RadSLDP: Selective Local Differential Privacy for Radiology Vision-Language Models

Konghao Zhao[1†], Xinyang Xu[1†], Runhui Xu[1†], Yutaka Natsuaki[2], Wenjie Xiong[3], Sai Praneeth Karimireddy[1], and Ruishan Liu[1,2*]

[1] Department of Computer Science, Viterbi School of Engineering, University of Southern California, Los Angeles, CA, USA
[2] Department of Radiation Oncology, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA
[3] Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA
ruishanl@usc.edu

Abstract. Medical Vision–Language Models (VLMs) for radiology report generation are often fine-tuned by external developers due to hospitals' limited computational infrastructure. It requires sensitive patient data to be transmitted to untrusted training environments, which creates privacy risks. Standard Differential Privacy protects the trained model but assumes trusted access to raw data, making it unsuitable for untrusted external training. Local Differential Privacy (LDP) enables privacy-preserving training in untrusted settings by perturbing all text token embeddings before transmission but significantly degrading clinical utility. We observe that clinical reports contain both image-inferable and non-inferable content and define the non-inferable portion as non-visual clinical context (NVCC), such as demographics and medical history, which constitutes the primary privacy risk. To preserve privacy while maintaining utility, we propose RadSLDP, a selective LDP mechanism that perturbs only expert-validated NVCC token embeddings, with each protected token embedding satisfying formal per-token (ε, δ)-LDP via the Gaussian mechanism. As a result, RadSLDP matches or outperforms uniform LDP in 94% of comparisons, including 29/36 improvements and 5/36 ties. It improves BLEU-4 by 24–77% and CheXbert Micro-F1 by 87–108% on ReXGradient dataset. RadSLDP also exhibits lower performance variability across privacy budgets, enabling practical privacy-preserving VLM fine-tuning in untrusted settings.

Keywords: Local Differential Privacy · Vision-Language Models · Radiology Report Generation.

1 Introduction

Medical Vision-Language Models (VLMs) have demonstrated remarkable capability in generating radiology reports from chest X-ray images, offering the

[†] Equal contribution, [*] Corresponding author

potential to reduce radiologists' workload and improve diagnostic consistency. However, modern VLMs comprise billions of parameters, making fine-tuning computationally prohibitive for most hospital systems with limited infrastructure. Hospitals must therefore rely on external cloud providers or model developers, requiring transmission of raw training data to untrusted environments. However, training datasets often contain sensitive patient information, including demographics, medical history, and clinical indications. Such outsourcing introduces fundamental privacy risks, as external trainers gain direct access to sensitive patient information, potentially violating regulatory frameworks (Fig. 1a).

Differential Privacy (DP) [6] is the standard approach for training neural networks with formal privacy guarantees by ensuring that model outputs are approximately invariant to the inclusion of any individual training sample. DP-SGD [1] achieves this in neural networks through per-sample gradient clipping and noise injection, and has been applied to vision and language tasks [37,15,32]. Recent methods further apply and improve this [31,27,36] in multi-modal training. However, DP-SGD assumes a trusted training environment and protects only the released model, making it unsuitable for untrusted trainer settings common in medical VLM development.

Local Differential Privacy (LDP) protects data in the untrusted trainer setting by perturbing inputs before they leave the trusted domain [5]. Prior work applies LDP to text through binary embedding perturbation [18], sentence-level mechanisms [4], and private inference methods [8]. However, existing approaches apply uniform noise to all tokens, resulting in substantial degradation in downstream task performance [11].

This limitation highlights an important structural property of radiology VLM training data. Clinical text accompanying chest X-rays contains two fundamentally different types of information. Image-inferable content, such as descriptions of cardiomegaly or device placement, reflects clinical findings visible in the radiograph. In contrast, image-uninferable content, including patient demographics and medical history, cannot be derived from the image alone. We define this portion as Non-Visual Clinical Context (NVCC) which constitutes the primary privacy risk.

This raises another question of whether de-identified chest radiographs carry residual privacy risk. In standard data-sharing workflows, hospitals remove all categories of identifiers specified by HIPAA Safe Harbor [28] before releasing imaging data, and public datasets such as MIMIC-CXR [13], CheXpert Plus [3], and ReXGradient [35] follow these protocols. Unlike head CTs or full-body MRIs, standard chest radiographs lack directly recognizable features such as facial structure. Prior work suggests that re-identification becomes plausible primarily in settings that enable cross-dataset linkage, for example when an adversary has access to identified reference images [22]. In typical outsourced training workflows, hospitals may share de-identified radiographs for model development but do not share institution-held identified image repositories with external parties; this limits the auxiliary data needed for cross-dataset linkage attacks. This moti-

vates selectively perturbing NVCC tokens to protect patient-specific information while preserving diagnostic signal from the image.

Prior work has explored selective LDP to reduce the utility loss of uniform noise by restricting perturbation to sensitive content. Shi et al. [24] introduced this for RNN/LSTM architectures. Just Fine-tune Twice [25] extends it to Transformers via two-phase training. Policy-based embedding-level noise injection [30], adaptive token-weighted noise [33], split-and-denoise architectures [19], and gradient-based selective fine-tuning [34] further develop this direction. However, all existing selective DP methods target unimodal text-only models. We are the first to formalize and evaluate selective LDP for radiology VLM fine-tuning.

With this insight, we propose RadSLDP, a selective embedding-level LDP mechanism for privacy-preserving radiology VLM fine-tuning. To identify sensitive content, we develop a multi-stage annotation pipeline: clinical indications from three large-scale datasets are decomposed into 141,129 semantically unique units, clustered into 347 unique subcategories, and annotated for image-inferability using an LLM, followed by review from a certified medical physicist. RadSLDP then operates within the trusted hospital environment: for each training sample, a policy function identifies NVCC tokens. Their embeddings are clipped to bound ℓ_2-sensitivity and perturbed with calibrated Gaussian noise. In contrast, image-inferable token embeddings are transmitted unmodified. Each protected token embedding satisfies formal per-token (ε, δ)-LDP via the Gaussian mechanism [6], with guarantees that hold independently of the downstream model architecture.

We validate RadSLDP across three chest X-ray datasets and three VLM architectures. At per-token ε=8, RadSLDP matches or outperforms uniform LDP in 94% of comparisons (34/36), with BLEU-4 improvements of 25–77% and CheXbert Micro-F1 gains of 87–108% on the ReXGradient dataset (Fig. 1c). RadSLDP also exhibits lower performance variability across privacy budgets than uniform LDP across all three architectures. Our main contributions are as follows:

- We study radiology VLM fine-tuning with an untrusted external trainer, and curate an expert-validated NVCC taxonomy to derive token-level sensitivity labels across three datasets.
- We propose RadSLDP, a model-agnostic selective LDP mechanism that perturbs only NVCC token embeddings with formal per-token (ε, δ)-LDP guarantees per protected token.
- We conduct extensive experiments across three large-scale datasets, demonstrating that selective perturbation achieves substantial utility gains over uniform LDP.

2 Methodology

2.1 Problem Formulation

We consider a setting where a trusted hospital holds a dataset $\mathcal{D} = \{(I_i, X_i, Y_i)\}_{i=1}^{N}$, and I_i is a 2D chest X-ray image, X_i is the associated clinical context text, and Y_i

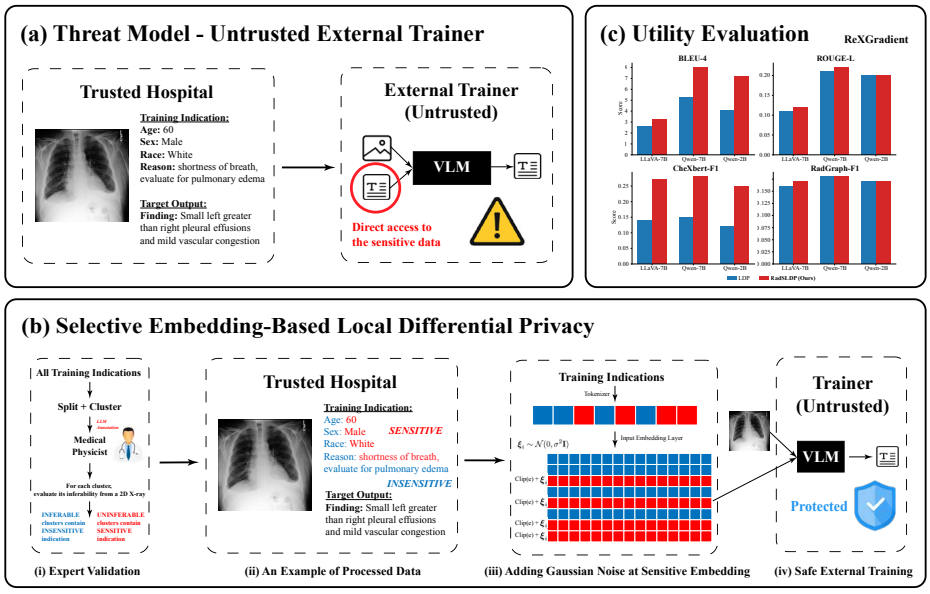

Fig. 1. Overview of RadSLDP for privacy-preserving clinical VLM fine-tuning.

is the target radiology report. The hospital lacks the computational resources to fine-tune a large VLM and must transmit training data to an untrusted external trainer. The goal is to protect sensitive patient attributes in X_i while preserving sufficient clinical signal to produce a high-quality report generation model.

2.2 Sensitive Patient Attribute Identification

By our observation, clinical context text X_i contains two fundamentally different types of information (NVCC and non-NVCC). To distinguish between these categories, we define a deterministic policy function $\pi : X \to \mathcal{W}$ that maps clinical context text to a set of sensitive words $\mathcal{W}$. The policy combines two complementary components: (1) a predefined rule that matches patient demographic attributes (e.g., age, sex, race, ethnicity), and (2) a data-driven annotation process.

We build our taxonomy using clinical indications pooled from three large, publicly available chest radiology datasets (see Sec. 3.1 for details), covering diverse institutions and clinical settings. We first normalize and split all clinical indications across three datasets into semantically coherent and unique units, yielding 141,129 unique parts. These parts are embedded using a domain-specific biomedical encoder based on PubMedBERT [9] which are then clustered using HDBSCAN [2] into 347 subcategories, each representing a group of semantically similar clinical expressions. We then label each subcategory for image-inferability using GPT-5.2 [21]. To further validate the labels, a certified medical physicist reviews all subcategories; 342 out of 347 labels are accepted and the remaining

Table 1. Taxonomy of clinical indications. For the example subcategory, "✓" denotes image-inferable and "×" denotes non-inferable. "# Sub." indicates the number of subcategories and "# Indi." indicates the number of unique granular indications under each summarized category.

Summarized Category (13)	**Example Subcategory (347)**	**# Sub.**	**# Indi.**
Thoracic Pathology Evaluation	Evaluation for pneumonia or pneumothorax (✓)	135	54,799
Medical Devices, Lines & Tubes	Status post chest tube placement or removal (✓)	41	19,527
Demographic & Administrative Context	Patient age (×)	36	17,603
Post-Procedural & Perioperative Context	Cesarean section delivery or birth (×)	27	10,151
Social History, Risk Factors & Exposures	History or presence of HIV/AIDS infection (×)	14	9,787
Non-Thoracic Conditions Prompting Chest Imaging	Nausea and vomiting symptoms (×)	16	5,865
Cardiovascular & Hemodynamic Conditions	Evaluation for mediastinal widening (✓)	22	5,196
Respiratory Support & Airway Management	Hypoxemic respiratory failure or distress (×)	12	4,695
Trauma, Injury & Hemorrhage History	Rib fracture(s) (✓)	11	4,338
Clinical Symptoms & Functional Status	Generalized or progressive weakness (×)	14	3,598
Neurologic & Cognitive Conditions	Altered mental status (×)	9	2,610
Laboratory & Physiologic Abnormalities	Evaluation or monitoring of leukocytosis (×)	8	2,542
Infectious, Inflammatory & Immune-Related	Evaluation or management of cellulitis (×)	2	418
Total		**347**	**141,129**

ones are corrected (Fig. 1b). The image-inferability label of each subcategory is propagated to all member phrases within its cluster.

These image non-inferable indications, together with predefined demographic attributes, constitute the sensitive set (Table 1). Across all datasets, an average of 10.6 out of 28.7 tokens per sample (37%) are identified as sensitive NVCC tokens. The resulting taxonomy captures recurring semantic patterns in radiology indications and supports incremental maintenance: new indications can be embedded and either assigned to existing subcategories or grouped into new clusters for review, without changing previously validated labels. For substantially out-of-distribution indications, additional expert annotation may be required before deployment.

2.3 Selective Embedding-Level Local Differential Privacy

Given the sensitive word set $\mathcal{W}_i = \pi(X_i)$ produced by the policy function, we tokenize the input as $\texttt{ids} = \text{Tok}(X_i)$ of length L_i and construct a binary mask $m_i = \text{Mask}(\mathcal{W}_i, \texttt{ids}) \in \{0,1\}^{L_i}$. For each token position $k \in \{1, \ldots, L_i\}$, $m_{i,k} = 1$ indicates that token position k corresponds to a sensitive NVCC attribute and $m_{i,k} = 0$ indicates an image-inferable token that is left unperturbed under the stated outsourced-training threat model.

Rather than applying uniform noise to all token embeddings as in standard LDP, we selectively perturb only the embeddings at sensitive positions identified by the policy mask m_i. For each sample (I_i, X_i, Y_i) in a training batch, the mechanism proceeds as follows.

Encoding. The image and clinical text are independently encoded to produce visual embeddings $E_i^V = f_V(I_i; \theta_V)$ and text embeddings $E_i^X = f_X(X_i; \theta_X) \in \mathbb{R}^{L_i \times d}$, where d is the encoding dimension. Critically, f_X denotes the non-trainable token embedding layer, which maps each token independently to a fixed vector.

Selective perturbation. For $k \in \{1, \ldots, L_i\}$, if $m_{i,k} = 1$ (sensitive), we clip the embedding to bound its ℓ_2-norm and add calibrated Gaussian noise:

$$\tilde{e}_{i,k} = E^X_{i,k} \cdot \min\left(1, \frac{C}{\|E^X_{i,k}\|_2}\right), \qquad E'^X_{i,k} = \tilde{e}_{i,k} + z, \quad z \sim \mathcal{N}(0, \sigma^2 I_d), \tag{1}$$

where C is the clipping norm. If $m_{i,k} = 0$ (non-sensitive), the embedding is passed through unmodified: $E'^X_{i,k} = E^X_{i,k}$ (Fig. 1b).

Fusion and training. The privatized text embeddings E'^X_i are fused with visual embeddings varied by model architecture, and the model is then trained with standard gradient descent on the report generation loss. The external trainer receives only the privatized embeddings and performs standard fine-tuning without access to the original sensitive information.

2.4 Privacy Analysis

Definition 1 (per-token (ε, δ)-Local Differential Privacy). *A randomized mechanism $\mathcal{M}$ operating on token embeddings satisfies per-token (ε, δ)-local differential privacy if, for any two possible embeddings $e, e' \in \mathbb{R}^d$ and any measurable set $\mathcal{S}$:*

$$\Pr[\mathcal{M}(e) \in \mathcal{S}] \leq e^{\varepsilon} \Pr[\mathcal{M}(e') \in \mathcal{S}] + \delta, \qquad \varepsilon > 0,\ \delta \in [0, 1). \tag{2}$$

Theorem 1. *Let $e \in \mathbb{R}^d$ denote the embedding of a sensitive token. The mechanism*

$$\mathcal{M}(e) = \tilde{e} + z, \qquad \tilde{e} = e \cdot \min\left(1, \frac{C}{\|e\|_2}\right), \quad z \sim \mathcal{N}(0, \sigma^2 I_d), \tag{3}$$

satisfies per-token (ε, δ)-local differential privacy for any $\varepsilon > 0$ and $\delta \in (0, 1)$ provided

$$\sigma \ \geq \ \frac{2C}{\varepsilon} \sqrt{2 \ln\left(\frac{1.25}{\delta}\right)}. \tag{4}$$

Proof. After clipping, $\|\tilde{e}\|_2 \leq C$ for any input e. For any two possible token embeddings e, e' with clipped representations $\tilde{e}, \tilde{e}'$, the ℓ_2 sensitivity satisfies

$$\Delta_2 = \|\tilde{e} - \tilde{e}'\|_2 \leq \|\tilde{e}\|_2 + \|\tilde{e}'\|_2 \leq 2C. \tag{5}$$

The mechanism $\mathcal{M}(e) = \tilde{e} + z$ with $z \sim \mathcal{N}(0, \sigma^2 I_d)$ is therefore the Gaussian mechanism with sensitivity $\Delta_2 \leq 2C$. By the standard Gaussian mechanism bound [6], choosing σ as above ensures per-token (ε, δ)-differential privacy. Since the perturbation is applied locally to each embedding prior to any transmission or external computation, the mechanism satisfies per-token (ε, δ)-local differential privacy by definition.

Remark 1 (Composition and model independence). Theorem 1 establishes a per-token (ε, δ)-LDP guarantee for each protected (NVCC) token embedding. Under the record-level adjacency where unprotected content is fixed and only protected NVCC tokens may differ, releasing k independently privatized protected embeddings yields a conservative $(k\varepsilon, k\delta)$-LDP record-level bound by sequential composition [7] (with tighter bounds via advanced composition [14] or Rényi DP [20]). Accordingly, when experiments report utility at a fixed per-token (ε, δ), the induced conservative record-level bound varies across records through their realized k.

In this paper, we focus on per-token LDP because it matches the privacy-critical interface of the setting: NVCC embeddings are privatized locally before transmission, and leakage in clinical text arises from inference on localized sensitive spans. If record-level privacy is instead the target, our selectivity also improves record-level accounting relative to a uniform token-level LDP baseline. Under conservative sequential composition, perturbing all L tokens leads to a $(L\varepsilon, L\delta)$ record-level bound, whereas perturbing only the k protected tokens leads to $(k\varepsilon, k\delta)$. Since $k < L$, our method composes over fewer releases and therefore allows less noise per protected output under matched record-level privacy budgets. Concretely, under a matched record-level budget $(\varepsilon_{\text{rec}}, \delta_{\text{rec}})$, uniform LDP allocates $(\varepsilon_{\text{rec}}/L, \delta_{\text{rec}}/L)$ per token, while RadSLDP allocates $(\varepsilon_{\text{rec}}/k, \delta_{\text{rec}}/k)$ only to protected NVCC tokens. Since $k = 10.6 < L = 28.7$ on average, RadSLDP applies less noise per protected token under the same record-level bound, improving utility while preserving the matched record-level guarantee.

3 Experiments

3.1 Experiment Setup

Datasets and Evaluation We evaluate on three public chest X-ray datasets. CheXpert Plus [3] contains 187,711 radiology studies with text radiology reports and patients' demographics; ReXGradient-160K [35] contains 160,000 radiology studies within 3 US health systems; MIMIC-CXR [13] contains 227,835 imaging studies presented to the Beth Israel Deaconess Medical Center Emergency Department. Together, these datasets comprise over 500,000 studies with paired radiographs and reports containing clinically rich indication text and demographic attributes. Model utility is evaluated using lexical overlap metrics BLEU-4 [23] and ROUGE-L [16], and clinically oriented metrics CheXbert Micro-F1 [26] and RadGraph F1 [12].

Implementation. We fine-tune LLaVA-1.5-7B [17], Qwen2-VL-7B and 2B [29] using LoRA [10] (rank 32) for 2 epochs with learning rate 2×10^{-4}, maximum sequence length 4096, clipping norm C=1.0, and privacy parameter δ=$1/n$, where n is the training-set size of the corresponding dataset, on 8 NVIDIA A6000 GPUs. Training is implemented in PyTorch using the Hugging Face Trainer with the AdamW optimizer, warmup ratio 0.03, and bf16 mixed precision.

Table 2. Utility results across models and datasets. ⋆*DP-SGD and non-private training assume access to unperturbed training data and therefore do not apply to the external-training privacy setting considered in this work; we report them for reference only.*

			ReXGradient				CheXpert PLUS				MIMIC-CXR			
Model	**Training**	ε	BLEU↑	RG-L↑	Chex-F1↑	Rad-F1↑	BLEU↑	RG-L↑	Chex-F1↑	Rad-F1↑	BLEU↑	RG-L↑	Chex-F1↑	Rad-F1↑
LLaVA-7b	Non-private⋆	-	3.98	0.13	0.29	0.18	5.99	0.17	0.53	0.21	5.04	0.16	0.39	0.19
	DP-SGD⋆	8	1.95	0.12	0.18	0.15	1.05	0.13	0.30	0.08	3.17	0.15	0.18	0.17
	LDP	8	2.60	0.11	0.14	0.16	4.30	0.16	0.46	0.19	4.32	0.15	0.37	0.17
	RadSLDP	8	**3.26**	**0.12**	**0.27**	**0.17**	**4.54**	**0.17**	**0.51**	**0.20**	**4.52**	0.15	**0.40**	**0.19**
Qwen-7b	Non-private⋆	-	10.42	0.24	0.36	0.19	10.08	0.22	0.49	0.20	7.23	0.23	0.36	0.19
	DP-SGD⋆	8	4.62	0.19	0.18	0.15	1.51	0.16	0.30	0.09	3.27	0.19	0.16	0.14
	LDP	8	5.30	0.21	0.15	0.18	7.27	0.19	0.49	0.19	**6.50**	0.22	**0.35**	0.18
	RadSLDP	8	**7.99**	**0.22**	**0.28**	0.18	**7.86**	0.19	**0.51**	**0.20**	6.06	**0.23**	0.30	**0.19**
Qwen-2b	Non-private⋆	-	9.10	0.23	0.30	0.18	8.86	0.21	0.40	0.19	7.25	0.22	0.30	0.18
	DP-SGD⋆	8	4.47	0.18	0.17	0.14	2.31	0.19	0.25	0.12	3.81	0.19	0.11	0.14
	LDP	8	4.06	0.20	0.12	0.17	5.50	0.17	0.42	0.15	6.06	0.21	0.28	0.16
	RadSLDP	8	**7.20**	0.20	**0.25**	0.17	**5.88**	**0.18**	**0.43**	**0.18**	**6.72**	**0.22**	**0.32**	**0.18**

Table 3. Effect of per-token privacy budget ε on utility on the ReXGradient dataset. $\varepsilon = \infty$ represents clipping only. Std is computed over $\varepsilon \in \{4, 8, \infty\}$.

		$\varepsilon = 4$	$\varepsilon = 8$	$\varepsilon = \infty$	**Overall**
Model	**Training**	BLEU↑	BLEU↑	BLEU↑	BLEU Std$_\varepsilon$ ↓
LLaVA-7B	LDP	2.36	2.60	2.91	0.23
	Ours	**3.32**	**3.26**	**3.48**	**0.09**
Qwen2-7B	LDP	5.09	5.30	9.52	2.04
	Ours	**7.99**	**7.99**	**9.60**	**0.76**
Qwen2-2B	LDP	3.40	4.06	7.40	1.75
	Ours	**7.10**	**7.20**	**8.05**	**0.43**

3.2 Results

Overall Utility Comparison. Table 2 reports utility at per-token ε=8. We compare against LDP that adds uniform noise to all input embeddings as the primary baseline, as it is the only mechanism that satisfies our untrusted trainer threat model. To the best of our knowledge, no prior work has applied LDP to VLM fine-tuning; we therefore implement uniform embedding-level LDP using the same Gaussian mechanism and privacy budget as RadSLDP, applied to all input tokens rather than selectively. Across all 36 comparisons, RadSLDP outperforms LDP in 29 cases (80.5%) and matches its performance in 5 cases (14%), with improvements most pronounced on ReXGradient: 24–77% in BLEU-4 and 87–108% in Micro-F1 across all architectures. Notably, Qwen2-2B improves from 0.12 to 0.25 in Micro-F1 (+108%) and from 4.06 to 7.20 in BLEU-4 (+77%) on ReXGradient. On CheXpert Plus and MIMIC-CXR, RadSLDP consistently matches or exceeds LDP, indicating that selective perturbation preserves clinically relevant information that uniform noise can disrupt.

Stability Across Privacy Budgets. Table 3 examines the effect of varying $\varepsilon \in \{4, 8, \infty\}$ on ReXGradient. RadSLDP exhibits lower BLEU-4 standard deviation than LDP across all three architectures: LLaVA-7B (0.09 vs. 0.23),

Qwen2-7B (0.76 vs. 2.04), and Qwen2-2B (0.43 vs. 1.75). Because RadSLDP perturbs only NVCC tokens, changes in the noise scale affect a smaller portion of the input, producing more stable performance as ε varies. In contrast, uniform LDP applies noise to all tokens, amplifying the effect of any change in privacy budget across the entire input. Both methods show a general increasing trend as ε grows, consistent with prior work.

4 Conclusion

We presented RadSLDP, a selective embedding-level local differential privacy mechanism for privacy-preserving fine-tuning of medical VLMs in untrusted training settings. By perturbing only Non-Visual Clinical Context (NVCC) token embeddings while leaving image-inferable content intact, RadSLDP provides formal per-token (ε, δ)-LDP guarantees independent of the downstream model architecture. Across three chest X-ray datasets and three VLM architectures, RadSLDP matches or outperforms uniform LDP in 94% of comparisons (34/36), achieving up to 108% Micro-F1 improvement on ReXGradient with lower performance variability across privacy budgets due to perturbing fewer tokens. To support reproducibility, the source code and our sensitivity annotations comprising 347 expert-validated subcategories derived from 141,129 clinical indications are available at: https://github.com/Epiphany625/RadSLDP.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abadi, M., Chu, A., Goodfellow, I., et al.: Deep learning with differential privacy. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS). pp. 308–318 (2016)
2. Campello, R.J., Moulavi, D., Sander, J.: Density-based clustering based on hierarchical density estimates. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) (2013)
3. Chambon, P., Delbrouck, J.B., Sounack, T., et al.: Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats (2024)
4. Du, M., Yue, X., Chow, S.S.M., Sun, H.: Sanitizing sentence embeddings (and labels) for local differential privacy. In: Proceedings of the ACM Web Conference. pp. 2349–2359 (2023)
5. Duchi, J.C., Jordan, M.I., Wainwright, M.J.: Local privacy and statistical minimax rates. In: 2013 IEEE 54th Annual Symposium on Foundations of Computer Science. pp. 429–438 (2013)
6. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Theory of Cryptography Conference (TCC). pp. 265–284. Springer (2006)
7. Dwork, C., Roth, A.: The Algorithmic Foundations of Differential Privacy, vol. 9. Now Publishers (2014)

8. Flemings, J., Razaviyayn, M., Annavaram, M.: Differentially private next-token prediction of large language models. In: North American Chapter of the Association for Computational Linguistics (NAACL). pp. 4390–4404 (2024)
9. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing. ACM Trans. Comput. Healthcare **3**(1) (Oct 2021)
10. Hu, E.J., Shen, Y., Wallis, P., et al.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR) (2022)
11. Hu, L., et al.: Differentially private natural language models: Recent advances and future directions pp. 478–499 (Mar 2024)
12. Jain, S., Agrawal, A., Saporta, A., et al.: RadGraph: Extracting clinical entities and relations from radiology reports. In: Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track (2021)
13. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., et al.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data **6** (2019)
14. Kairouz, P., Oh, S., Viswanath, P.: The composition theorem for differential privacy. In: International Conference on Machine Learning (ICML). pp. 1376–1385. PMLR (2015)
15. Kaissis, G.A., Ziller, A., Passerat-Palmbach, J., et al.: End-to-end privacy preserving deep learning on multi-institutional medical imaging. Nature Machine Intelligence **3**, 473–484 (2021)
16. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81 (2004)
17. Liu, H., Li, C., Wu, Q., et al.: Visual instruction tuning. In: Advances in Neural Information Processing Systems (NeurIPS) (2024)
18. Lyu, L., He, X., Li, Y.: Differentially private representation for nlp: Formal guarantee and an empirical study on privacy and fairness. In: Findings of the Association for Computational Linguistics: EMNLP. pp. 2355–2365 (2020)
19. Mai, P., Yan, R., Huang, Z., et al.: Split-and-denoise: Protect large language model inference with local differential privacy. In: Proceedings of the 41st International Conference on Machine Learning (ICML). pp. 34431–34461 (2024)
20. Mironov, I.: Rényi differential privacy. In: IEEE Computer Security Foundations Symposium (CSF). pp. 263–275. IEEE (2017)
21. OpenAI: GPT (version 5.2). Large Language Model (2025), https://platform.openai.com/
22. Packhäuser, K., Gündel, S., Münster, N., et al.: Deep learning-based patient reidentification is able to exploit the biometric nature of medical chest x-ray data. Scientific Reports **12**, 14851 (2022)
23. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A method for automatic evaluation of machine translation. In: Annual Meeting of the Association for Computational Linguistics (ACL). pp. 311–318 (2002)
24. Shi, W., Cui, A., Li, E., et al.: Selective differential privacy for language modeling. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) (2022)
25. Shi, W., Shea, R., Chen, S., et al.: Just fine-tune twice: Selective differential privacy for large language models. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 6327–6340 (2022)
26. Smit, A., Jain, S., Rajpurkar, P., et al.: CheXbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In:

Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1500–1519 (2020)
27. Tran, L., Sun, W., Patterson, S., Milanova, A.: Privacy-preserving personalized federated prompt learning for multimodal large language models. In: The Thirteenth International Conference on Learning Representations (2025)
28. U.S. Department of Health and Human Services: Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) privacy rule (2012)
29. Wang, P., Bai, S., Tan, S., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
30. Wang, T., Zhai, L., Yang, T., et al.: Selective privacy-preserving framework for large language models fine-tuning. Information Sciences **682**, 121000 (2024)
31. Wei, Q., Li, J., You, Z., et al.: Dual-priv pruning: Efficient differential private fine-tuning in multimodal large language models. arXiv preprint arXiv:2506.07077 (2025)
32. Yu, D., Naik, S., Backurs, A., et al.: Differentially private fine-tuning of language models. arXiv preprint arXiv:2110.06500 (2021)
33. Yu, M., Singh, A., Li, S., Cao, Y.: Adaptive token-weighted differential privacy for LLMs: Not all tokens require equal protection. arXiv preprint arXiv:2509.23246 (2025)
34. Zhang, R.: Privacy-oriented text generation in LLMs via selective fine-tuning and semantic attention masks. Journal of Computer Technology and Software **4**(8) (2025)
35. Zhang, X., Acosta, J.N., Miller, J., et al.: Rexgradient-160k: A large-scale publicly available dataset of chest radiographs with free-text reports (2025)
36. Zheng, L., Cao, Y., Yoshikawa, M., et al.: Sensitivity-aware differential privacy for federated medical imaging. Sensors **25**(9), 2847 (2025)
37. Ziller, A., Usynin, D., Braren, R., et al.: Medical imaging deep learning with differential privacy. Scientific Reports **11**, 13524 (2021)