row of Table 1.

Table 5. DP fine-tuning evaluation on ImageNet-1K. We compare (Syn)-ViP and ViP with existing DP training methods (DP-ResNet-18, DP-NFNet, and TAN) on ImageNet-1K. The (ϵ', δ') represents the privacy budget w.r.t. the ImageNet dataset.

Model	(ϵ', δ')-DP	Top-1 Accuracy
DP-ResNet-18 (Kurakin et al., 2022)	$(13.2, 10^{-6})$	6.2%
DP-NFNet (De et al., 2022)	$(8, 8 \cdot 10^{-7})$	32.4%
TAN (Sander et al., 2022)	$(8, 8 \cdot 10^{-7})$	39.2%
(Syn)-ViP	$(8, 8 \cdot 10^{-7})$	48.9% $\pm$ 0.2
ViP	$(8, 8 \cdot 10^{-7})$	**50.3%** $\pm$ 0.3

B.4. Additional Experiments on the Classification Task

Comparison with non-private MAE. To gain a better understanding of the gap between non-private training and private training, we use the same synthetic pre-trained model as initialization and perform DP-AdamW training on LAION233M with $\sigma = 0.0$[8]. We keep most of the training parameters the same except for setting the sampling ratio to $q = 4096/n$ and the number of iterations $T = 60,000$[9]. We then evaluate the linear probing (few-shot fine-tuning) performance of the trained model and provide the results in Table 6 (Table 7).

For linear probing, our *ViP* model closes more than half the gap between the *(Syn)-ViP* model and the non-private MAE model. With a more refined training recipe, it is plausible that the gap can be reduced even further, allowing DP-trained foundation vision models to rival non-privately trained ones on certain downstream tasks. In the context of few-shot fine-tuning, a comparison between private learning and the non-private MAE model reveals considerable potential for improvement in the private learning approach.

Comparison with ViP trained on de-duplicated LAION-2B. Recent work has demonstrated that there exist duplicated samples in the LAION dataset, which poses copyright and privacy challenges for foundation models trained on LAION. Therefore, we also pre-train our proposed ViP model on a de-duplicated subset of LAION-2B (Schuhmann et al., 2022), denoted by Dedup-LAION-245M, which consists of a similar number of training samples (245 million) as the one we mainly consider in this work. We summarize the linear probing performance of the ViP pre-trained on Dedup-LAION-245M in Table 6. We find the ViP model pre-trained on the de-duplicated LAION achieves similar performance as the one trained on LAION-400M (Schuhmann et al., 2021).

Table 6. Linear probing evaluation on downstream classification. We compare *ViP* and *(Syn)-ViP* with (non-private) MAE (He et al., 2022).

Model	Pre-train dataset	DP?	SSL?	ImageNet-1K[‡]	Places-365	Places-205	iNat-2021
(non-private) MAE	LAION-233M	✗	✓	60.5%	48.3%	51.8%	38.5%
(Syn)-ViP	LAION-233M	✓	✓	49.8%	43.2%	45.8%	32.4%
ViP	LAION-233M	✓	✓	55.7%	46.1%	48.5%	38.1%
ViP	Dedup-LAION-245M	✓	✓	55.5%	46.3%	48.1%	38.0%

Linear probing evaluation of ViP with different model sizes. We study the scaling behavior of ViP and (Syn)-ViP through linear probing. As shown in Table 9, we compare the performance of ViP and (Syn)-ViP with different model sizes. The performance of ViP consistently improves across all datasets as the model size increases. In contrast, increasing the model size from MAE-Base to MAE-Large results in less than 1% improvement in top-1 accuracy for (Syn)-ViP. These findings further underscore the effectiveness of our proposed ViP training recipe for scaling up model size in private pre-training.

[8] In this case, the $\epsilon = +\infty$ for the (ϵ, δ)-DP.

[9] While the trained model may not necessarily achieve optimal performance, our main purpose is to present a non-private model that follows a similar training setup, with the exception of setting the noise to zero. This allows us to compare its performance to the private model.

Table 7. Fine-tuning evaluation on few-shot downstream classification. We compare *ViP* and *(Syn)-ViP* with (non-private) MAE (He et al., 2022).

Model	Aircraft			Caltech-101			CIFAR-100		
	10-shot	20-shot	30-shot	5-shot	10-shot	30-shot	5-shot	10-shot	30-shot
(non-private) MAE	36.78%	56.82%	66.20%	72.93%	84.50%	92.78%	34.38%	47.98%	62.88%
(Syn)-ViP	21.79%	46.85%	58.45%	60.51%	76.21%	88.48%	27.62%	38.96%	55.84%
ViP	**31.62%**	**53.05%**	**64.26%**	**68.05%**	**79.03%**	**88.90%**	**30.73%**	**40.95%**	**57.52%**

Table 8. Linear probing evaluation of pre-trained *ViP-LAION* with different privacy budget. Here we evaluate the linear probing on the ImageNet-1k dataset. We vary the privacy budget epsilon (ϵ) from 2.0 to $+\infty$, where our default privacy budget is $\epsilon = 8.0$ and we use $\epsilon = +\infty$ to denote the non-private MAE model. The ϵ represents the privacy budget w.r.t. the LAION dataset. For $\epsilon \in \{2.0, 4.0\}$, we use the same training hyperparameters as the default ViP ($\epsilon = 8.0$) except for the noise multiplier σ and total iterations T. We let $(\sigma = 0.787, T = 1500)$ for $\epsilon = 2.0$ and $(\sigma = 0.603, T = 3000)$ for $\epsilon = 4.0$.

Model	Downstream dataset	$\epsilon = 2.0$	$\epsilon = 4.0$	$\epsilon = 8.0$	$\epsilon = +\infty$
ViP-LAION	ImageNet-1k	51.4%	53.8%	55.7%	60.5%

Table 9. Additional linear probing evaluation on downstream classification (ViP with different model sizes).

Model	# parameters	Backbone	ImageNet-1K	Places-365	Places-205	iNat-2021
(Syn)-ViP-S	61.6M	MAE-Small	46.0%	40.9%	43.2%	28.3%
(Syn)-ViP-B	99.0M	MAE-Base	49.8%	43.2%	45.8%	32.4%
(Syn)-ViP-L	233.3M	MAE-Large	50.2%	43.3%	46.5%	32.7%
ViP-S	61.6M	MAE-Small	49.6%	42.4%	44.7%	30.0%
ViP-B	99.0M	MAE-Base	55.7%	46.1%	48.5%	38.1%
ViP-L	233.3M	MAE-Large	**58.0%**	**48.5%**	**50.8%**	**40.6%**

B.5. ViP Ablation Experiments

We study the effect of dataset size and batch size in *ViP* pre-training, and evaluate different models with linear probing and fine-tuning on ImageNet-1K. We consider the *ViP-Base* setting and the results are summarized in Table 10.

Table 10. Ablation studies on the effect of dataset size and batch size. The first row shows the result of (Syn)-ViP, which is the common starting point for all models in the subsequent rows. Difference in performance compared to (Syn)-ViP is shown in parentheses. See text for details. (‡ represents linear probing evaluation and ◇ represents 10-shot fine-tuning evaluation.)

Model	Batch Size	# Train data	Noise σ	ImageNet-1K ‡	Places-365 ‡	iNat-2021‡	Aircraft◇	CIFAR-100◇
(Syn)-ViP	-	-	-	49.8%	43.2%	32.4%	21.8%	39.0%
ViP	98,304	2M	2.50	52.6% (+2.8%)	44.8% (+1.6%)	37.0% (+4.6%)	29.1% (+7.3%)	39.9% (+0.9%)
ViP	98,304	23M	0.66	53.7% (+3.9%)	45.2% (+2.0%)	37.6% (+5.2%)	31.5% (+9.7%)	40.5% (+1.5%)
ViP	98,304	233M	0.48	55.7% (+5.9%)	46.1% (+2.9%)	38.1% (+5.7%)	31.6% (+9.8%)	41.0% (+2.0%)
ViP	8,192	233M	0.41	43.9% (- 5.9%)	41.0% (- 2.2%)	27.6% (- 4.8%)	15.0% (- 6.8%)	39.2% (+0.2%)
ViP	32,768	233M	0.45	53.0% (+3.2%)	45.1% (+1.9%)	36.2% (+3.8%)	30.0% (+8.2%)	40.3% (+1.3%)
ViP	98,304	233M	0.48	55.7% (+5.9%)	46.1% (+2.9%)	38.1% (+5.7%)	31.6% (+9.8%)	41.0% (+2.0%)

We study the effect of MAE-decoder depth and MAE-masking ratio in *ViP* pre-training, and evaluate different models with linear probing on ImageNet-1K. We consider the *ViP-Base* setting and the results are summarized in Table 11.

Table 11. Ablation studies on the effect of decoder depth and masking ratio in MAE.

Model	decoder depth	masking ratio	ImageNet-1K
ViP (default)	4	0.75	55.7%
ViP	*1*	0.75	43.4%
ViP	*2*	0.75	51.7%
ViP	*8*	0.75	50.1%
ViP	4	*0.25*	53.5%
ViP	4	*0.5*	54.7%