Private Data protected Distillation model for image captioning via multi-teachers latent space feature aggregation

Mengze Li*, Zihan Huang*, Ziqi Tan, Zhengqing Fang, Zhou Zhao, Kun Kuang, Shengyu Zhang, Fei Wu

***Abstract*—Neural network based image captioning automatically generates natural language descriptions for images, but it is likely to cause private vision-language data leakage in real-world applications, such as medical images and diagnostic language reports in clinical trials. In this work, we study privacy-preserving image captioning with differential privacy, *i.e.*, protecting image captioning neural networks from data leakage. Generic differential privacy methods, which are mostly designed for regression/classification tasks under constrained environments, would inevitably incur excessive privacy budget for image captioning due to the complex nature of multi-model understanding and sequence generation task under natural unconstrained environments. To bridge the gap, we propose a Private Data protected Distillation framework (PDD) for image captioning. PDD detaches the privacy protection from the training process to alleviate excessive privacy budget incurred by the challenging multi-modal learning. Technically, PDD distills non-sensitive image captioning knowledge from multiple sensitive teachers with differential privacy on an auxiliary, small-scale, unlabeled, and non-sensitive dataset. To accommodate privacy protection for the sequence generation task, PDD encapsulates a Variational Privacy-preserving Pseudo-labeling Module to generate privacy-protected captions with multiple teachers for student training. Extensive experiments on image captioning benchmarks validate that PDD can achieve effective privacy protection effect with competitive image captioning performance.**

***Index Terms*—image captioning, differential privacy**

I. Introduction

IMAGE captioning is a long-standing computer vision task that aims to summarize images via natural language descriptions with a wide range of real-world applications [1], [2] In many applications, the image and language data used for training models can be privacy-sensitive, such as the medical images and the diagnosis reports of patients. Despite the remarkable progress made due to the rapid development of deep neural network, prior image captioning methods are arguably at risk of training data leakage according to findings in the generic domain [3], [4]. Therefore, it is necessary to study privacy protection for the image captioning task, preventing potential leakage of sensitive vision-language data while still maintain the usability of the model.

Recently, privacy protection for deep neural networks has drawn significant attention in the literature. Typically, differential privacy is an effective strategy, which has rigorous theory and excellent performance [5], [6]. Most of previous methods [5], [6] achieve differential privacy via intervening backpropagation process during model training. In particular, each training iteration requires privacy protection (*e.g.*, injecting gaussian noises on gradients) such that the model could not memorize the training data. Even so, a few privacy information continues to remain and consumes the privacy budget. Obviously, the budget size is tightly bounded by the number of training iterations. Previous methods mostly focus on image recognition task under constrained environments, such as hand-written digit recognition. Such tasks typically require limited number of training iterations where privacy protection could be achieved under a pre-defined privacy budget. However, it becomes non-trivial to apply existing differential privacy deep learning methods in the image captioning task due to the complex nature of multi-modal understanding and the sequence generation task under real-world natural environments.

In this work, we analyze the critical challenges and propose rational solutions for achieving differential privacy in image captioning. The first challenge comes from the multi-modal understanding under real-world natural scenarios, which inevitably requires significantly more training iteration for convergence. Existing methods that achieve differential privacy in model **training stage** would incur excessive privacy budget or inferior privacy protection under a limited budget, which might be less practical. To bridge the gap, rather than implementing protection in the training stage, we achieve differential privacy by introducing an auxiliary, small-scale, unlabeled, and non-sensitive data set to generate privacy-protected pseudo labels. To achieve this target, we resort to knowledge distillation techniques, where multiple teachers are trained on the original sensitive vision-language data samples, and generate pseudo captions for images in the auxiliary set with differential privacy. Since the images and captions themselves in the auxiliary set are privacy-protected, training the student model requires no additional privacy budget and thus addresses the first challenge. Another challenge relates to the pseudo caption generation with privacy protection. Note that we adopt multiple teachers trained on non-overlapping sensitive subsets such that any change in a particular sensitive vision-language data would at most affect one teacher. As a result, the pseudo captions aggregated from the output of multiple teachers will not rely on any single sensitive point, which greatly reduces the risk of privacy leakage. However, the aggregation of language sequences generated by multiple teacher models is non-trivial due to the discrete nature of

All authors are from the School of Computer Science, Zhejiang University. email: huanzh@zju.edu.cn

* equal contribution

language tokens and the sequential dependencies between tokens. To bridge the gap, we propose to perform multi-sequence aggregation, privacy preservation, and non-private feature correction in the continuous latent space.

Technically, we propose a Private Data protected Distillation framework (PDD) for image captioning via variational multi-teacher aggregation. The essence of the PDD framework is the **differential privacy distillation**, where we train multiple image captioning teachers on non-overlapping subsets of the original sensitive vision-language dataset, and one image captioning student on an auxiliary, small-scale, unlabeled, and non-sensitive dataset labeled by multiple teachers with differential privacy. To achieve effective labeling for the sequence generation task, we propose the **Variational Privacy-preserving Pseudo-labeling Module**, where we aggregate the sequences generated by multiple teachers in a variational latent space, protect the latent factors with gaussian noises according to differential privacy, and reconstruct the privacy-preserving pseudo captions as labels via Variational Auto-Encoder (VAE). In addition, a feature augmenter module is designed to fix disturbed non-sensitive information, and prevent larger variance of semantic deviation in the latent space aggregation and privacy protection.

We evaluate the PDD framework on two widely used image captioning benchmark datasets, *i.e.*, MS-COCO [7] and Flicker30k [8], which contain lots of sensitive human portraits and personal behavior habits. The performance is measured under pre-defined privacy budgets and four widely used standard metrics: BLEU@N [9], Meteor [10], Rouge-L [11] and CIDEr [12]. Extensive experiments show that our model can achieve effective privacy protection and competitive performance for this task, and an adjustable balance between privacy and accuracy.

The main contributions of this paper are summarized as:

- To the best of our knowledge, we take the initiative to investigate private data protection for the image captioning task. We analyze the task-specific challenges, and propose a novel Private Data protected Distillation framework to detach privacy protection from model training.
- A Variational Privacy-preserving Pseudo-labeling Module is proposed to aggregate the captions generated by multiple teachers in a variational latent space under differential privacy guarantee. In addition, a feature augmenter module corrects the non-private information and prevents potential semantic disturbance incurred by the latent space aggregation and privacy-protection, thus achieving an adjustable balance between privacy and performance.
- We conduct experiments on the image captioning task using the MS-COCO dataset and the Flicker30k. The extensive experiments show the effectiveness of the model.

II. Related work

A. Image captioning

The image captioning task has received extensive attention from researchers and has made a series of progress in recent years. In the early literature, most of papers use relatively primitive feature extraction methods combined with simple feature classifiers, such as [13], [14]. These methods are early attempts by researchers on this task. Although the accuracy is not high, many enlightening research ideas have been designed. With the development of deep learning, the framework of CNN combined with RNN is introduced into the model design of the image captioning task. [15] adopts the CNN to extract the features of the images and the LSTM to generate the captions by analyzing the features. [16] proposes three types of LSTM under the framework of CNN combined with RNN. The methods of this stage lack a mechanism corresponding to words and image regions, resulting in limited accuracy. [17], as a landmark paper, combines top-down and bottom-up attention into the design of the image captioning model, which greatly improves the accuracy. A large number of subsequent papers, such as [18], [19], continue to develop and improve accuracy on this basis, and a series of progress has been made. At this stage, the focus of research is shifting from improving accuracy to solving concrete problems, which are found during the practical process. Considering that different applications in complex scenes have different concerns and need to generate different captions, [20] designs a control image captioning framework. [21] design a model based on adversarial learning to generate multi-type captions, which can satisfy the language style requirements of the scene. However, the research on the protection of privacy for the image captioning task is still at a blank stage. In this paper, we propose a new image captioning model to solve it.

B. Privacy protection

As the problem of privacy leakage becomes more and more serious, more and more researchers invest in the research of privacy protection technology. The K-anonymity algorithm [22] requires that every sample of the dataset is indistinguishable from at least K-1 samples before publishing the dataset. Then, the l-diversity [23] and other algorithms are designed based on it to resist attack methods such as background knowledge attacks and homogenization attacks. However, such methods still rely heavily on the setting of background knowledge mastered by the attacker. The differential privacy [24], a privacy protection algorithm, can effectively solve this problem. How to combine it with powerful deep learning is a hot topic nowadays. [5] is the first paper to research this problem, which designs a general method that combines deep learning with differential privacy. The later papers mostly design algorithms for specific problems, such as [25], [26] focus on the classification task, and [6], [27] focus on the GAN task. But there is a lack of research on the image captioning task at this stage. And there is a large gap between previous research and the needs of the image captioning task.

III. Preliminaries

Differential privacy [24], [28], [29] provides a strong standard for privacy protection of data-driven algorithms, which refer to our PDD in this paper. Differential privacy, by its definition, is defined on two adjacent datasets which usually differ only by a single training example. A variant of differential privacy [30] is defined as follows:

Definition 1 (Differential Privacy)**.** A randomized algorithm $f : \mathcal{X} \rightarrow \mathcal{Y}$ with domain $\mathcal{X}$ and range $\mathcal{Y}$ satisfies (ϵ, δ)-differential privacy ((ϵ, δ)-DP) if for any two adjacent datasets $X, X' \in \mathcal{X}$ and any output set $Y \in \mathcal{Y}$ it holds that:

$$Pr[f(X) \in Y] \leq e^{\epsilon} Pr[f(X') \in Y] + \delta, \tag{1}$$

where $\epsilon > 0$ is the privacy budget and δ stands for the possibility that pure $(\epsilon, 0)$-DP might be broken. The smaller ϵ, δ means better privacy protection. In practice, a small δ can bring increment in the privacy budget ϵ.

Compared to (ϵ, δ)-DP, a strictly stronger privacy definition, Rényi differential privacy (RDP) is proposed by Mrionov [31]. In specific, RDP provides a more convenient and tighter way to bound the cumulative privacy loss of a sequence of adaptive and heterogeneous mechanisms.

Definition 2 (Rényi divergence)**.** For two distributions P and Q, the Rényi divergence of order α is defined as

$$D_{\alpha}(P \| Q) \frac{1}{\alpha - 1} \log \mathbb{E}_{x \sim Q} \left(\frac{P(x)}{Q(x)} \right)^{\alpha} \tag{2}$$

Definition 3 ((α, ϵ)-RDP)**.** A randomized mechanism $f : \mathcal{X} \rightarrow \mathcal{Y}$ is said to satisfy (α, ϵ)-RDP with $\alpha > 1$, if for any two adjacent datasets $X, X' \in \mathcal{X}$, it holds that

$$D_{\alpha}(f(X) \| f(X')) \leq \epsilon. \tag{3}$$

RDP is defined by Rényi divergence for $\alpha > 1$ and is a generalization of pure $(\epsilon, 0)$-DP. (α, ϵ)-RDP is equivalent to $(\epsilon, 0)$-DP when $\alpha = \infty$. Mrionov [31] also provides an easy way to convert RDP into (ϵ, δ)-DP for any $0 < \delta < 1$.

Lemma 1 (From RDP to (ϵ, δ)-DP)**.** If a mechanism f satisfies (α, ϵ)-RDP, then f satisfies $(\epsilon_f(\alpha), \delta)$-DP for any $\delta \in (0, 1)$ where $\epsilon_f(\alpha) = \epsilon + \frac{\log 1/\delta}{\alpha - 1}$.

However, in real practice, data processing is usually composed of a sequence of procedures. It is essential to bound the total privacy loss with each procedure a differentially private mechanism. RDP provides such a natural way to do composition.

Lemma 2 (Composition)**.** Let $f = (f_1, \cdots, f_t)$ be a sequence of adaptive mechanisms such that for any $i \in [t]$, $f_i : \mathcal{X} \times \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_{i-1} \rightarrow \mathcal{Y}_i$ satisfies (α, ϵ_i)-RDP, then f satisfies (α, ϵ)-RDP with $\epsilon = \sum_{i=1}^{t} \epsilon_i$.

To guarantee differential privacy in each mechanism, a common practice is adding Gaussian noise.

Lemma 3 (Gaussian Mechanism [32])**.** For a mechanism f with $\|f(X) - f(X')\|_2 \leq \Delta_2$ for any adjacent datasets X, X', the Gaussian mechanism is defined as

$$\mathbf{G}_{\sigma} f(X) = f(X) + \mathcal{N}(0, \sigma^2), \tag{4}$$

which satisfies $(\alpha, \epsilon_f(\alpha))$-RDP with $\epsilon_f(\alpha) = \frac{\alpha \Delta_2^2}{2\sigma^2}$.

Algorithm 1 The training process of the PDD model.

```
Prepare
initialize the Teacher Module $TM$
initialize the Privacy Filter module $PF$ = { Encoder, Aggregator, Privacy Barrier $PB$, Feature Augmenter module $FA$, Decoder }
initialize the Student Module $SM$
Input: private dataset $D$
Input: public text dataset $T$, public image dataset $I$
train $TM$ on the $D$
train $PF$ on the $T$
$C^{in} \leftarrow TM(i_m)$, where the image $i_m \in I$ and the caption set
    $C^{in} = \{c_1^{in}, ..., c_K^{in}\}$
$F^{in} \leftarrow Encoder(C^{in})$
    where the feature set $V^{in} = \{v_1^{in}, ..., v_K^{in}\}$
$v_{sum} \leftarrow Aggregator(V^{in})$
$v_{avg} \leftarrow PB(v_{sum})$
$V^{En} \leftarrow FA(v_{avg})$
    where the feature set $V^{En} = \{v_1^{En}, ..., v_N^{En}\}$
$C^{out} \leftarrow Decoder(V^{out})$ where the caption set $C^{out} = \{c_1^{out}, ..., c_N^{out}\}$
train $SM$ with the image $i_m$ and the generated corresponding caption set $C^{out}$
return $SM$
```

Notably, RDP has wide applications [26], [33] and is also closely related to moment accountant proposed by Abadi [5]. Under the framework of RDP, we are able to formulate the privacy analysis of PDD in section 4.3.

IV. Method

As shown in Figure 1, the PDD model contains three modules: teacher module, student module, and privacy filtering module. During the training process, these modules cooperate and are trained alternately.

Specifically, for the training settings, referring to the previous differential privacy model [25] and combining the characteristics of the image captioning task, we assume that there is a private image captioning dataset, a public text dataset, and a public image dataset. As shown in Algorithm 1, the training of the PDD model contains three processes:

- The teacher module is fully trained on the private dataset, which carries sensitive private information. Then, the module masters the ability to generate corresponding captions based on input images.
- We generate labels for public images using the teacher module and the privacy filter module. During this process, we first generate the captions for the public unlabeled images using the trained teacher module. And then, the privacy filter module, which is trained on the public text dataset, aggregates the captions output by the child modules of the teacher module to new captions as the pseudo labels for the public images. At this step, to realize the differential privacy guarantee, the middle features extracted from the input captions are eliminated sensitive

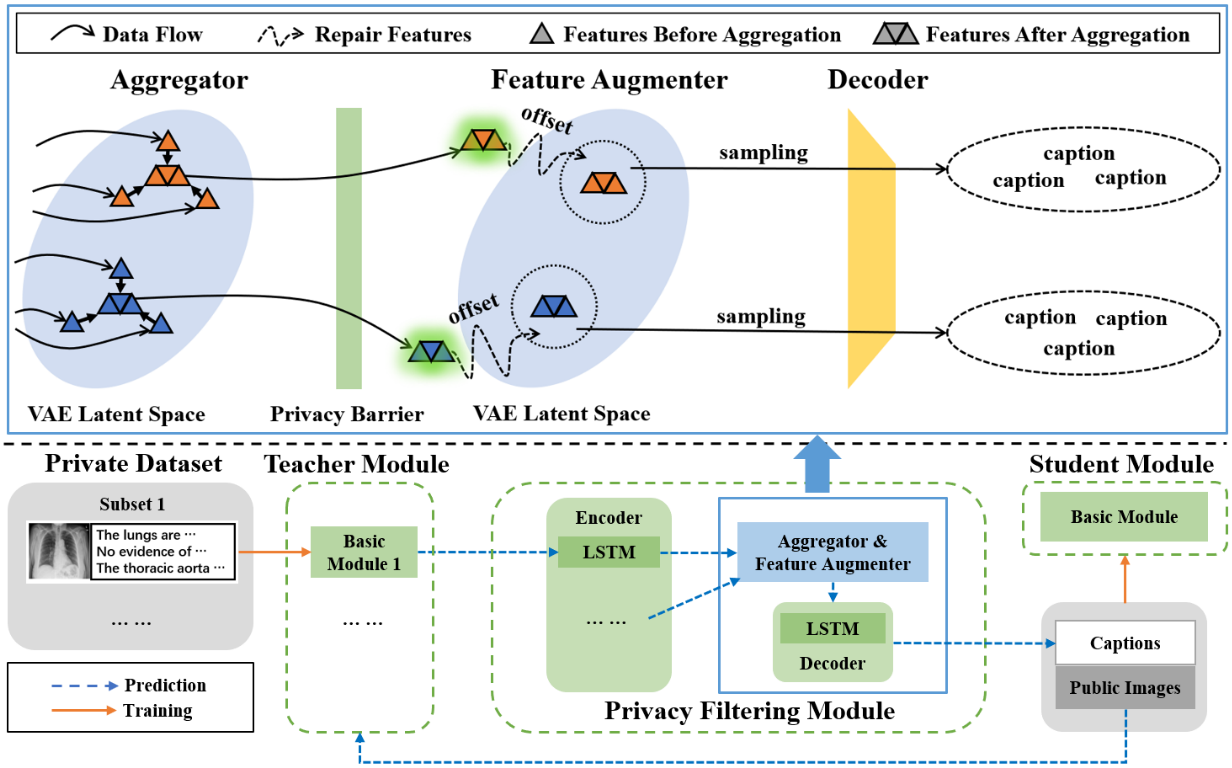

Fig. 1. The network architecture of the PDD. This model consists of a teacher module, a privacy filtering module, and a student module. The teacher module is trained on the private dataset and calculates out the captions for the public images. The privacy filtering module, which is the key part of the PDD, is responsible for privacy protection. In the process, it aggregates the caption information in the continuous latent space by summing the features directly, eliminates the privacy information of the features generated by aggregation, and finally calculates the offset to repair non-privacy information of the features. The student module is trained on the public image dataset and is made public.

information by adding Gaussian noise. Then, the feature augmenter module repairs the non-private information of these features disturbed by noise and augments features to more diverse features.

- The student module is trained on the public image dataset labeled by the teacher module and the privacy filter module together. After training, the model is made public. The private information in these image labels is effectively eliminated by the privacy filtering module. Thus, the student module can be trained fully using these labels without worrying about privacy issues.

In this section, we introduce the modules of the PDD in detail. Among them, the privacy filtering module as the key part is highlighted.

A. *Teacher Module and Student Module*

The teacher module and the student module contains N and *1* same basic image captioning modules respectively. For the teacher module, its N basic modules are trained on N datasets obtained by separating the private dataset, following the previous differential privacy model [25]. After training, it generates labels for the public dataset independently.

The basic module is designed referring to an existing model. Considering the purpose of selecting an appropriate reference method is getting a reliable example to verify the performance of the framework, we select a landmark classic model [17] as an example to verify the effectiveness of privacy protection. The basic module consists of a feature extractor and a text generator. Given an image *I*, the feature extractor extracts K image features $V = \{v_1, ..., v_K\}, v_k \in R^D$, from different regions of the image, and then, the text generator process the features V to generate the captions.

For the feature extractor, in the different scenes, we can use different models. If the object detection task is not feasible, we choose CNN to analyze the images.

$$F = CNN(I) \tag{5}$$

Where the $F \in R^{M*M*D}$ is the feature map of image *I*. We expand F into K D-dimensional vectors as V and input it to the next part, where $K = M*M$. Because the key information in the captions mainly comes from the objects in the images, if the object detection task is feasible, we choose Faster R-CNN to extract the features to localize the objects with bounding boxes and get the feature representation of them.

$$V = Faster\ R-CNN(I) \tag{6}$$

The text generator is designed by introducing the top-down attention mechanism. Specifically, this module consists of three parts: top-down attention LSTM, attention calculation, and language LSTM.

The top-down attention LSTM is used to calculate the key information that needs to be paid attention to in the next step.

$$h_t^1 = LSTM([h_{t-1}^2, \overline{v}, W_d], h_{t-1}^1) \tag{7}$$

Where the h_t^1 is the hidden state of the top-down attention LSTM at the time t, the h_t^2 is the hidden state of the language LSTM at the time t, the $\overline{v}$ is the mean of the V, and the W_d is the word generated from the time $t-1$.

The attention calculation part is designed to calculate the degree of attention to the feature v_k extracted from different areas of the image.

$$\alpha_t = softmax(wv_i + wh_t^1) \tag{8}$$

$$\hat{v}_t = \sum_{i=1}^{K} \alpha_{i,t} v_i \tag{9}$$

The language LSTM deals with the features $\hat{v}_t$ and the hidden state of the top-down attention LSTM h_t^1. Then, it outputs the word.

$$W_d = LSTM([\hat{v}_t, h_t^1], h_{t-1}^2) \tag{10}$$

B. Privacy Filtering Module

The main function of the privacy filtering module is to transmit the captions generated by the teacher module and protect the private information in them. To achieve this goal, we firstly aggregate the input captions to generate new captions, which prevents each output caption from relying on the detail of a single input caption. Then, we add Gaussian random noise to realize the differential privacy guarantee. However, the Gaussian noise not only eliminates the private information but also damages the non-private useful information. Thus, repairing the information of the middle features damaged by noise is another task. To complete this, we design the feature augmenter module for the privacy filtering module. In addition, to prevent the student module from overfitting, the feature augmenter module is also responsible for enhancing the features to generate diverse features, which are decoded into diverse captions as pseudo labels for the public images.

Specifically, we first need to realize the captions aggregation. The caption features can be added directly in the continuous latent space which is generated by the Variational Auto-Encoder (VAE) [34]. Traditional VAE consists of an encoder and a decoder. The encoder generates means and variances, which are used to take Gaussian sampling in the continuous latent space and get many latent variables. The decoder is able to transform the latent variables into the text. Based on this, we use the LSTM to encode multiple captions into means and variances. Then, we limit the modulus length of the means to C and sum them directly to generate a new feature vector v_{sum}.

$$v_{sum} = \sum_{i=0}^{N} clip(w_{mean} * LSTM(c_i)), \tag{11}$$

where the c_i is the caption output by the i-th teacher, w_{mean} is the learnable parameter, and the function $clip(.)$ limits the magnitude of the vector to the hyperparameter β.

Secondly, considering that the feature v_{sum} still contains lots of private information, adding noise to it to protect privacy is needed. The standard deviation δ of the Gaussian noise is calculated out by using and expanding differential privacy theory.

$$v_{avg} = (v_{sum} + N(0, \delta^2))/N \tag{12}$$

Finally, after getting the feature vector v_{avg}, the feature augmenter module of the privacy filter module repairs the non-private information of it, and enhances the features to generate a diverse feature set V_{En} by sampling in the continuous latent space built up by the variational auto-encoder.

$$V_{En} = Feature\ augmenter(v_{avg}) \tag{13}$$

Then, the decoder, which is a LSTM model, decodes V_{En} into caption set C_{De}.

$$C_{De} = LSTM(V_{En}) \tag{14}$$

For the feature augmenter module, it uses multiple fully connected layers to process v_{avg} and generates the offset vector v_{offset}. and the variance vector v_{var}.

$$v_{offset} = FC_{offset}(v_{avg}) \tag{15}$$

$$v_{var} = FC_{var}(v_{avg}) \tag{16}$$

The offset is added to the feature vector v_{avg} for repairing non-privacy information.

$$v_{rep} = v_{avg} + v_{offset} \tag{17}$$

After that, with the repaired feature vector v_{rep} as the mean and the vector v_{var} as the variance, we take Gaussian sampling in the continuous latent space generated by the variational auto-encoder to get the feature set V_{En}.

The self-supervised training process of the privacy filtering module is carried out in two steps. Because the function of the feature augmenter module is based on the exploration of the latent space, we first train the encoder and the decoder to build up the latent space, and then, train the feature augmenter to master the ability to repair and enhance features.

For the self-supervised training phase 1, the training target of the module are: 1) the distribution of the feature set S, which is generated by using the means v_{mean} and the variances v_{var} to take Gaussian sampling in latent space, obeys the normal distribution; 2) the decoder decodes the feature set S into the text similar to the input as much as possible.

$$L_{p1} = -D_{KL}(q(S|text)||p(S)) + E_{q(S|text)}[log\ p(text|S)] \tag{18}$$

For the self-supervised training phase 2, the feature augmenter is added to the training process and trained on the condition where other parameters of the privacy filter module are fixed. During training phase 2, the forward propagation process is similar to the inference process. Firstly, we inject Gaussian noise to the mean vector v_{mean} to simulate the process of implementing the differential privacy guarantee and generate a new mean vector. Secondly, we use multiple fully connected layers to deal with the v_{mean} for generating the new variance vector $\tilde{v}_{var}$ and the offset vector v_{offset}, which is used to repair the mean vector v_{mean}. Finally, we use the repaired mean vector $\tilde{v}_{mean}$ and the new variance vector $\tilde{v}_{var}$

to take Gaussian sampling in the continuous latent space to generate the new feature set $\tilde{S}$, which are decoded to text. The training targets of the feature augmenter module are: 1) the feature set $\tilde{S}$ sampled using the repaired mean and the new variance can be decoded more accurately; 2) the variance $\tilde{v}_{var}$ is constrained to be near the specified size.

$$L_{p2} = -log\ \tilde{v}_{var} + E_{q(\tilde{S}|text)}[log\ p(text|\tilde{S})] \tag{19}$$

C. Differential Privacy Guarantee of PDD

We now analyze the differential privacy guarantee of our proposed PDD. In our model, we add Gaussian noise in the process of generating captions for public images, which is the only process that reveals privacy information. According to the definition of DP, generating captions for every image would take some certain privacy budget according to the Gaussian mechanism. Based on the composition theorem of RDP, we can analyze the process and prove that our model meets the requirements of differential privacy.

Theorem 1. Suppose there are T images in the public dataset to be labeled, the sensitivity for the PDD is Δ_2 and the variance of the added Gaussian noise is σ^2. Then, for any $0 < \delta < 1$, there exists some ϵ that PDD satisfies (ϵ, δ)-DP.

Besides RDP, moments accountant (MA) [5] is also a common practice in differential privacy's application in deep learning [25], [27], which tracks privacy budget by summing the moments of all queries. In order to get a lower potential privacy cost of PDD, we also compute the privacy budget computed by MA. However, we found that RDP and MA always share the same privacy cost in our model. We prove the following theorem.

Theorem 2 (Equivalence of RDP and MA)**.** In the problem setting of Theorem 1, RDP and MA share the same privacy cost.

Theorem 1 and Theorem 2 together guarantee differential privacy property of PDD and goodness of our proved privacy bound. The detailed proofs of both Theorem 1 and Theorem 2 can be found in the Appendix.

V. Experiment

A. Experiment Setting

We describe the dataset overview, implementation details, and evaluation metrics in this section. In view of the limited paper space, we introduce the key parts of them, and the rest is shown in the appendix. For the convenience of description, we use the abbreviation to represent the modules of the PDD in the diagrams and the tables of the experiment section. The meaning of the abbreviation is described in detail in Table I. All experiments are conducted on a Linux server with 4 NVIDIA 1080Ti.

TABLE I
Abbreviation meaning table.

Abbreviations	Meanings
TM	teacher module
PF	privacy filter module
TM-PF	teacher module and privacy filter module joint use
FA	feature augmenter module
SM	student module
BM	basic module
Noisy-SGD	baseline of the PDD
Upper-Bound(PDD)	upper bound of the PDD
Upper-Bound	upper bound of all the privacy protection models

1) Dataset Description: We check the performance of the PDD on the two popular image captioning benchmark datasets, MS-COCO [7] and Flicker30k [8].

- **MS-COCO.** MS-COCO is a large-scale dataset, which is widely used in the image captioning task. The dataset contains 123,287 images, and for every image, there are 5 text annotations. There are two splits for the MS-COCO: the office split and the Karpathy split [35]. Considering the test set of the office split is not public, we use the Karpathy split, which is widely adopted by lots of image captioning papers. For the Karpathy split, the number of images in the train/val/test set is 113287/5000/5000.
- **Flickr 30K.** There are $31,783$ images with the language description in the Flickr 30K, which are collected from Flickr. Image-query pairs in this dataset mainly describe the daily activities of human beings. In addition, 5 language queries are labeled for each image. We follow the official split.

2) Implementation Details: For the basic module, we use the Faster R-CNN as the feature extractor, which is pretrained on the MS-COCO dataset [7]. During training the basic module, the batch size is 100, the learning rate is 0.00005, and the learning rate decays every 5 rounds. For the teacher module and the student module, the number of the teacher module's basic module, the scale of the noise, and the data volume of the public images used by the student module are related to the privacy cost. Like learning rate and other hyperparameters that need to be tried and adjusted for different application scenarios, we adjust these hyperparameters under different privacy requirements. In addition, the δ in differential privacy is set 10^{-5}. In order to facilitate the expression of privacy budget, we take log_{10} for it. For the privacy filter module, the data volume of the text dataset for training this module is 11112, which is divided from the image captioning dataset. During training, the batch size is 100, the learning rate is 0.5.

3) Evaluation Metrics: We use BLEU@N [9], METEOR [10], ROUGE-L [11], and CIDEr [12] to evaluate the accuracy of PDD, which are standard automatic evaluations of the image captioning task. And we use privacy budget ϵ calculated by theoretical derivation to evaluating the performance of privacy protection. For convenience, we take the logarithm of ϵ to the base 10 and denote it as log_ϵ.

TABLE II
COMPARISON WITH BASELINES ON MS-COCO AND FLICKER30K UNDER DIFFERENT PRIVACY BUDGETS. THE NOISY-SGD MODEL IS TRAINED USING THE PRIVATE DATASET AND THE PUBLIC DATASET TOGETHER. THE PDD REPRESENTS THE PERFORMANCE OF THE PDD'S PUBLIC MODULE (STUDENT MODULE), WHICH IS TRAINED ON THE PUBLIC IMAGES LABELED BY THE JOINT USE OF THE TEACHER MODULE AND THE PRIVACY FILTER MODULE. THE UPPER-BOUND(PDD) REPRESENTS THE STUDENT MODULE TRAINED ON THE PUBLIC IMAGES WITH MANUAL LABELS.

Dataset	Methods	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L	CIDEr	log_ϵ
MS-COCO	Noisy-SGD	2.98e-4	2.61e-12	1.27e-14	1.57e-13	2.39e-2	1.69e-1	1.21e-4	5.2
	PDD	**6.89e-1**	**5.10e-1**	**3.63e-1**	**2.51e-1**	**2.12e-1**	**5.02e-1**	**6.94e-1**	**5.2**
	Upper-Bound(PDD)	7.14e-1	5.46e-1	4.02e-1	2.90e-1	2.32e-1	5.25e-1	8.39e-1	-
Flicker30k	Noisy-SGD	2.08e-3	1.32e-4	5.45e-10	1.14e-12	1.76e-2	1.55e-1	3.90e-4	5.7
	PDD	**5.82e-1**	**3.62e-1**	**2.19e-1**	**1.36e-1**	**1.39e-1**	**3.83e-1**	**1.70e-1**	**5.7**
	Upper-Bound(PDD)	6.02e-1	4.04e-1	2.60e-1	1.66e-1	1.64e-1	4.15e-1	2.98e-1	-
Dataset	Methods	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L	CIDEr	log_ϵ
MS-COCO	Noisy-SGD	5.40e-7	5.35e-9	2.55e-14	5.70e-17	1.07e-2	8.36e-2	6.53e-5	1.3
	PDD	**6.07e-1**	**4.10e-1**	**2.58e-1**	**1.58e-1**	**1.72e-1**	**4.43e-1**	**4.24e-1**	**1.3**
	Upper-Bound(PDD)	6.66e-1	4.92e-1	3.46e-1	2.32e-1	2.02e-1	4.91e-1	6.20e-1	-
Flicker30k	Noisy-SGD	4.64e-15	1.91e-23	3.16e-26	1.32e-27	1.94e-3	2.84e-2	7.28e-5	1.6
	PDD	**5.24e-1**	**3.03e-1**	**1.70e-1**	**9.85e-2**	**1.17e-1**	**3.16e-1**	**9.08e-2**	**1.6**
	Upper-Bound(PDD)	6.02e-1	4.03e-1	2.60e-1	1.66e-1	1.64e-1	4.15e-1	2.98e-1	-

4) Comparison Methods Selection: Because our PDD model is the first one realizing the differential privacy guarantee specifically for the image captioning task, there is no same type of algorithms that can be directly compared. Thus, we choose the Noisy-SGD as the baseline, which is the general differential privacy algorithm suitable for almost all deep learning models.

B. Performance Comparisons

Considering that different levels of privacy protection in different scenarios are needed, we compare our PDD model with the baselines under different privacy budgets on two datasets. The experiment results of all methods are shown in Table II. We can observe from the table:

- The score of the PDD is higher than the Noisy-SGD on all evaluation metrics, which proves that our PDD is the optimal model that can be directly applied to the image captioning task at this stage.
- The performance of the PDD is close to the Upper-Bound(PDD). It proves that the quality of the non-private captions produced by the joint use of the teacher module and the privacy filter module is close to manual labels when the proportion of private data is large enough and exceeds 70% on two datasets. It shows the private filter module is well designed and the distillation framework is suitable for the image captioning task. In addition, for different privacy protection requirements, the performance of the Upper-Bound(PDD) is different. This is because captions generated for each public image increase the privacy cost. Thus, for different privacy protection requirements, the number of labeled public images used to train the student module is different.
- The PDD achieves the above two points on two different datasets. It proves that the model has a strong generalization ability and can handle different application scenarios of image captioning tasks.
- The PDD achieves the first point and the second point, especially for the relatively large privacy budget. It proves the PDD is suitable for different privacy protection needs.

C. Privacy Budget Changes

In the previous section, we only evaluate the PDD under a relatively large privacy budget and a small privacy budget. In order to evaluate the performance of the PDD in more different privacy protection requirements in-depth, we experiment under more different privacy budget log_ϵ settings and draw the accuracy curve. The changing trends on all evaluation metrics are consistent, we choose BLUE4 and METEOR as examples to analyze. The results as shown in Figure 2. In it, the solid lines represent the performance of the PDD and the baselines.

Compared with the Noisy-SGD, the PDD is better under all privacy budget settings on all evaluation metrics. It further proves the superiority of the PDD at this stage. In addition, the performance difference between the PDD and the Upper-Bound(PDD) is no more than 0.020 on the BLEU4 and no more than 0.015 on the METEOR. The close performance further proves the quality of the labels for the public images generated by the PDD is close to the manual labels when the number of the private data is more than 60000 on the MS-COCO. It proves the design of the PDD is reasonable.

We can also observe that the accuracy of PDD increases fast with slight increase of privacy cost. This is because as the privacy budget increases, the scale of the labeled public image dataset increases and the scale of the noise decreases, so the student module can get sufficient training by using more labeled data and more correct labels. Due to the rapid increase in accuracy, the BLEU4 is larger than 0.30 and the METEOR is larger than 0.23, when the privacy cost is no more than 4. The competitive accuracy proves that the PDD is suitable for scenarios with high privacy protection requirements. As the privacy cost increases, the performance of the PDD is approaching the Upper-Bound(PDD). And finally, the difference between the Upper-Bound(PDD) and PDD is no more than 0.01 on the BLEU4 and no more than 0.01 on the METEOR. The close performance proves the PDD can realize a small cost of accuracy for relatively low privacy protection requirements. In summary, the PDD is adapted to different scenes, and the degree of privacy protection can be set by the users to meet the requirements of privacy protection and accuracy.

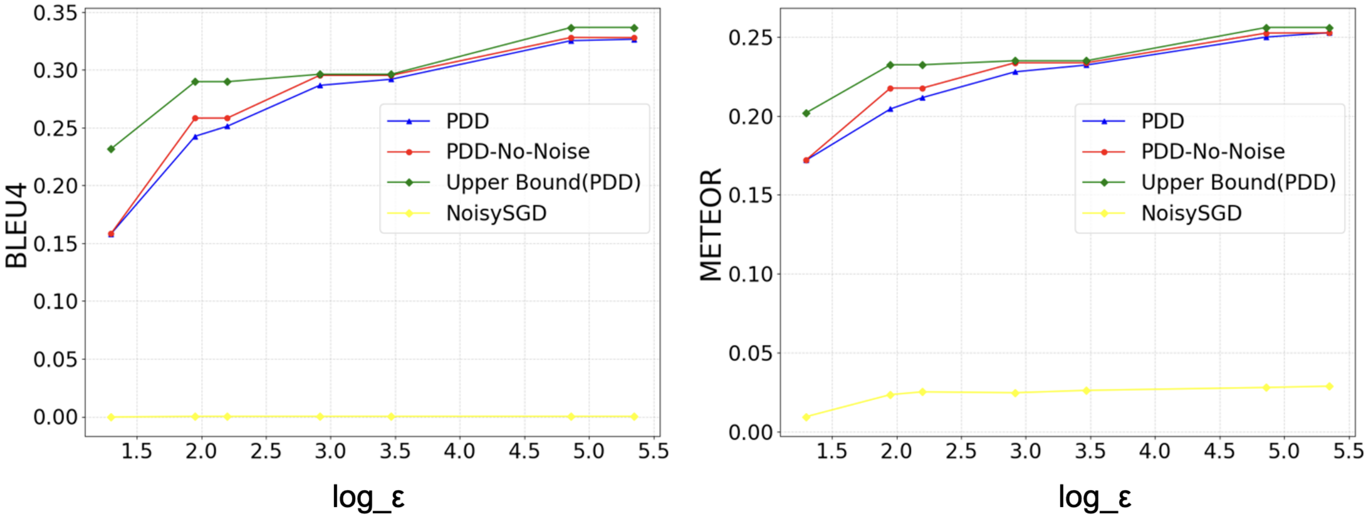

Fig. 2. Experimental results with different privacy budget settings on MS-COCO for BLUE4 and METEOR. PDD-No-Noise represents that no Gaussian noise is added during the training process of the PDD model.

TABLE III
ABLATION STUDY ON THE MS-COCO DATASET.

Methods	BLEU4	METEOR	ROUGE-L	CIDEr
PDD	1.58e-1	1.72e-1	4.43e-1	4.24e-1
PDD(w.o. FA)	1.56e-1	1.71e-1	4.44e-1	4.24e-1
PDD(w.o. VAE)	4.98e-14	1.76e-2	1.31e-1	9.67e-5

D. Ablation Study

The performance of PDD has been proven in the above experiment. In order to detail the effects of the key module and the privacy filter module, we design the ablation study, which includes the Variational Auto-Encoder (VAE) and the Feature Augmenter (FA). After removing the VAE, we directly add noise to the features of the generated language to realize privacy protection. When the privacy budget log_ϵ is 1.3, the performance of each module on the MS-COCO dataset is shown in Table III. From Table III, we can observe:

- Without the variational auto-encoder, the model accuracy drops dramatically. It proves the continuous latent space generated by the variational auto-encoder of our PDD is suitable for the straightforward fusion of features under privacy protection.
- The performance of modules used with the feature augmenter module is better than the modules without the feature augmenter module, which proves the feature augmenter module effectively corrects the non-privacy information in the middle features and effectively prevents the student module from overfitting.

VI. Conclusion

In this paper, we study the privacy protection problem for the image captioning task, and innovatively propose a privacy protection method, PDD. It consists of teacher modules, a privacy filter module and a student module. Among them, the privacy filter module firstly aggregates the caption information generated by the teacher sub-module in the continuous space. It makes each output caption not dependent on the sensitive details of a single input. Then, the privacy filter module eliminates the privacy of it with the help of differential privacy technology. During the process of eliminating the sensitive information, the key non-privacy information is damaged. Thus, we design a feature augmenter module for the privacy filter module to repair it. The experiments on MS-COCO and Flicker30k datasets verify the reliability of PDD.

References

[1] A. E. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng, “Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs,” *arXiv preprint arXiv:1901.07042*, 2019.

[2] G. Liu, T.-M. H. Hsu, M. McDermott, W. Boag, W.-H. Weng, P. Szolovits, and M. Ghassemi, “Clinically accurate chest x-ray report generation,” in *Machine Learning for Healthcare Conference*. PMLR, 2019, pp. 249–269.

[3] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, “Membership inference attacks against machine learning models,” in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 3–18.

[4] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song, “The secret sharer: Evaluating and testing unintended memorization in neural networks,” in *28th {USENIX} Security Symposium ({USENIX} Security 19)*, 2019, pp. 267–284.

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

[6] L. Xie, K. Lin, S. Wang, F. Wang, and J. Zhou, “Differentially private generative adversarial network,” *arXiv preprint arXiv:1802.06739*, 2018.

[7] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[8] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, “Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2641–2649.

[9] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[10] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proceedings of the ninth workshop on statistical machine translation*, 2014, pp. 376–380.

[11] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[12] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[13] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in *European conference on computer vision*. Springer, 2010, pp. 15–29.

[14] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Babytalk: Understanding and generating simple image descriptions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2891–2903, 2013.

[15] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[16] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[17] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.

[18] Y. Qin, J. Du, Y. Zhang, and H. Lu, "Look back and predict forward in image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8367–8375.

[19] Y. Zhou, M. Wang, D. Liu, Z. Hu, and H. Zhang, "More grounded image captioning by distilling image-text matching model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4777–4786.

[20] M. Cornia, L. Baraldi, and R. Cucchiara, "Show, control and tell: A framework for generating controllable and grounded captions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8307–8316.

[21] L. Guo, J. Liu, P. Yao, J. Li, and H. Lu, "Mscap: Multi-style image captioning with unpaired stylized text," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4204–4213.

[22] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[23] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, pp. 3–es, 2007.

[24] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of cryptography conference*. Springer, 2006, pp. 265–284.

[25] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," *arXiv preprint arXiv:1610.05755*, 2016.

[26] Y. Zhu, X. Yu, M. Chandraker, and Y.-X. Wang, "Private-knn: Practical differential privacy for computer vision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 854–11 862.

[27] J. Jordon, J. Yoon, and M. Van Der Schaar, "Pate-gan: Generating synthetic data with differential privacy guarantees," in *International Conference on Learning Representations*, 2018.

[28] C. Dwork, "A firm foundation for private data analysis," *Communications of the ACM*, vol. 54, no. 1, pp. 86–95, 2011.

[29] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy." *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211–407, 2014.

[30] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," in *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 2006, pp. 486–503.

[31] I. Mironov, "Rényi differential privacy," in *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*. IEEE, 2017, pp. 263–275.

[32] M. Bun and T. Steinke, "Concentrated differential privacy: Simplifications, extensions, and lower bounds," in *Theory of Cryptography Conference*. Springer, 2016, pp. 635–658.

[33] N. Papernot, S. Song, I. Mironov, A. Raghunathan, K. Talwar, and Ú. Erlingsson, "Scalable private learning with pate," *arXiv preprint arXiv:1802.08908*, 2018.

[34] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[35] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

Appendix
Theorem Proof and Experimental Supplementary

A. Proof for Theorem 1

We begin by restating Theorem 1.

Theorem 3. *Suppose there are T images in the public dataset to be labeled, the sensitivity for the PDD is Δ_2 and the variance of the added Gaussian noise is σ^2. Then, for any $0 < \delta < 1$, PDD satisfies (ϵ, δ)-DP.*

Proof. Note that for a single public image, first we aggregate the teachers' output captions into a caption feature, then we sample feature points in the feature space and decode them. The output of label generation could be deemed as the aggregated caption feature with sensitivity Δ_2^2 instead of the final labeled captions. Thus, for a single image labels generation process as a query, according to Lemma 3, the process satisfies $(\alpha, \frac{\alpha\Delta_2^2}{2\sigma^2})$-RDP for any $\alpha > 1$. With T images in total, Lemma 2 provides a total privacy cost bound for the labeling process as a composition process that satisfies $(\alpha, \frac{T\alpha\Delta_2^2}{2\sigma^2})$-RDP. Then with the help of Lemma 1, PDD satisfies $(\epsilon_g(\alpha), \delta)$-DP for any $\alpha > 1$ and $0 < \delta < 1$, where

$$\epsilon_g(\alpha) = \frac{T\alpha\Delta_2^2}{2\sigma^2} + \frac{\log 1/\delta}{\alpha - 1}. \tag{20}$$

Finally, when we choose $\alpha = 1 + \frac{\sigma}{\Delta_2}\sqrt{\frac{2}{T}\log\frac{1}{\delta}}$, we get

$$\epsilon_g(\alpha) = \frac{\Delta_2\sqrt{T}}{\sigma}(\sqrt{2\log\frac{1}{\delta}} + \frac{\Delta_2\sqrt{T}}{2\sigma}), \tag{21}$$

which is the minimum value of $\epsilon_g(\alpha)$ as the lowest privacy bound of PDD. □

B. Proof for Theorem 2

First we provide the basic definition and theorems of moments accountant.

Definition 4 (Privacy Loss)**.** Let $f : \mathcal{X} \to \mathcal{Y}$ be a randomized algorithm with domain $\mathcal{X}$ and range $\mathcal{Y}$, aux be an auxiliary input and $X, X' \in \mathcal{X}$ be two adjacent datasets. For an outcome $o \in \mathcal{Y}$, the privacy loss at o is defined as:

$$c(o; f, aux, X, X') \log \frac{Pr[f(aux, X) = o]}{Pr[f(aux, X') = o]}. \tag{22}$$

Definition 5 (Moments Accountant)**.** Let f be a randomized algorithm, The moments accountant is defined as:

$$\alpha_f(l) \max_{aux,X,X'} \alpha_f(l; aux, X, X'). \tag{23}$$

where $\alpha_f(l; aux, X, X') \log \mathbb{E}_{o\sim f(aux,X)}[\exp(lc(o; f, aux, X, X'))]$ is the log of the moment generating function of the privacy loss random variable.

With the definition of moments accountant, there are also two theorems to bound the privacy loss.

Theorem 4 (Composability)**.** Suppose a randomized algorithm f consists of a sequence of adaptive algorithms $f_1, ..., f_T$ where $f_i : \prod_{j=1}^{i-1} \mathcal{Y}_j \times \mathcal{X} \to \mathcal{Y}_i$. Then, for any l

$$\alpha_f(l) \leq \sum_{i=1}^{T} \alpha_{f_i}(l). \tag{24}$$

Theorem 5 (Tail Bound)**.** Let f be a randomized algorithm. For any $\epsilon > 0$, f is (ϵ, δ)-Dp for

$$\delta = \min_l \exp(\alpha_f(l) - l\epsilon) \tag{25}$$

Then we restate Theorem 2 and give the proof.

Theorem 6 (Equivalence of RDP and MA)**.** In the problem setting of Theorem 1, RDP and MA share the same privacy cost.

Proof. For Take the definition of privacy loss into moments accountant, we have

$$\begin{aligned}\alpha_f(l) &= \max_{aux,X,X'} \alpha_f(l; aux, X, X') \\ &= \max_{aux,X,X'} \log \mathbb{E}_{o\sim f(aux,X)}[\exp(lc(o; f, aux, X, X'))] \\ &= \max_{aux,X,X'} \log \mathbb{E}_{o\sim f(aux,X)} \left(\frac{Pr[f(aux, X) = o]}{Pr[f(aux, X') = o]}\right)^l \\ &= \max_{aux,X,X'} \log \mathbb{E}_{o\sim f(aux,X')} \left(\frac{Pr[f(aux, X) = o]}{Pr[f(aux, X') = o]}\right)^{l+1}.\end{aligned}$$

Remind the form of Rényi divergence, we have

$$\alpha_f(l) = \max_{aux,X,X'} lD_{l+1}(f(X) \| f(X')). \tag{26}$$

According to Lemma lem3 and Definition def3, for a single public image label generation process f_i, we have

$$D_{l+1}(f_i(X) \| f_i(X')) \leq \frac{(l+1)\Delta_2^2}{2\sigma^2}. \tag{27}$$

Due to the arbitrariness of X, X', we have

$$\alpha_{f_i}(l) \leq \frac{l(l+1)\Delta_2^2}{2\sigma^2}. \tag{28}$$

According to Theorem thm4, for the whole process of T images' label generation, we have

$$\alpha_f(l) \leq \sum_{i=1}^{T} \alpha_{f_i}(l) \leq \frac{Tl(l+1)\Delta_2^2}{2\sigma^2}. \tag{29}$$

By taking logarithm to Equation tailbound, we have

$$\log \delta = \min_l(\alpha_f(l) - l\epsilon) = \min_l(\frac{Tl(l+1)\Delta_2^2}{2\sigma^2} - l\epsilon) \tag{30}$$

Through extreme value calculation of the right side of Equation 30, we could reach the same form of ϵ as 21. □

C. *Experimental Supplementary Materials*

We use the MS-COCO dataset and the Flicker30k to test the performance of our model. The MS-COCO dataset is a large and rich object detection segmentation and captioning dataset. This dataset aims to scene understanding, which is mainly selected from complex daily scenes. MS-COCO contains lots of versions. The one we used is Captioning 2015, which is published in 2014. In it, each caption describes one image content, and each image has at least 5 caption annotations. Flickr 30K consists of images collected from Flickr. Image-query pairs in this dataset mainly describe the daily activities of human beings. To evaluate the performance of the PDD, we choose the BLEU@N, METEOR, ROUGH-L, CIDEr, and the privacy budget. BLEU is an evaluation metric, whose full name is Bilingual Evaluation Understudy. It is designed for evaluating the translation result on behalf of the person. In the machine translation task and image/video captioning task, the BLEU@N is very commonly accepted by researchers and often used to evaluate the difference between the generated text and the target text. Specifically, this method calculates the N-grams of the two sentences and counts the number of matches. The value of it is between 0.0 and 1.0. The larger the value, the higher the degree of matching between the texts. METEOR is proposed in 2005 and it is designed to make up for the shortcomings of BLEU4. The main content of this evaluation metric is that it makes a word-to-word matching between the candidate and the reference, and every word is matched with zero or one word. Specifically, it contains two stages. In the first stage, all the possible matches are listed. In the second stage, choose the one with the most successful words among all the possible matches. According to the degree of the matching, the METEOR can be calculated. Different from the above two methods, the ROUGE-L is not based on the N-grams but uses the longest common sub-column. It makes the ROUGE-L capture the structural features in sentences. CIDEr is designed specifically for the image captioning task. It is also based on the N-grams. However, CIDEr is added the TP-IDF to re-weight the different N-gram. In this way, it is avoided that some common but not informative N-grams are scored high when evaluating caption. Privacy budget is the evaluation metrics for degree of the privacy protection using differential privacy. For different algorithms, the calculation formula is different. For the PDD, the formula is proven in the last section. Considering the actual needs, we use a logarithm to express them. This value mainly depends on the relative size to reflect on the privacy protection capabilities of different algorithms. And it is adjusted according to the different privacy budgets of the usage scenarios. In an environment that does not require high-strength privacy protection, setting it larger results in better accuracy. Considering that there are no obvious advantages and disadvantages of the evaluation metrics, we use them together referring to existing researches to fully test the effect of the model. The situation is as we expected. Our model achieves the best balance between performance and privacy protection.