Differentially Private Representation Learning via Image Captioning

Tom Sander[1,2,⋆], **Yaodong Yu**[1,3,⋆], **Maziar Sanjabi**[1], **Alain Durmus**[2], **Yi Ma**[3], **Kamalika Chaudhuri**[1], **Chuan Guo**[1]

[1]Meta, [2]École polytechnique, [3]UC Berkeley

[⋆]Equal contributions

Differentially private (DP) machine learning is considered the gold-standard solution for training a model from sensitive data while still preserving privacy. However, a major barrier to achieving this ideal is its sub-optimal privacy-accuracy trade-off, which is particularly visible in DP representation learning. Specifically, it has been shown that under modest privacy budgets, most models learn representations that are not significantly better than hand-crafted features. In this work, we show that effective DP representation learning can be done via image captioning and scaling up to internet-scale multimodal datasets. Through a series of engineering tricks, we successfully train a DP image captioner (DP-Cap) on a 233M subset of LAION-2B *from scratch* using a reasonable amount of computation, and obtaining unprecedented high-quality image features that can be used in a variety of downstream vision and vision-language tasks. For example, under a privacy budget of $\varepsilon = 8$, a linear classifier trained on top of learned DP-Cap features attains 65.8% accuracy on ImageNet-1K, considerably improving the previous SOTA of 56.5%. Our work challenges the prevailing sentiment that high-utility DP representation learning cannot be achieved by training from scratch.

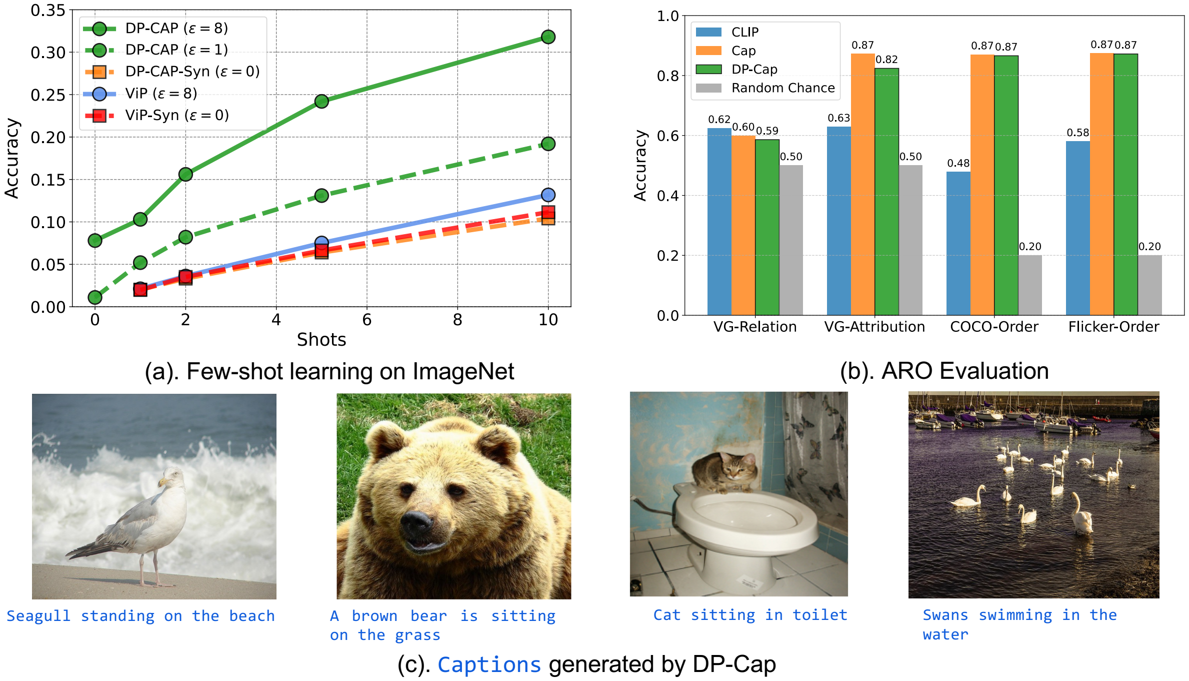
Date: March 6, 2024

Correspondence: tomsander@meta.com, yyu@eecs.berkeley.edu, chuanguo@meta.com

1 Introduction

Differentially private (DP; Dwork et al. (2006)) model training is an effective strategy for privacy-preserving ML on sensitive data. For most optimization-based learning algorithms, DP-SGD (Song et al., 2013; Abadi et al., 2016) can be readily applied to obtain models with rigorous DP guarantee. Regrettably, DP training has also been marred by a sub-optimal privacy-utility trade-off, with model utility severely lagging behind their non-private counterpart (Jayaraman and Evans, 2019; Tramer and Boneh, 2020; Kurakin et al., 2022). At the core of this unfavorable trade-off is the *difficulty of DP representation learning*. Tramer and Boneh (2020) showed that when DP training from scratch under a low-to-moderate privacy budget, most models learn representations with a quality worse than even handcrafted features. These observations naturally lead to the research question: *"How does one learn useful representations with DP training?"*

One plausible reason for the failure of prior attempts at DP representation learning is the lack of training data. Indeed, DP limits the information content of each training sample via the privacy budget ε, inducing a privacy-accuracy-sample size tradeoff; thus a substantially larger training dataset is required to extract the same amount of information to train the model. As the vast majority of prior work only utilize small to moderate scale classification datasets such as CIFAR-10 (Krizhevsky et al., 2009) and ImageNet (Deng et al., 2009b), the amount of training data is simply insufficient for learning high-quality representations under DP (Tramer and Boneh, 2020). Yu et al. (2023) made partial progress towards this through *self-supervised learning* (SSL) on *internet-scale data*. By training a masked autoencoder (MAE; He et al. (2022)) using DP-SGD on a 233M subset of the LAION-2B dataset (Schuhmann et al., 2022), the model learned image representations that are on-par with non-private AlexNet (Krizhevsky et al., 2012) trained on ImageNet—the first deep learning model to outperform handcrafted features and a major cornerstone for representation learning. However, the MAE objective promotes the model to learn extraneous details in the image that may not be helpful for obtaining generalizable representations, limiting the potential of this approach for DP.



(a). Few-shot learning on ImageNet

(b). ARO Evaluation

(c). Captions generated by DP-Cap

Figure 1 (a) Few-shot ImageNet-1K linear probe accuracy comparison between DP-Cap (ours) and ViP (Yu et al., 2023) (previous SOTA). DP-Cap learns better image representations using the same training data and privacy budget, and considerably surpasses synthetic initialization (*syn*). (b) Compositional understanding evaluation on the ARO benchmark (Yuksekgonul et al., 2022). DP-Cap performance is close to non-private Cap and outperforms non-private CLIP. (c) Captions generated by DP-Cap on images from the MS-COCO 2017 (Lin et al., 2015) test set.

We adopt a different approach of DP training via image captioning on internet-scale multimodal datasets. The reason is twofold: **1.** Text caption provides a concise summary of the training image and serves as better supervision compared to image-only SSL (Tschannen et al., 2023). Under the constraint on information content from DP, we hypothesize that it provides substantially more efficient information extraction under DP training. **2.** Image captioning is well-aligned with the prerequisites of DP-SGD such as having an instance-separable loss. We apply this method on a 233M subset of LAION-2B to train a DP image captioning model (DP-Cap), whose learned representations surpass previous SOTA—ViP (Yu et al., 2023)—by a large margin. As depicted in Figure 1(a), our model trained with a privacy budget of $\varepsilon = 8$ shows substantial improvements on downstream tasks compared to ViP, both trained on the same dataset. To achieve this, we also made crucial improvements to the efficiency of the DP training pipeline, **reducing the compute cost by close to 5**× on the largest model.

The image representations learned by DP-Cap also exhibit strong performance for multimodal tasks that require alignment of image and text features, the first occurrence for models trained from scratch with DP; see Figure 1(b). As a qualitative evaluation, we also use the trained DP-Cap model to caption several images from the MS-COCO 2017 (Lin et al., 2015) test set in Figure 1(c) and Appendix B.3. The resulting captions are grammatically correct and semantically coherent, while (close to) accurately describing contents of the image; this is interesting because our model has only been exposed to language supervision from LAION, which are far from being flawless. Our results suggest that DP training on internet-scale multimodal datasets can be a viable approach for obtaining high-utility learned representations.

2 Background and Related Work

Vision-language pre-training. Many modern ML datasets such as Conceptual Captions (Changpinyo et al., 2021), LAION (Schuhmann et al., 2021) and DataComp (Gadre et al., 2023) consist of aligned image-text pairs where the image and text contain roughly similar semantic information. One can leverage the aligned nature of the training data to pre-train *vision-language models* (VLMs) that connect the two modalities, whose representations perform more general multi-modal tasks. Contrastive learning-based techniques such as

CLIP (Radford et al., 2021) and BLIP (Li et al., 2022) are also applicable for pre-training VLMs. Doing so not only learns high-quality image and text representations but also introduces new multi-modal capabilities such as cross-modal retrieval and zero-shot prediction (Radford et al., 2021). Recent work by Tschannen et al. (2023) shows an image captioning approach (predicting text captions from images) is a viable alternative to contrastive learning and can lead to models with robust performance.

Differential privacy (Dwork et al., 2006). In the following, we denote by $\mathcal{M}$ a randomized learning algorithm, which takes a dataset $\mathcal{D}$ containing N samples and produces a machine learning model $\boldsymbol{\theta}$ through the process $\mathcal{M}(\mathcal{D})$. A randomized mechanism $\mathcal{M}$ is (ε, δ)-DP if, for any two adjacent datasets $\mathcal{D}$ and $\mathcal{D}'$ differing by a single sample, and for any subset $\mathcal{O} \subset \mathbf{Im}(\mathcal{M})$:

$$\mathbf{P}[\mathcal{M}(\mathcal{D}) \in \mathcal{O}] \leq \mathbf{P}[\mathcal{M}(\mathcal{D}') \in \mathcal{O}] \exp(\varepsilon) + \delta. \tag{1}$$

We adopt the leave-one-out notion of adjacency in this work, *i.e.*, $\mathcal{D} = \mathcal{D}' \cup \{\mathbf{x}\}$ for some sample $\mathbf{x}$ or vice versa. DP bounds the extent to which any potential adversary can infer information about the dataset $\mathcal{D}$ after observing the algorithm's output. In the context of ML, this implies that if we obtain the model $\boldsymbol{\theta}$ through a DP training algorithm $\mathcal{M}$ then its training data is provably difficult to recover or infer (Balle et al., 2022; Guo et al., 2022, 2023).

DP-SGD (Song et al., 2013; Abadi et al., 2016) is predominant differentially private algorithm for training deep neural networks (DNNs). At each gradient step k, a batch $\mathcal{B}_k$ is sampled where each example from the training data is chosen randomly with probability $q = B/N$, where B represents the average batch size. For $C > 0$, define the clipping function for any $X \in \mathbb{R}^d$ by $\text{clip}_C(X) = C \cdot X/\|X\|$ if $\|X\| \geq C$ and $\text{clip}_C(X) = X$ otherwise. Given model parameters $\boldsymbol{\theta}_k$, DP-SGD defines the update $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta_k \widetilde{\mathbf{g}}_k$ where η_k is the step size and $\widetilde{\mathbf{g}}_k$ is given by:

$$\widetilde{\mathbf{g}}_k := \frac{1}{B}\left[\sum_{i \in \mathcal{B}_k} \text{clip}_C\left(\nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}_k)\right) + \mathcal{N}\left(0, C^2\sigma^2\mathbf{I}\right)\right], \tag{2}$$

where $\ell_i(\boldsymbol{\theta})$ is the per-sample loss function evaluated at sample $\mathbf{x}_i$. We also use the term "DP-SGD" loosely to refer to the category of gradient-based optimization algorithms that operate on the noisy gradient, *e.g.*, Adam (Kingma and Ba, 2014). The privacy analysis of DP-SGD relies on composition of multiple steps. One particularly powerful analysis framework amenable to such compositions relies on a variant of DP called Rényi differential privacy (RDP) (Mironov, 2017). An advantage of RDP is its additive composition property, where the privacy guarantees of a sequence of mechanisms can be combined with amplification by subsampling (Wang et al., 2019) and then translated to (ε, δ)-DP (Balle et al., 2020; Gopi et al., 2021). In this work, we adopt this accounting technique.

Scaling up DP-SGD training. DP training is a theoretically and empirically proven remedy against unintended training data memorization. Even models with large ε (*e.g.*, $\varepsilon = 100$) can empirically defend against privacy attacks (Carlini et al., 2021; Guo et al., 2023). Despite its great appeal, DP training also carries a significant drawback of large drop in model utility (Abadi et al., 2016; Tramer and Boneh, 2020). For example, the SOTA performance on ImageNet when training from scratch with a DP guarantee of $\varepsilon = 8$ is 39.2% (Sander et al., 2023); in comparison, the non-private performance on ImageNet when training from scratch can reach 88% (Touvron et al., 2022) or higher. This degradation in model utility also translates to poorly learned representations, as Tramer and Boneh (2020) showed that even handcrafted features can rival ones learned through DP training.

Yu et al. (2023) made the first step towards obtaining high-utility learned representations through scaling DP training. They proposed self-supervised learning (SSL) on internet-scale data as a solution for the privacy-utility trade-off in DP representation learning. Among the numerous SSL algorithms, the authors observed that the reconstruction-based approach of masked autoencoder (MAE; He et al. (2022)) is compatible with the requirements of DP-SGD. By leveraging weight initialization through synthetic pre-training, the authors were able to obtain high-utility learned representations at a strict privacy budget of $\varepsilon = 8$. Compared to ViP (Yu et al., 2023), we demonstrate that the image captioning approach (see Section 3.1) learns much better image representations by utilizing the additional text supervision.

3 Approach

We describe in detail our approach of DP representation learning via image captioning. We first argue why image captioning is intuitively a suitable objective for obtaining better image representations via DP-SGD training (section 3.1). Then, we elucidate the technical challenges that we resolved to make DP training viable and effective for image captioning (section 3.2).

3.1 DP Representation Learning via Image Captioning

Why is vision-language pre-training suitable? Given image-text aligned datasets, prior works (Radford et al., 2021; Li et al., 2022; Tschannen et al., 2023) showed that pre-training using language supervision is an appealing option for non-private representation learning. We hypothesize that this is true for DP representation learning as well. Compared to image-only supervision, language supervision contains a more condensed summary of the image content, allowing the model to ignore irrelevant details such as background and focus on objects of interest and their relationships. This is especially helpful for DP since the model needs to extract as much useful information as possible from each sample given the privacy budget ε. Captioning could thus enhance the privacy-utility-sample size trade-off in DP, considering it requires less information per sample.

In addition, we show that vision-language pre-training supports a very large batch size, much larger than what is typically used in image-only pre-training (Radford et al., 2021; Li et al., 2022; Yu et al., 2022). This subtle aspect is in fact crucial for reducing the effective noise in DP-SGD (Li et al., 2021), which allows the model parameters to converge to a stable solution with lower training loss (see Section 3.2).

Vision-language pre-training via image captioning. Perhaps the most popular approach for vision-language pre-training is contrastive language image pre-training (CLIP; Radford et al. (2021)) as well as its variants (Mu et al., 2022; Li et al., 2023). However, the contrastive loss used in these methods is not an additive function over the samples, *i.e.*, it cannot be written in the form $\sum_i \ell_i$, where ℓ_i depends only on the i-th sample. Thus, DP-SGD (*cf.* equation 2) cannot be directly applied.

Unlike contrastive learning, the image captioning approach (Sariyildiz et al., 2020; Desai and Johnson, 2021; Tschannen et al., 2023) aligns well with DP-SGD training. Specifically, an image captioner is trained to predict captions based on their corresponding images. The training objective of the image captioner for one image-text pair $[\mathbf{x}^{\text{img}}, \mathbf{z}^{\text{text}}]$ is to minimize over $\boldsymbol{\theta} := \{\boldsymbol{\theta}_{\textsf{enc}}, \boldsymbol{\theta}_{\textsf{dec}}\}$ the following loss:

$$L_{\text{Cap}}([\mathbf{x}^{\text{img}}, \mathbf{z}^{\text{text}}]; \boldsymbol{\theta}) := \frac{1}{T} \sum_{t=0}^{T-1} \ell_{\text{CE}}\Big(z_{t+1}^{\text{text}}, \varphi\big(\underbrace{\psi(\mathbf{x}^{\text{img}}; \boldsymbol{\theta}_{\textsf{enc}})}_{\text{image embedding}}, \underbrace{z_1^{\text{text}}, \ldots, z_t^{\text{text}}}_{\text{first } t \text{ tokens}}; \boldsymbol{\theta}_{\textsf{dec}}\big)\Big), \tag{3}$$

where $\mathbf{z}^{\text{text}}$ denotes the caption token sequence $\{z_1^{\text{text}}, \ldots, z_T^{\text{text}}\}$ and the image captioner consists of two parts: the image encoder $\psi(\cdot; \boldsymbol{\theta}_{\textsf{enc}})$ and the text decoder $\varphi(\cdot; \boldsymbol{\theta}_{\textsf{dec}})$. The rationale behind the design of equation 3 is that the image encoder maps the input image $\mathbf{x}^{\text{img}}$ to an embedding vector, and the text decoder takes the image embedding $\psi(\mathbf{x}^{\text{img}}; \boldsymbol{\theta}_{\textsf{enc}})$ and the first t caption tokens $\{z_1^{\text{text}}, \ldots, z_t^{\text{text}}\}$ as inputs and predicts the next caption token z_{t+1}^{text}. Both the encoder and decoder are trained to maximize the log-likelihood of the correct next token. Equation 3 corresponds to the loss function for an image-text pair $[\mathbf{x}^{\text{img}}, \mathbf{z}^{\text{text}}]$; summing over all the samples in a batch gives the complete empirical loss in an additive form, which is directly compatible with DP-SGD. We discuss the sense of the privacy guarantees in Appendix A.1.

3.2 Strategy for Effective DP Training

Although image captioning has demonstrated impressive representation learning capabilities in the non-private regime, adapting it to DP training requires careful considerations. To obtain a useful pre-trained model, one needs to train for a sufficient number of steps under a low effective noise, both of which are at odds with obtaining a strong privacy guarantee. We detail the strategy we used to handle this trade-off when training the image captioner.

Sufficient number of training steps. We address this challenge via synthetic pre-training. We first observe that image representations learned by *DP-Cap (random init)* outperform those of *ViP (random init)* as evidenced

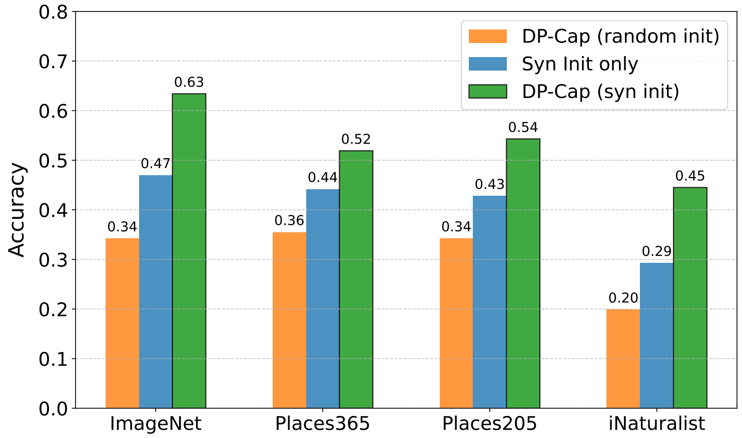
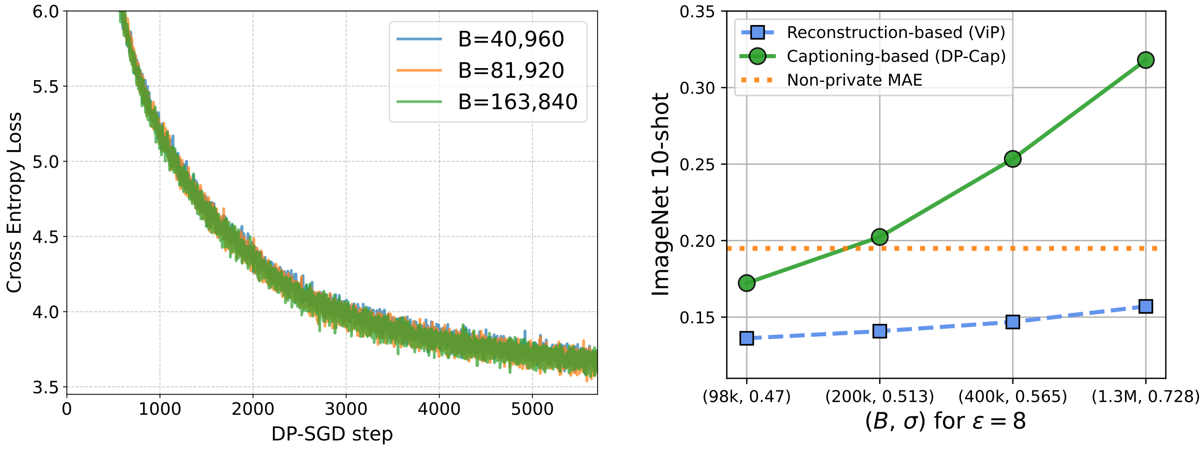


Figure 2 Impact of synthetic initialization on the DP-Cap model. The learned image representation benefits substantially from initializing on the Shaders21k dataset. The accuracy gap between DP-Cap (random init) and DP-Cap (syn init) can be as large as 24% for ImageNet linear probing.

in table 9 in App. B. Interestingly, Yu et al. (2023) have shown that synthetic images consisting of only textures can provide a better initialization for their reconstruction-based model without any privacy risk. With this initialization, the model can focus on learning dataset-specific properties rather than low-level image properties such as edge detectors, therefore expending privacy budget in a more optimal manner. We adapt this technique of pre-training on the Shaders21k dataset (Baradad et al., 2022) for initialization (see Appendix A.1 for details) and observe that it is even more effective with DP-Cap compared to ViP. As shown in Fig. 2, our *DP-Cap (syn init)* improves over *DP-Cap (random init)* by more than 24% on ImageNet-1k linear probing. It also improves over synthetic initialization alone (*Syn-init*) by more than 14% on ImageNet-1k linear probing, whereas the gain in ViP is smaller than 6%.

Using extreme batch sizes to reduce effective noise. Li et al. (2021) first showed that increasing the batch size in DP-SGD often improves the privacy-utility trade-off. This is because the effective noise added to the average gradient has magnitude σ/B (*cf.* equation 2), and that increasing B, rather than decreasing σ, results in better privacy guarantees according to conventional accounting techniques (Bun and Steinke, 2016; Dwork and Rothblum, 2016; Mironov et al., 2019; Sander et al., 2023). However, under supervised learning, increasing the batch size beyond a certain limit can lead to training instability under both private and non-private training. Specifically, Sander et al. (2023, 2024) observed that when training a classifier from scratch with DP-SGD on



(a) DP-Cap scales gracefully to extreme batch sizes

(b) Performance improves consistently with the batch size

Figure 3 (a) We fix the effective noise $\sigma/B = 5.6 \times 10^{-7}$ (corresponding to our (B, σ) = (1.3M, 0.728)) and show that the loss is remarkably consistent across different batch sizes, allowing us to effectively scale up batch size to improve the SNR. (b) Performance from 4 sets of parameters that provide $\varepsilon = 8$, with constant number of steps 5708. From batch size 98k (used in ViP (Yu et al., 2023)), to our 1.3M batch size. In contrast to ViP, DP-Cap successfully leverages the better SNR and learns features that achieve substantially better 10-shot accuracy on ImageNet even compared to a **non-private** MAE (He et al., 2022) trained on the same dataset (see Appendix A.1).

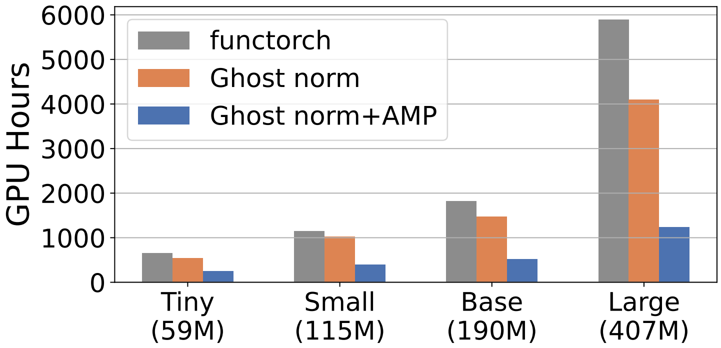


Figure 4 Number of GPU hours to train DP-Cap for one epoch on 233M samples. For the Large model, we achieve a close to 5× reduction.

Method	**Model**			
	Tiny	Small	Base	Large
functorch	36	20	12	5
Ghost norm	350	230	180	64
Ghost norm+AMP	420	310	225	90

Table 1 Max physical batch size under different per-sample gradient computation methods.

ImageNet, at a fixed number of steps S and fixed effective noise σ/B, the performance decreases significantly when the batch size becomes too large: for $B \in [128, 16384]$, a drop of 10% in top-1 accuracy was observed.

Intriguingly, we find that vision-language pre-training on internet-scale datasets can tolerate *extreme* batch sizes, *e.g.* $B = 1\text{M}$. In Figure 3(a), we compare the loss behaviors when scaling the batch size for DP-Cap. We fix the effective noise σ/B while varying the batch size. In stark contrast to the previous observation from Sander et al. (2023), the loss trajectory is identical across different batch sizes. With this observation, we are able to successfully scale up the batch size for DP-Cap to as large as $B = 1.3\text{M}$, achieving an effective noise of 5.6×10^{-7}, almost 10 times smaller than the effective noise of ViP in Yu et al. (2023). Training DP-Cap under such a small effective noise allows it to extract information from the training dataset more efficiently under the DP constraint. In Figure 3, we show that with $B = 1.3\text{M}$, the representation learned by DP-Cap even outperforms *non-private* MAE trained on the same dataset.

Improving training pipeline efficiency. DP training is known to be computationally inefficient due to factors such as per-sample gradient computation (Lee and Kifer, 2020; Li et al., 2021). Training on internet-scale datasets using extreme batch sizes further complicates this issue. For example, a naive implementation of DP-Cap with per-sample gradient computation using `functorch` would take approximately 61 days(!) on 128 NVIDIA V100 GPUs. We made significant efficiency improvements to the training pipeline using two techniques: ghost norm (Li et al., 2021) and the automatic mixed precision (AMP) package in PyTorch. Combining these two techniques with DP-SGD requires careful considerations to ensure both a correct DP guarantee as well as numerical stability. We detail the implementation of these two techniques in Appendix A.2.

In Figure 4 we compare the compute cost of different gradient computation methods: `functorch`, ghost norm and ghost norm+AMP. The number of GPU hours is estimated on a single NVIDIA V100 GPU with 32GB memory using 100K samples. The improvement is especially notable for our largest model: Using ghost norm+AMP, we achieve a 4.7× speedup compared to `functorch` and 3.3× speedup compared to ghost norm alone, which amounts to a reduction from 61 days to 13 days when training for 32 epochs—a large but manageable compute cost. This improvement is due to both a more efficient forward-backward pass, as well as enabling a larger physical batch size; see Table 1. In addition, we adopt the TAN simulation framework (Sander et al., 2023) to reduce the compute cost during hyperparameter search. Due to the batch size scaling behavior depicted in Figure 3, TAN simulation is ideal for DP-Cap training and allows for rapid experimentation to identify promising methods before launching a full training run.

4 Evaluation

We demonstrate the representation learning capabilities of DP-Cap on both vision (**V**) and vision-language (**V-L**) downstream tasks. For all evaluations, the DP-Cap model is first pre-trained using DP-SGD on a subset of LAION-2B (Schuhmann et al., 2022), and then fine-tuned non-privately on a downstream dataset.

Table 2 Linear probing evaluation on downstream classification. Results for DP-NFNet, TAN, AlexNet and SimCLR are obtained from Yu et al. (2023). For ViP (Yu et al., 2023), we train with the same privacy parameters as for DP-Cap on the deduplicated dataset. More details are given in Appendix A.1.

Model	pretraining data	DP?	ImageNet-1K	Places-365	Places-205	iNat-2021
DP-NFNet	ImageNet-1K	✓	45.3%	40.1%	39.2%	28.2%
TAN	ImageNet-1K	✓	49.0%	40.5%	38.2%	31.7%
AlexNet	ImageNet-1K	✗	56.5%	39.8%	35.1%	23.7%
SimCLR	ImageNet-1K	✗	67.5%	46.8%	49.3%	34.8%
Cap	Dedup-LAION-233M	✗	77.5%	56.3%	63.9%	63.9%
ViP	Dedup-LAION-233M	✓	56.5%	47.7%	49.6%	38.2%
DP-Cap	Dedup-LAION-233M	✓	63.4%	51.9%	54.3%	44.5%

Table 3 Performance of DP-Cap on zero-shot classification and compositional understanding (ARO). CLIP's zero-shot results are obtained from Radford et al. (2021) (base model). For ARO, see Appendix A.3.2.

Model	DP?	Zero-shot			ARO			
		ImageNet-1k	CIFAR10	CIFAR100	VGR	VGA	COCO	Flickr
Random Chance	-	0.1%	10%	1%	50%	50%	20%	20%
CLIP	✗	62.2%	91.3%	65.1%	62.4%	62.9%	47.8%	58.0%
Cap	✗	25.2%	90.0%	37.4%	59.9%	87.2%	87.0%	87.4%
DP-Cap	✓	7.8%	54.4%	16.4%	58.6%	82.4%	86.6%	87.2%

4.1 Downstream Tasks

Linear probing (V). We train a linear classifier on top of learned representations and evaluate its accuracy. We consider both full linear probing using the full downstream dataset, as well as few-shot linear probing, which subsamples the downstream dataset down to K samples per class. Few-shot linear probing is especially useful for evaluating learned representations since the model must rely heavily on the generalizability of representations in order to perform well under data scarcity.

Zero-shot image classification (V-L) is one of the most widely used methodologies for evaluating vision-language models (Radford et al., 2021). A strong zero-shot performance suggests that the image representation aligns well to text. We perform zero-shot classification using the DP-Cap image encoder and text decoder by evaluating the likelihood of captions of the form "this is a photo of a [label]". We enumerate over different labels and predict the class that has the highest likelihood; see Section A.3.1 for full details.

ARO (Attribution, Relation, and Order) (V-L). The ARO benchmark (Yuksekgonul et al., 2022) can be used to gauge the adeptness of VLMs in understanding the compositional relationship between objects and attributes. A strong performance on ARO suggests that the learned image representation encodes semantic relationships such as "the horse is eating the grass" vs. "the grass is eating the horse".

4.2 Experimental Setup

We present an overview of the experimental setup; refer to Appendix A for additional details.

Datasets. Following the approach introduced by Yu et al. (2023), we first pre-train on the Shader21k dataset (Baradad et al., 2022) of synthetic images. We then train with DP-SGD on a subset comprising 233 million deduplicated (using SemDeDup (Abbas et al., 2023)), NSFW-filtered and face-blurred (using an approach similar to Yang et al. (2021)) image-caption pairs from the (English-only) LAION-2B dataset (Schuhmann et al., 2022). We refer to this dataset as Dedup-LAION-233M.

We use the ImageNet-1K (Deng et al., 2009a; Russakovsky et al., 2014), CIFAR-10/100 (Krizhevsky et al.,

Table 4 Ablation studies on the effect of dataset size and privacy budget ε on DP-Cap (base).

ε	σ	# Data	# Steps	B	ImageNet-1K				ARO (**V-L**)			
					0-shot (**V-L**)	1-shot (**V**)	2-shot (**V**)	10-shot (**V**)	VGR	VGA	COCO	Flickr
$+\infty$	0	233M	60,000	40,960	25.2%	27.0%	37.2%	57.9%	59.9%	87.2%	87.0%	87.4%
8.0	0.728	233M	5708	1.3M	7.8%	10.3%	15.6%	31.8%	58.6%	82.4%	86.6%	87.2%
2.0	1.18	233M	2854	1.3M	3.2%	7.0%	10.8%	23.9%	58.5%	79.7%	85.3%	86.6%
1.0	1.5	233M	1427	1.3M	1.1%	5.2%	8.2%	19.9%	58.3%	75.6%	83.9%	85.3%
8.0	0.728	23M	5708	130K	0.7%	3.4%	5.3%	13.3%	58.3%	76.2%	84.9%	85.9%
8.0	0.728	2.3M	5708	13K	0.1%	1.8%	2.9%	8.1%	57.6%	66.4%	79.5%	82.0%

2009), Places-365/205 (Zhou et al., 2014) and iNaturalist-2021 (Van Horn et al., 2021) image classification datasets to assess the performance of learned image representations via full linear probing, few-shot linear probing, and zero-shot prediction. For vision-language tasks, we employ the Visual Genome Attribution (VGA), Visual Genome Relation (VGR), COCO-order (Lin et al., 2015) and Flickr-30k (Plummer et al., 2016) datasets from the ARO benchmark (Yuksekgonul et al., 2022). Finally, we evaluate image captioning using the MS-COCO 2017 (Lin et al., 2015) test set; result is shown in Figure 1(c) and Appendix B.3.

Model and training. We use a transformer architecture (Vaswani et al., 2017) for both the encoder and the decoder of DP-Cap, where the decoder applies causal cross-attention; see Section A.1 and Tschannen et al. (2023) for details. For privacy accounting we use Rényi DP composition along with privacy amplification via Poisson subsampling (Mironov et al., 2019), and convert to DP using Balle et al. (2020) through the PyTorch-based Opacus library (Yousefpour et al., 2021), targeting $\delta = 1/N$ where N represents the number of training samples. We refer to the non-private counterpart of DP-Cap trained on the same Dedup-LAION-233M dataset as "Cap".

4.3 Main Results

Linear probing evaluation (V). We assess the performance of the vision encoder on downstream tasks via linear probing. In Fig 1(a), we compare the performance of DP-Cap and ViP (Yu et al., 2023) on ImageNet-1k few-shot linear probing. DP-Cap significantly improves over ViP, with up to ×2.5 better performance across different shots. In addition, we evaluate the full linear probing accuracy of DP-Cap, ViP and other baselines in Table 2. DP-Cap outperforms ViP and other DP models, including TAN (Sander et al., 2023) and DP-NFNet (De et al., 2022), across all tasks. DP-Cap even outperforms non-private AlexNet (Krizhevsky et al., 2012) and except on ImageNet, SimCLR (Chen et al., 2020) (both were trained on ImageNet). We provide additional results for fine-tuning on downstream datasets in Table 10 (App. B), also showing improvements over competing methods.

Zero-shot performance (V-L). In the left three columns of Table 3, we evaluate the zero-shot performance of DP-Cap compared to non-private Cap and CLIP/BLIP on ImageNet-1k and CIFAR10/100. Contrastive methods such as CLIP and BLIP have demonstrated greater suitability for zero-shot prediction compared to image captioning approaches (Tschannen et al., 2023), which is evident by the disparity between the performance of Cap and CLIP/BLIP. Nevertheless, we observe that DP-Cap achieves noteworthy zero-shot classification performance that is significantly above random chance, and stands as the first DP model to do so. This accomplishment marks a promising milestone for DP training, although there remains a substantial performance gap between DP-Cap and Cap.

Attribution, Relation, and Order (ARO) evaluation (V-L). Contrastive-based methods such as CLIP often exhibit behavior akin to bag-of-words models (Yuksekgonul et al., 2022; Tejankar et al., 2021; Basu et al., 2023), making them less adept at performing well on the ARO benchmark. Remarkably, DP-Cap significantly outperforms *non-private* CLIP in this context (see Fig 1(b) and Table 3), and even achieves performance close to that of *non-private* Cap. Our result shows that DP training can be particularly effective for learning complex compositional relationships.

Table 5 Ablation studies on the effect of model size. We compare ViP and DP-Cap's number of encoder parameters. More details about the DP-Cap models can be found in Table 6.

Model	Config	# parameters	ImageNet-1K (Vision)					ARO (Vision-Language)			
			1-shot	2-shot	5-shot	10-shot	LP	VGR	VGA	COCO	Flickr
ViP	Base	86.6M	2.5%	4.2%	8.5%	14.3%	56.5%	/	/	/	/
DP-Cap	Tiny	22.0M	7.9%	12.1%	18.7%	25.2%	57.5%	58.6%	79.1%	85.7%	87.1%
DP-Cap	Small	49.0M	9.0%	14.0%	21.6%	28.9%	61.1%	59.1%	80.5%	86.0%	86.6%
DP-Cap	Base	86.6M	10.3%	15.6%	24.2%	31.8%	63.4%	58.6%	82.4%	86.6%	87.2%
DP-Cap	Large	407.3M	11.8%	17.5%	26.2%	34.0%	65.8%	59.5%	80.1%	86.6%	86.5%

4.4 Ablation Studies

We perform ablation studies on the scaling behavior of DP-Cap with respect to the dataset size, privacy budget and model size. In Appendix B, we show additional results on image captioning and on the impact of compute budget.

Scaling dataset size. We show that dataset scaling is crucial for effectively training DP-Cap as it results in better SNR under the same privacy budget (see Figure 5). We randomly subsample 1% and 10% of the Dedup-LAION-233M dataset, which is used for training our default DP-Cap-Base model in Table 2 (denoted by Dedup-LAION-2M and Dedup-LAION-23M). We set the batch size to $B/100$ for Dedup-LAION-2M and $B/10$ for Dedup-LAION-23M, respectively. This allows the model to be trained for the same number of steps across the different datasets, although at a much larger effective noise level. As shown in Table 4, the number of training samples is critical for achieving strong performance for DP-Cap models: the zero-shot performance of our model trained on 1% of the dataset achieves random zero-shot performance on ImageNet and much worse accuracy across the board on ARO.

Impact of the privacy budget ε. We also investigate the performance of DP-Cap under lower privacy budgets ($\varepsilon = 1$ and $\varepsilon = 2$), employing the same batch size of 1.3 million. The outcomes of these experiments are presented in Table 4. As anticipated, the utility of our model does exhibit a decline with decreasing ε. However, the performance degradation is relatively minor for the learned representation, with 10-shot ImageNet performance decreasing from 31.8% ($\varepsilon = 8$) to 19.9% ($\varepsilon = 1$). More surprisingly, the performance impact on ARO is nearly negligible. It is noteworthy that both models continue to outperform previous state-of-the-art DP models trained with $\varepsilon = 8$ (see Figure 1). This phenomenon can be attributed to the relatively small effective noise resulting from the extreme batch size, which for $\varepsilon = 1$ remains five times smaller than that used in Yu et al. (2023).

Scaling model size. Scaling up the model size is one of the most effective approaches for training better non-private foundation models (Brown et al., 2020; Bommasani et al., 2021; Touvron et al., 2023). However, conventional wisdom suggests that scaling up model size does not improve utility in DP training since more model parameters will lead to lower signal-to-noise ratio[1]. To test this hypothesis, we train DP-Cap with different model sizes (Tiny, Small, Base, Large) using the same hyperparameters and evaluate their performance in Table 5,11; see Table 6 for details about different model sizes. We observe consistent improvements when scaling up the model from DP-Cap-Tiny to DP-Cap-Large. Our observation suggests that DP-Cap has strong model scaling behavior even with DP-SGD training.

5 Discussion and Future Work

We demonstrated that DP representation learning via image captioning is viable. In particular, image captioning is an ideal objective that supports both per-sample loss and large batch training—two critical ingredients in DP-SGD. When applied to the Dedup-LAION-233M dataset, the trained model learns useful image representations for downstream tasks and exhibits strong multi-modal capabilities.

[1] This is because the added noise has L_2 norm $\approx \sigma C\sqrt{d}/B$, where d is the number of model parameters, whereas the gradient norm is constrained to C regardless of model size.

Through our study we also identify three open problems in the general direction of DP pre-training of large-scale foundation models that are difficult to handle with existing techniques: **1.** While we made notable efficiency improvements to support extreme batch sizes, it remains computationally demanding compared to non-private training. Is it possible to do DP training without the use of extreme batch sizes? **2.** Are there more parameter-efficient architectures that provide a better privacy-utility trade-off under data scaling? **3.** Contrastive learning remains state-of-the-art for learning features for downstream tasks such as retrieval. What techniques can enable effective DP contrastive learning?

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*, 2023.

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *European Conference on Computer Vision*, pages 456–473. Springer, 2022.

Borja Balle, Gilles Barthe, Marco Gaboardi, Justin Hsu, and Tetsuya Sato. Hypothesis testing interpretations and Rényi differential privacy. In *International Conference on Artificial Intelligence and Statistics*, pages 2496–2506. PMLR, 2020.

Borja Balle, Giovanni Cherubin, and Jamie Hayes. Reconstructing training data with informed adversaries. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1138–1156. IEEE, 2022.

Manel Baradad, Richard Chen, Jonas Wulff, Tongzhou Wang, Rogerio Feris, Antonio Torralba, and Phillip Isola. Procedural image programs for representation learning. *Advances in Neural Information Processing Systems*, 35: 6450–6462, 2022.

Samyadeep Basu, Maziar Sanjabi, Daniela Massiceti, Shell Xu Hu, and Soheil Feizi. Augmenting clip with improved visio-linguistic reasoning. *arXiv preprint arXiv:2307.09233*, 2023.

Leonard Berrada, Soham De, Judy Hanwen Shen, Jamie Hayes, Robert Stanforth, David Stutz, Pushmeet Kohli, Samuel L Smith, and Borja Balle. Unlocking accuracy and fairness in differentially private image classification. *arXiv preprint arXiv:2308.10888*, 2023.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. *arXiv preprint arXiv:2112.03570*, 2021.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020. https://arxiv.org/abs/2002.05709.

Soham De, Leonard Berrada, Jamie Hayes, Samuel L. Smith, and Borja Balle. Unlocking high-accuracy differentially private image classification through scale, 2022. https://arxiv.org/abs/2204.13650.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009a. doi: 10.1109/CVPR.2009.5206848.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009b.

Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11162–11173, 2021.

Cynthia Dwork and Guy N Rothblum. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, page 265–284, 2006.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023.

Sivakanth Gopi, Yin Tat Lee, and Lukas Wutschitz. Numerical composition of differential privacy. *Advances in Neural Information Processing Systems*, 34:11631–11642, 2021.

Chuan Guo, Brian Karrer, Kamalika Chaudhuri, and Laurens van der Maaten. Bounding training data reconstruction in private (deep) learning. In *International Conference on Machine Learning*, pages 8056–8071. PMLR, 2022.

Chuan Guo, Alexandre Sablayrolles, and Maziar Sanjabi. Analyzing privacy leakage in machine learning via multiple hypothesis testing: A lesson from fano. In *International Conference on Machine Learning*, pages 11998–12011. PMLR, 2023.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

Bargav Jayaraman and David Evans. Evaluating differentially private machine learning in practice. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 1895–1912, 2019.

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

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Alexey Kurakin, Shuang Song, Steve Chien, Roxana Geambasu, Andreas Terzis, and Abhradeep Thakurta. Toward training at imagenet scale with differential privacy. *arXiv preprint arXiv:2201.12328*, 2022.

Jaewoo Lee and Daniel Kifer. Scaling up differentially private deep learning with fast per-example gradient clipping. *arXiv preprint arXiv:2009.03106*, 2020.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners, 2021.

Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23390–23400, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

Jingcheng Liu and Kunal Talwar. Private selection from private candidates. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pages 298–309, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

Julien Mairal. Cyanure: An open-source toolbox for empirical risk minimization for python, c++, and soon more. *arXiv preprint arXiv:1912.08165*, 2019.

H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963*, 2017.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE, 2017.

Ilya Mironov, Kunal Talwar, and Li Zhang. Rényi differential privacy of the Sampled Gaussian Mechanism. *arXiv preprint arXiv:1908.10530*, 2019.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, pages 529–544. Springer, 2022.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, 2016.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Alfréd Rényi. 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, pages 547–562. University of California Press, 1961.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2014. https://arxiv.org/abs/1409.0575.

Tom Sander, Pierre Stock, and Alexandre Sablayrolles. Tan without a burn: Scaling laws of dp-sgd. In *International Conference on Machine Learning*, pages 29937–29949. PMLR, 2023.

Tom Sander, Maxime Sylvestre, and Alain Durmus. Implicit bias in noisy-sgd: With applications to differentially private training, 2024.

Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 153–170. Springer, 2020.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. *ArXiv*, abs/2210.08402, 2022. https://api.semanticscholar.org/CorpusID:252917726.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (S&P)*, pages 3–18, 2017.

Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 245–248, 2013.

Ajinkya Tejankar, Maziar Sanjabi, Bichen Wu, Saining Xie, Madian Khabsa, Hamed Pirsiavash, and Hamed Firooz. A fistful of words: Learning transferable visual models from bag-of-words supervision. *arXiv preprint arXiv:2112.13884*, 2021.

Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *European Conference on Computer Vision*, pages 516–533. Springer, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Florian Tramer and Dan Boneh. Differentially private learning needs better features (or much more data). 2020.

Michael Tschannen, Manoj Kumar, Andreas Steiner, Xiaohua Zhai, Neil Houlsby, and Lucas Beyer. Image captioners are scalable vision learners too. *arXiv preprint arXiv:2306.07915*, 2023.

Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12884–12893, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation, 2015.

Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled rényi differential privacy and analytical moments accountant. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1226–1235. PMLR, 2019.

Kaiyu Yang, Jacqueline Yau, Li Fei-Fei, Jia Deng, and Olga Russakovsky. A study of face obfuscation in imagenet. *CoRR*, abs/2103.06191, 2021. https://arxiv.org/abs/2103.06191.

Yang You, Igor Gitman, and Boris Ginsburg. Scaling SGD batch size to 32k for imagenet training. *CoRR*, abs/1708.03888, 2017. http://arxiv.org/abs/1708.03888.

Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. Opacus: User-friendly differential privacy library in PyTorch, 2021.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A. Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. Differentially private fine-tuning of language models, 2021.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Yaodong Yu, Maziar Sanjabi, Yi Ma, Kamalika Chaudhuri, and Chuan Guo. Vip: A differentially private foundation model for computer vision. *arXiv preprint arXiv:2306.08842*, 2023.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2022.

Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. *Advances in neural information processing systems*, 27, 2014.

Appendix

A Implementation Details

Table 6 Details of transformer backbone variants used in DP-Cap.

Model	Encoder depth	Encoder width	Decoder depth	Decoder width	# parameters (encoder & decoder)
DP-Cap-Tiny	12	384	6	384	59M
DP-Cap-Small	12	576	6	576	115M
DP-Cap-Base	12	768	6	768	190M
DP-Cap-Large	24	1024	6	768	407M

A.1 Training Details

DP accounting. We use RDP accounting with subsampling from the Opacus library (Yousefpour et al., 2021). Let D_α denote the Rényi divergence of order α (Rényi, 1961), and let

$$g_\alpha(\sigma, q) := D_\alpha((1-q)\mathcal{N}(0,\sigma^2) + q\,\mathcal{N}(1,\sigma^2) \,\|\, \mathcal{N}(0,\sigma^2)). \tag{4}$$

Then, from Mironov et al. (2019), performing S steps of DP-SGD satisfies (ε, δ)-DP with:

$$\varepsilon := \min_\alpha \left\{ S \cdot g_\alpha(\sigma, q) + \frac{\log(1/\delta)}{\alpha - 1} \right\}. \tag{5}$$

The quantity $g_\alpha(\sigma, q)$ can be upper bounded mathematically or derived numerically: we use the Opacus (Yousefpour et al., 2021) library for accounting in our work.

Regarding the DP guarantee, ε-DP bounds the amount of information extracted from each training sample by ε. Notably, for DP-Cap, each sample is made of {image + caption}, while ViP (Yu et al., 2023) utilizes {image} only. Consequently, DP-Cap inherently offers an equivalent or better privacy guarantee for each image. One way to see it is to note that DP provides protection against membership inference attacks (Shokri et al., 2017). Suppose ε-DP upper bounds the success rate of a membership inference attack (when given the image-text pair) against DP-Cap as $\leq p$. Then the MIA success rate when given only the image can be at most p since the attacker has strictly less information. This is exactly the upper bound for the success rate of a membership inference attack against ViP. In other words, any attacker that can attack the image+caption model (such as DP-Cap) can also attack the image-only model (such as ViP).

On the other hand, since the {image+caption} models utilize the caption, the privacy leakage from the text part of the image-caption pair is non-zero for $\varepsilon > 0$. It is worth noting that in our set up since we use DP, we protect the captions with the same ε-DP guarantee. Thus, the privacy protection for DP-Cap is neither strictly stronger nor strictly weaker than that for ViP, so the two privacy notions are not directly comparable.

Model details and task description. We utilize a transformer architecture (Vaswani et al., 2017) DP-Cap. This captioning model uses a text decoder that generates captions in an auto-regressive manner, utilizing a full attention mechanism on the vision encoder's output, as well as causal attention on the text. This architecture is closely aligned with the Cap architecture introduced in Tschannen et al. (2023). See Table 6 for details about the transformer architecture for different sizes. All results utilize the *base* model with the exception of the comparison in Table 6.

Hyperparameters. Our choice of gradient clipping factor is $C = 1$, as we did not observe any performance improvement with other values. We always use AdamW (Loshchilov and Hutter, 2018) for training. We use a learning rate of 5.12×10^{-4}. The learning rate is kept constant across batch sizes for TAN simulations and for the performance comparison in Figure 3 as the effective noise is kept constant in these cases (Sander et al., 2023). We use a maximum length of 40 tokens to process the LAION captions. We use a linear schedule, with 40% of warm-up iterations, and 2× the entire training as decay horizon. As opposed to what was previously observed (De et al., 2022; Sander et al., 2023), the learning rate schedule played an important role for us with DP-SGD training. We use a weight decay of 0.05. These choices come from hyperparameter search using TAN simulation with our base model. Following the standard practice (Berrada et al., 2023; De et al., 2022; Li et al., 2021; Yu et al., 2023; Sander et al., 2023), we do not count hyperparameter search within our privacy budget. Liu and Talwar (2019) have shown that hyperparameter search might not incur observable privacy loss.

Pre-training DP-Cap on the synthetic dataset. Compared to the encoder and decoder architecture design used in masked autoencoders (MAE) (He et al., 2022), the two main differences of the image captioning model used in this paper are: (1) The output of the encoder is fed into the decoder via cross-attention (Vaswani et al., 2017) in Cap; and (2) The self-attention used in the Cap decoder is causal self-attention. Similar to Yu et al. (2023), we apply the synthetic image dataset, Shaders21k (Baradad et al., 2022), to pre-train the DP-Cap model via MAE-based training. We follow most of the training setups used in ViP synthetic pre-training (Yu et al., 2023), except that we feed the output of the encoder to the decoder via cross-attention. The training loss of the synthetic pre-training in this paper is still the reconstruction loss used in MAE (He et al., 2022), and we did not leverage real-world text data for pre-training. After the model is pre-trained on Shaders21k, we change the self-attention to causal self-attention in the decoder, and replace the final layer (for pixel-wise reconstruction) of the decoder with the (randomly initialized) decoding layer for next word prediction. After making these modifications, we apply DP-SGD to pre-train our DP-Cap model with standard image captioning training objectives (see Section 3.1).

Pre-training ViP. To conduct a comparison with training on an identical datasets, we follow the methodology outlined in (Yu et al., 2023) to train with DP-SGD a MAE-based model, but with a change in the training data from LAION-233M to Dedup-LAION-223M, and use the same encoder's synthetic initialization as for DP-Cap. We further examine the linear probing performance on ImageNet and observe a 2% between the original model and the one trained on the deduplicated dataset. In addition, to corroborate the observation made in Figure 3, which suggests that the MAE-based method struggles to effectively harness massive batch sizes for achieving low effective noise in DP-SGD, we also train ViP models with larger batches, up to using the exact privacy parameters employed for DP-Cap (under $\varepsilon = 8$) with a notably large batch size of 1.3 million, and showcase the results in Table 7. For full linear probing, we observe only a small improvement over the original ViP model that was trained with batch size 98k. The success of DP-Cap is not solely attributed to its appropriate privacy parameters but is also a consequence of its remarkable ability to leverage the small effective noised induced by extremely large batch sizes.

A.2 Computation cost

Mixed Precision Package & Ghost Norm. DP-SGD introduces additional computational overhead compared to non-private training, primarily due to the computation of per-sample gradient norms. By employing the ghost norm technique of Li et al. (2021), we have successfully reduced the computational cost by up to one third with the Large Model (see Figure 4) compared to using `functorch`. The `torch.amp` package offers convenient methods for mixed precision, significantly speeding up operations like linear layers. However, it often leads to NaNs due to low precision handling of extreme values. While one can skip a step that led to NaNs in normal training, combining AMP with Ghost Norm is more complex. Ghost Norm requires two backward passes. In the first pass, per-sample gradient norms are computed. If one gradient norm is NaN, it contaminates the entire batch, leading to a NaN in the second backward pass. This issue is particularly prevalent in our setting with a batch size of 1.3M, as even a minuscule proportion of computations leading to NaNs can cause problems. To address this, we propose two solutions:

- **Loss Scaler**: We employ a similar trick to the standard use of AMP to reduce the number of NaNs. This involves dynamically upscaling and downscaling the loss with `torch.cuda.amp.GradScaler`. The same factor is used before the first and the second backward, and is updated based on the outputs of the second backward only.
- **Clipping to 0**: If any per-sample gradient norm computation results in a NaN value after the first backward, we set its clipping coefficient (the multiplicative coefficient in front of the corresponding per-sample loss for the second backward, as detailed in Li et al. (2021)) to 0 for the second backward. In this case, we do not update the loss scaling factor.

Table 7 Set-ups for our training for ViP, MAE (He et al., 2022) and DP-Cap: ImageNet-1k linear probing.

Model	pretraining data	(B, σ, S)	ImageNet-1K				
			1-shot	2-shot	5-shot	10-shot	full
ViP (Yu et al., 2023)	LAION-233M	(98k, 0.48, 6000)	2.5%	4.1%	8.5%	14.2%	55.7%
ViP	Dedup-LAION-233M	(98k, 0.474, 5708)	2.3%	3.9%	8.0%	13.6%	53.4%
ViP	Dedup-LAION-233M	(200k, 0.513, 5708)	2.5%	4.1%	8.3%	14.1%	54.0%
ViP	Dedup-LAION-233M	(400k, 0.564, 5708)	2.5%	4.2%	8.7%	14.7%	55.2%
ViP	Dedup-LAION-233M	(1.3M, 0.728, 5708)	2.7%	4.6%	9.4%	15.7%	56.5%
MAE (Non private)	Dedup-LAION-233M	(40960, 0, 40000)	3.4%	5.8%	11.8%	19.5%	62.5%
DP-Cap	Dedup-LAION-233M	(98k, 0.474, 5708)	4.2%	6.9%	11.4%	17.2%	50.6%
DP-Cap	Dedup-LAION-233M	(200k, 0.513, 5708)	5.6%	8.9%	14.5%	20.2%	54.2%
DP-Cap	Dedup-LAION-233M	(400k, 0.564, 5708)	7.6%	11.5%	18.5%	25.3%	59.1%
DP-Cap	Dedup-LAION-233M	(1.3M, 0.728, 5708)	10.3%	15.6%	24.2%	31.8%	63.4%

It's worth noting that the second solution is entirely valid for DP: instead of clipping the per-sample gradient to a norm C, it clips it to 0 in cases where computation results in a NaN value. This approach effectively mitigates the issue of NaN contamination in large batches. Overall, We have successfully reduced the computational cost by a factor 5 for the Large Model compared to `functorch`.

TAN simulation. Crucially, to achieve a favorable privacy-utility trade-off, DP-SGD necessitates training with massive batches over a substantial number of steps to achieve a good privacy-utility trade-off, as elaborated in Section 3.2. All our hyperparameter search were performed using the TAN simulation (Sander et al., 2023) for one epoch on our Dedup-LAION-233M. For our $\varepsilon = 8$ models, we limited training to 32 epochs, a process that took 5 days utilizing 128 V100 GPUs for the Base model.

While we have tried to reduced it as much as possible, training DP-Cap imposed a considerable energy consumption, resulting in elevated CO2 emissions. Our intention in releasing these models is to contribute to the mitigation of future carbon emissions, as the training has already been completed.

A.3 Evaluation Details

A.3.1 Details about Zero-shot Image Classification

While methods employing contrastive learning, such as CLIP, excel in this task, captioning methods exhibit comparatively lower performance, and with greater computational demands during evaluation. To evaluate a captioning model's zero-shot performance, we employ two distinct strategies:

- **Tree-based search**: We initiate caption generation with a prompt like "this is a photo of a " and greedily select the most likely next token among those that lead to valid completions within the true label set. The process continues until an End of Sentence (EOS) token is reached. For instance, if there are only two labels starting with "car": "car [EOS]" and "carpet [EOS]", and the initial predicted token is "car". Then the text decoder will predict the next token among "[EOS]" and "pet". If, among these two, "[EOS]" is chosen, and "car [EOS]" corresponds to the true label, then the zero-shot prediction is deemed correct.
- **Loss-based classification**: We assess, for each image, the probability of various captions that begin with "this is a photo of a [...]" where "[...]" is substituted with all feasible labels. Subsequently, we select the label that yields the most probable caption.

The "loss-based classification" comes with significantly higher computation costs as all the different captions have to be evaluated for each image (there representations is conditional to the image). For ImageNet, it implies 1000 forwards through the decoder for each image. We thus employ the tree-based search for presenting our findings in Table 3, although its greedy character with no backtracking is not optimal. Surprisingly, our preliminary experiments suggest the tree-based search gives comparable results.

Table 8 Compositional understanding (ARO): Results for CLIP (base) in Yuksekgonul et al. (2022) compared to our evaluation.

Model	ARO			
	VGR	VGA	COCO	Flickr
CLIP (eval from Yuksekgonul et al. (2022))	59%	63%	46%	60%
CLIP (our eval)	62.4%	62.9%	47.8%	58.0%

Table 9 Training from random initialization: Superiority of DP-Cap over ViP, both trained from random initialization.

Model	ImageNet-1K			
	1-shot	2-shot	10-shot	full
ViP ($\varepsilon = 8$)	0.1%	1.7%	6.1%	23.9%
DP-Cap ($\varepsilon = 8$)	5.6%	8.5%	18.8%	47.0%

A.3.2 Details about ARO Evaluation

We adhered to the protocol and code base established in Yuksekgonul et al. (2022) for re-evaluating CLIP's performance, and we observe slightly different results (see Table 8). For our captioning models, our approach involved computing the cross-entropy loss for all possible captions associated with each image and subsequently selecting the one with the lowest loss.

A.3.3 Details about Linear Probing and Fine-tuning Evaluation.

Few-shot linear probing is accomplished using the Cyanure library (Mairal, 2019). We use the same hyper parameters as in Assran et al. (2022). We adapted the MAE (He et al., 2022) code base for full linear probing, and we use the same hyperparameters as in Yu et al. (2023) (extract 12 layers of the image encoder, LARS optimizer (You et al., 2017) with base learning rate of 0.1, no weight decay and batch size of 16384).

B Additional Results

B.1 Additional experiments

Impact of the initialization (V). Our synthetic initialization for DP-Cap achieves less favorable results than the one from ViP reaches 50% (Yu et al., 2023); for instance, for full linear probing on ImageNet, they achieve 44% (Figure 2) and 50% respectively. However we have demonstrated that training with DP on top of synthetic initialization leads to significantly better results for DP-Cap compared to ViP for all the metrics; see Table 2, Table 10 and Figure 1. We observe that this superiority also appears when the models are trained from random initialization: as shown in Table 9, the improvement over ViP is even larger when training without synthetic initialization.

Fine-tuning (V). In Table 10, we present DP-Cap's performance in fine-tuning for few-shot evaluation. In contrast to the linear probing results shown in Table 2, the network is completely unfrozen. Therefore, we assess DP-Cap's capabilities primarily as a network initialization. Similarly to the linear probing results, we note a significant improvement in all metrics compared to previous DP vision backbones. Note that, similarly to linear probing comparison in Figure 1, we compare to non-private model performance which provides information about the performance gap between private models and non-private models. For fair comparison, we evaluate on the same same datasets than Yu et al. (2021).

Captioning task (V-L). We evaluate the image captioning performance of DP-Cap in comparison to non-private Cap. In Fig. 1(c), we present some (randomly chosen) captions generated by DP-Cap; more examples for DP-Cap and Cap can be found in Appendix B.3. Qualitatively, DP-Cap seems to generate reasonably good captions, similar to the ones generated by Cap. We also compare the two models quantitatively using the CIDEr metric (Vedantam et al., 2015) to evaluate the generated captions on the MS-COCO test set, and the results are summarized in the last column of Table 3. As DP-Cap and Cap are only trained on noisy captions from LAION, the CIDEr metric on MS-COCO is relatively low for both models. Moreover, despite the

Table 10 Fine-tuning evaluation on few-shot downstream classification.

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	23.3%	34.4%	41.4%	64.7%	73.6%	81.4%	29.7%	36.3%	49.3%
SimCLR	38.8%	56.9%	64.9%	81.7%	89.1%	94.5%	49.9%	60.2%	71.8%
TAN	22.8%	37.9%	46.0%	49.3%	66.4%	77.9%	21.3%	27.8%	42.4%
ViP	31.6%	53.1%	64.3%	68.1%	79.0%	88.9%	30.7%	41.0%	57.5%
DP-Cap	37.5%	57.9%	66.7%	70.3%	81.3%	90.0%	36.3%	46.3%	62.1%

Table 11 Ablation studies on the effect of model size for zero-shot prediction.

Model	Config	# parameters	DP?	Zero-shot		
				ImageNet-1k	CIFAR10	CIFAR100
Random Chance	-	-	-	0.1%	10%	1%
Cap	Base	86.6M	✗	25.2%	90.0%	37.4%
DP-Cap	Tiny	22.0M	✓	5.0%	46.5%	11.1%
DP-Cap	Small	49.0M	✓	6.9%	53.6%	17.1%
DP-Cap	Base	86.6M	✓	7.8%	54.4%	16.4%
DP-Cap	Large	407.3M	✓	9.2%	62.1%	24.0%

Table 12 Captioning evaluation on the MS-COCO test set of Cap and DP-Cap. For “fine-tuned", the model’s decoder is fine-tuned for one epoch on the MS-COCO train set (with the image encoder frozen).

Model	CIDEr score	
	original	fine-tuned
Cap	29.9	79.2
DP-Cap	15.7	51.3

similar performance between DP-Cap and Cap on ARO, the gap is much more significant for the captioning evaluation. Given these results, it is plausible that even though DP-Cap attains remarkably compositional understanding capabilities, its ability to generate text is still limited.

We also fine-tune Cap and DP-Cap’s decoders (while freezing the encoder) for one epoch on the MS-COCO train set, and assess the improvement in CIDEr scores in Table 12 to showcase the quality of the image representations and decoder initialization from the pre-training stage. The captions in Figure 1 and Appendix B.3 are generated using models that were *not* trained on MS-COCO.

What can we do with more compute budget? We restricted training the DP-Cap model for a compute budget of 32 epochs on the Dedup-LAION-233M dataset for each of our models with $\varepsilon = 8$. To fit the privacy budget while utilizing a batch size of 1.3 million and training for 32 epochs, RDP analysis yields $\sigma = 0.728$. However, we anticipate that further increasing the compute budget can yield even better models up to a certain limit: With the same ε and batch size, doubling the compute to 64 epochs only necessitates a 12% increase in σ. This increase enables twice as many steps to be performed with only a marginal increase in effective noise, potentially allowing the model to converge to a better solution.

In the absence of necessary compute for running this experiment, *we partially validate this hypothesis through the Total Amount of Noise (TAN) simulation*, training for the same number of gradient steps and with the same SNR per step, but using a $\times 32$ smaller batch size and $\times 32$ smaller σ to simulate at $\times 32$ lower compute. Our results in Table 13 indicate a significant performance improvement of 5% in 10-shot accuracy on ImageNet (compared to a similar simulation of the 32 epochs training). However, increasing the budget further to 128 epochs does not seem to enhance performance compared to the 64 epoch counterpart. Intuitively, the lower gradient SNR and larger number of gradient steps have opposite effects on optimization, and pushing past

Table 13 TAN simulation of the impact of the compute budget on the performance at fixed B.

	σ	
	0.81	0.95
Epochs	64 (×2)	128 (×4)
Effective noise σ/B	×1.12	×1.32
Predicted Final loss	-0.2	-0.2
Predicted 10-shot ImageNet	+5%	+5%

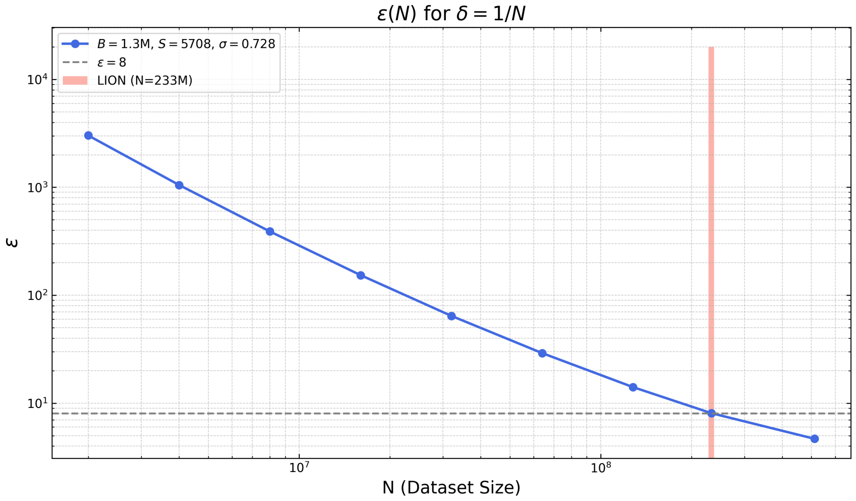


Figure 5 At fixed (B, σ, S), ε drastically reduces with the dataset size.

the "sweet spot" of training for 64 epochs at $\sigma = 0.81$ results in noisy steps that are unproductive for model convergence. To surpass the performance of the 64-epoch, 1.3-million batch size DP-Cap model, training with an even larger batch size appears necessary. We emphasize again that this result is derived through TAN simulation, and actual, compute-intensive training is required to fully validate this assertion.

B.2 More on the Impact of dataset size and privacy parameters

Dataset size. We emphasize here (again) the importance of having enough training data to achieve a good privacy-utility trade-off with DP-SGD. As depicted in Figure 5, increasing the number of training samples N while keeping the same number of equivalent DP-SGD steps (*i.e.*, keeping batch size B, noise σ, and number of update steps S constant) considerably reduces the privacy budget ε. Equivalently, having more data allows for an increase in the number of equivalent DP-SGD steps at fixed ε. Similar observations were also made by Tramer and Boneh (2020); McMahan et al. (2017). The abundance of pre-training data available for training foundation models thus proves highly compatible with DP requirements.

Batch size and σ. We wish to underscore the influence of batch size and σ on both the computational budget and model performance. As highlighted in Section 3.2, for a given target ε, elevating σ beyond 0.5 allows training for significantly more steps. In Figure 6, the blue, orange and green lines show the batch size (B) vs. compute trade-off (E) at a given σ. The lines are monotonically decreasing with B, signifying that the number of epochs E decreases when increasing B. When maintaining a fixed privacy budget $\varepsilon = 8$, even a marginal increase in σ from 0.48 to 0.728 (from blue to orange) translates to a remarkable increase ranging from 100 (for small batch sizes) to 100,000 (for very large batch sizes) times more gradient steps. Thus it is favorable to increase σ and B at the same time for better model convergence.

Meanwhile, doing so also incurs a higher computational cost: Under a 32-epoch budget on Dedup-LAION-233M with a batch size of 1.3 million, we had to cut the red curve in Figure 6, with $\sigma = 0.728$. As outlined in

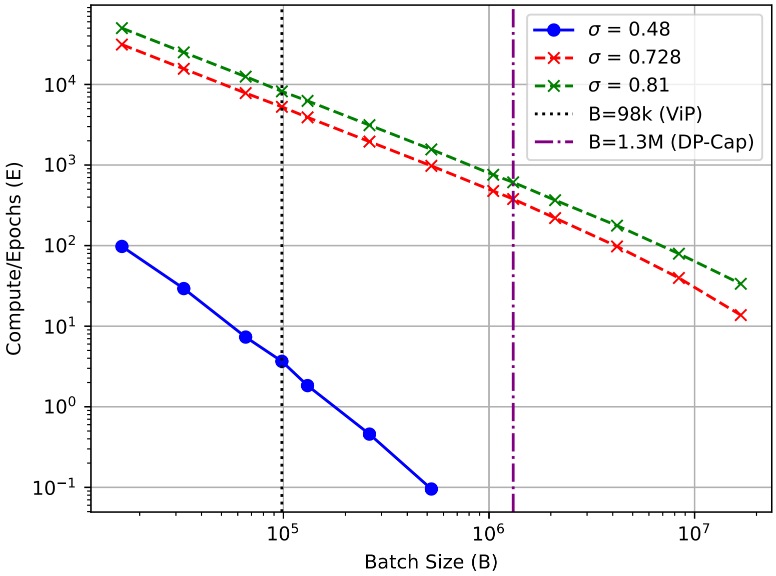
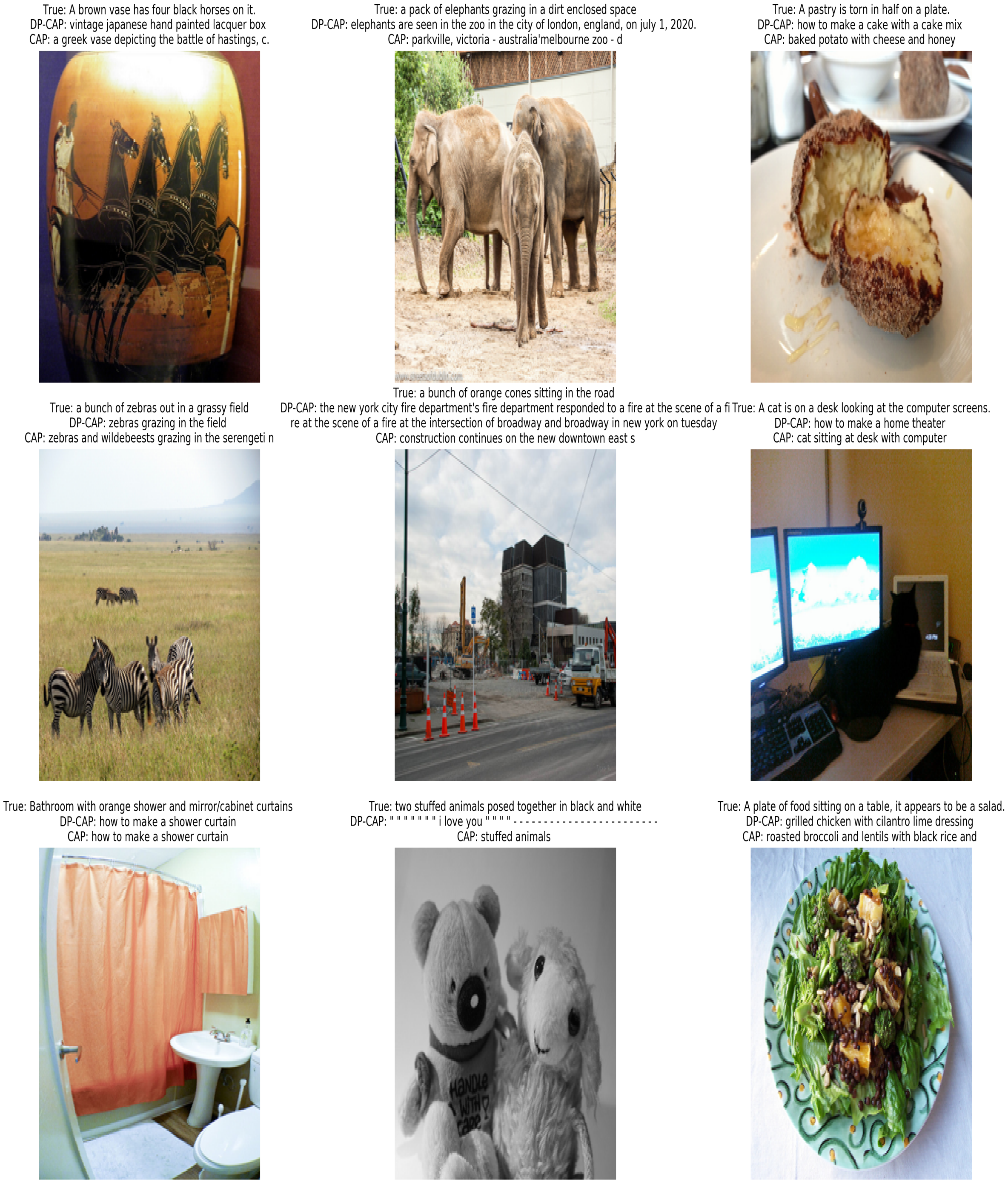


Figure 6 All points correspond to $\varepsilon = 8$ for a dataset of size $N = 233$M. At fixed ε and σ, the number of epochs decreases as the batch size increases.

Section 4.4, with twice this budget, we could have raised σ to 0.81 (green curve), with simulations indicating that this would have substantially improved performance. Additionally, Section 3.2 underscores that increasing the batch size is pivotal for achieving a high SNR while maintaining reasonable privacy guarantees. It is also crucial to note that at fixed ε, the compute budget is inversely proportional to the batch size. Therefore, increasing the batch size is beneficial for both SNR and computational efficiency. However, an excessively large batch size leads to fewer epochs and consequently a very limited number of training steps, which is detrimental to the training process (in addition to the difficulties of large batch training). For optimal privacy-utility-compute trade-off, a balance must be struck between computational resources, feasible batch size, and a reasonable number of training steps.

B.3 Image Caption Examples

In Figures 7 and 8, we show images from the MS-COCO 2017 test set and their corresponding captions generated by human annotator, Cap, and DP-Cap. Images in Figure 7 are selected randomly, whereas images in Figure 8 are randomly selected from the top 10% CIDEr score examples for DP-Cap. Qualitatively, the human-generated captions are more precise, whereas the captions generated by Cap and DP-Cap are more generic and sometimes contain factual errors. This is to be expected since Cap and DP-Cap are trained on LAION with much noisier text description and were *not* fine-tuned on MS-COCO. Nevertheless, DP-Cap still generates grammatically correct and (mostly) semantically coherent captions for unseen images.



True: A brown vase has four black horses on it.
DP-CAP: vintage japanese hand painted lacquer box
CAP: a greek vase depicting the battle of hastings, c.

True: a pack of elephants grazing in a dirt enclosed space
DP-CAP: elephants are seen in the zoo in the city of london, england, on july 1, 2020.
CAP: parkville, victoria - australia'melbourne zoo - d

True: A pastry is torn in half on a plate.
DP-CAP: how to make a cake with a cake mix
CAP: baked potato with cheese and honey

True: a bunch of zebras out in a grassy field
DP-CAP: zebras grazing in the field
CAP: zebras and wildebeests grazing in the serengeti n

True: a bunch of orange cones sitting in the road
DP-CAP: the new york city fire department's fire department responded to a fire at the scene of a fire at the scene of a fire at the intersection of broadway and broadway in new york on tuesday
CAP: construction continues on the new downtown east s

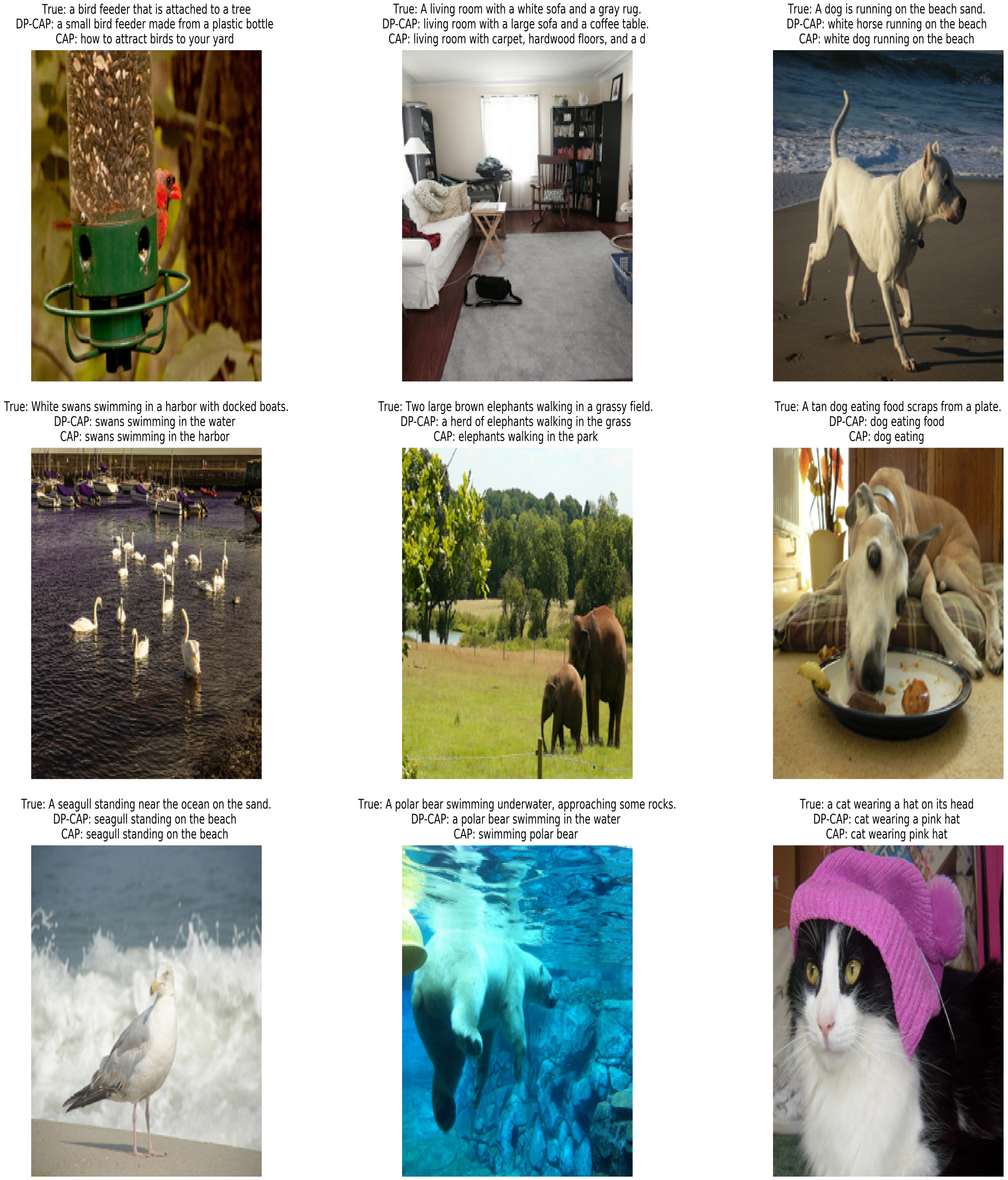
True: A cat is on a desk looking at the computer screens.
DP-CAP: how to make a home theater
CAP: cat sitting at desk with computer

True: Bathroom with orange shower and mirror/cabinet curtains
DP-CAP: how to make a shower curtain
CAP: how to make a shower curtain

True: two stuffed animals posed together in black and white
DP-CAP: " " " " " " " i love you " " " " -
CAP: stuffed animals

True: A plate of food sitting on a table, it appears to be a salad.
DP-CAP: grilled chicken with cilantro lime dressing
CAP: roasted broccoli and lentils with black rice and

Figure 7 Captions of randomly selected images from the MS-COCO 2017 test set.



True: a bird feeder that is attached to a tree
DP-CAP: a small bird feeder made from a plastic bottle
CAP: how to attract birds to your yard

True: A living room with a white sofa and a gray rug.
DP-CAP: living room with a large sofa and a coffee table.
CAP: living room with carpet, hardwood floors, and a d

True: A dog is running on the beach sand.
DP-CAP: white horse running on the beach
CAP: white dog running on the beach

True: White swans swimming in a harbor with docked boats.
DP-CAP: swans swimming in the water
CAP: swans swimming in the harbor

True: Two large brown elephants walking in a grassy field.
DP-CAP: a herd of elephants walking in the grass
CAP: elephants walking in the park

True: A tan dog eating food scraps from a plate.
DP-CAP: dog eating food
CAP: dog eating

True: A seagull standing near the ocean on the sand.
DP-CAP: seagull standing on the beach
CAP: seagull standing on the beach

True: A polar bear swimming underwater, approaching some rocks.
DP-CAP: a polar bear swimming in the water
CAP: swimming polar bear

True: a cat wearing a hat on its head
DP-CAP: cat wearing a pink hat
CAP: cat wearing pink hat

Figure 8 Captions of images (randomly picked among the top 10% CIDEr score of DP-Cap) from the MS-COCO 2017 test set.