BIGGER IS NOT BETTER UNDER DIFFERENTIAL PRIVACY: OPTIMIZATION FAILURE AT ELEVEN-BILLION SCALE IN VISION–LANGUAGE MODEL FINE-TUNING

Tzuen Su
Data Science Degree Program
National Taiwan University
r13946017@ntu.edu.tw

Li-Hong Guo
Data Science Degree Program
National Taiwan University
r13946015@ntu.edu.tw

Yangmi Su
Department of Computer Science and Information Engineering
National Central University
113522114@cc.ncu.edu.tw

Cheng-Yen Li
Data Science Degree Program
National Taiwan University
r13946003@ntu.edu.tw

ABSTRACT

Differential privacy (DP) is an appealing safeguard for adapting instruction-tuned vision–language models (VLMs), but its scaling behavior under standard private fine-tuning remains unclear. We present a focused negative result for DP-SGD fine-tuning with LoRA on two widely used backbones: PaLI-GEMMA-3B-PT-224 and LLaMA-3.2-11B-Vision-Instruct. We target ($\varepsilon \in \{1, 10, 100, 1000\}, \delta = 10^{-5}$) by tuning only the DP-SGD noise scale while keeping the data, epochs, batch size, clipping norm, and all non-private hyperparameters fixed; ε is computed with standard DP accounting. Across all budgets, the 3B model trains stably and stays close to non-private validation performance, whereas the 11B model becomes optimization-limited, with flat, high loss and collapsed lexical-overlap metrics. BERTScore drops much less, revealing a lexical–semantic mismatch that can obscure residual utility under DP. Our findings suggest that scaling DP-SGD+LoRA to 11B may require additional stability interventions beyond simply relaxing ε.

1 INTRODUCTION

Vision–language models (VLMs) have advanced rapidly through large-scale multimodal pretraining and instruction tuning, enabling strong performance on captioning and visual question answering (VQA). Recent systems include scaled multimodal models such as Flamingo Alayrac et al. (2022) and PaLI Chen et al. (2022), instruction-tuned open models such as LLaVA Liu et al. (2023), and efficient designs such as PaLI-Gemma Beyer et al. (2024). Larger instruction-following VLMs (e.g., LLaMA-3.2-11B-Vision-Instruct Grattafiori et al. (2024); Meta AI (2024)) further improve multimodal reasoning, motivating adaptation to domain-specific tasks such as medical VQA.

Even when fine-tuning uses public medical benchmarks, privacy-preserving training remains relevant because large models can memorize training examples and enable extraction-style attacks Carlini et al. (2021). Differential privacy (DP) offers a principled safeguard, but its practical behavior for modern instruction-tuned VLMs—especially as model scale increases—remains unclear. DP-SGD introduces gradient clipping and noise that substantially alter optimization dynamics, raising the practical question of when private adaptation becomes optimization-limited as models grow.

In this work, we study DP-SGD fine-tuning with LoRA on two widely used backbones of contrasting scale: PaLI-GEMMA-3B-PT-224 Beyer et al. (2024) and LLaMA-3.2-11B-Vision-Instruct Grattafiori et al. (2024); Meta AI (2024). Under comparable privacy budgets, we observe a scale-dependent failure: the 3B model trains stably with consistent loss reduction and preserved semantic similarity, whereas scaling the same recipe to 11B leads to persistent optimization breakdown under DP. This breakdown is characterized by flat, high training loss and a collapse of lexical-

overlap metrics, while semantic similarity degrades substantially less, revealing a lexical–semantic mismatch. Notably, this failure persists even as the privacy budget is relaxed over several orders of magnitude. These results show that DP-SGD+LoRA recipes effective at medium scale can enter an optimization-limited regime at larger model sizes, with stability emerging as the primary bottleneck.

2 Background and Motivation

We adopt a standard parameter-efficient fine-tuning setup based on LoRA (Hu et al., 2022), where backbone parameters are frozen and only low-rank adapter weights are updated. Differential privacy is enforced using DP-SGD (Abadi et al., 2016) under the standard DP definition (Dwork et al., 2014; 2006): per-example gradients are clipped, perturbed with Gaussian noise, and composed with standard accounting to obtain an overall (ε, δ) guarantee.

Although our experiments use a public benchmark, DP remains relevant as a widely used safeguard when adapting large models, since memorization can enable extraction-style attacks. In practice, DP-SGD modifies optimization through clipping and noise, making training increasingly sensitive to hyperparameters and model scale. Rather than proposing a new privacy mechanism, we empirically study how model scale affects optimization stability and utility under this standard DP-SGD+LoRA fine-tuning recipe.

3 Method: DP-SGD Fine-tuning with LoRA

We adopt a standard differentially private fine-tuning setup based on DP-SGD, applied only to LoRA adapter parameters. The pretrained vision and language backbone weights remain frozen throughout training, and only the low-rank LoRA matrices are updated. At each optimization step, we (i) compute per-example gradients with respect to the adapter parameters, (ii) clip each gradient to a fixed ℓ_2 norm bound C, (iii) average the clipped gradients over the minibatch, and (iv) add Gaussian noise with scale σC before applying the optimizer update. This procedure follows the conventional DP-SGD formulation, and privacy loss is tracked using standard accounting to report an overall (ε, δ) guarantee. We do not modify the DP mechanism or the LoRA architecture beyond this standard recipe. The full pseudocode is provided in Appendix A.

4 Dataset and Evaluation

We evaluate on the publicly available VQA-Med 2019 benchmark Ben Abacha et al. (2019). All models are fine-tuned using only the official training split, and performance is reported exclusively on the held-out validation split to measure generalization under privacy constraints. The same data preprocessing, training protocol, and evaluation procedure are used for both model scales.

We report standard generation metrics commonly used in medical visual question answering: BLEU Papineni et al. (2002), ROUGE (ROUGE-1/2/L variants) Lin (2004), and BERTScore (F1) Zhang et al. (2019). BLEU and ROUGE capture surface-level n-gram overlap, while BERTScore measures semantic similarity using contextual embeddings. Metrics are computed between each generated answer and its ground-truth reference on the validation set and then averaged across all validation examples.

5 Experimental Setup

We fine-tune and evaluate on the public VQA-Med 2019 benchmark (3,200 radiology images; 12,792 training QA pairs; 500 validation triplets). We train on the official training split and report results on the held-out validation split. Images are resized to 224×224 and normalized for the vision encoder. Each example is an (image, question, answer) triplet, and the model is trained to autoregressively generate the answer conditioned on the image and question.

We use LoRA for parameter-efficient adaptation, updating only adapter parameters while freezing the full vision and language backbones. Differential privacy is enforced with DP-SGD on LoRA parameters: per-example gradients are clipped to C and perturbed with Gaussian noise of scale σC before the optimizer update.

All non-private settings are fixed and only the noise scale is varied to obtain different privacy budgets (ε, δ). Basic stabilization attempts for the 11B model did not recover convergence. Unless noted otherwise, we use B=8, C=10, γ=5×10^{-4}, r=64, 2 epochs, and AdamW with mixed bfloat16 precision. Experiments run on NVIDIA GPUs (45 GB VRAM) using PyTorch, Transformers, PEFT, and BitsAndBytes; hyperparameters are summarized in Table 4.

6 Performance Evaluation

We compare DP-SGD fine-tuning on two vision–language model backbones LLaMA-3.2-11B-Vision-Instruct and PaLI-GEMMA-3B-PT-224 under privacy budgets ($\varepsilon \in \{1, 10, 100, 1000\}, \delta = 10^{-5}$) and a non-private baseline. All models are trained on the official VQA-Med 2019 training split and evaluated on the held-out validation set. We report BLEU and ROUGE (lexical overlap) together with BERTScore (semantic similarity), since DP training can alter answer verbosity and cause surface-form metrics to understate semantic fidelity. Quantitative results are shown in Tables 1, 2, and 3, and optimization behavior is illustrated via training loss curves (Figures 1 and 2).

Table 1: Validation results on VQA-Med 2019 for LLaMA-3.2-11B under different privacy budgets. BLEU: n-gram overlap; RL: ROUGE-L; R2: ROUGE-2; BP: BERTScore Precision; BR: BERTScore Recall; F1: BERTScore F1.

Privacy	**BLEU**	**RL**	**R2**	**BP**	**BR**	**F1**
$\varepsilon = 1$	0.0081	0.0661	0.0019	0.8145	0.8052	0.8093
$\varepsilon = 10$	0.0082	0.0671	0.0023	0.8144	0.8052	0.8093
$\varepsilon = 100$	0.0083	0.0673	0.0019	0.8148	0.8056	0.8096
$\varepsilon = 1000$	0.0086	0.0709	0.0019	0.8147	0.8062	0.8099
Full	0.4980	0.5962	0.2010	0.9368	0.9301	0.9330

Table 2: Validation results on VQA-Med 2019 for PaLI-GEMMA-3B under different privacy budgets. BLEU: n-gram overlap; RL: ROUGE-L; R2: ROUGE-2; BP: BERTScore Precision; BR: BERTScore Recall; F1: BERTScore F1.

Privacy	**BLEU**	**RL**	**R2**	**BP**	**BR**	**F1**
$\varepsilon = 1$	0.1140	0.1444	0.0000	0.9049	0.8346	0.8673
$\varepsilon = 10$	0.1160	0.1464	0.0000	0.9052	0.8352	0.8677
$\varepsilon = 100$	0.1180	0.1478	0.0000	0.9055	0.8352	0.8679
$\varepsilon = 1000$	0.1180	0.1489	0.0000	0.9056	0.8353	0.8680
Full	0.1160	0.1458	0.0000	0.9054	0.8348	0.8676

Table 3: Summary comparison of main metrics for strongest ($\varepsilon = 1$), weakest ($\varepsilon = 1000$) DP, and full training. BLEU: n-gram overlap; RL: ROUGE-L; BP: BERTScore Precision; F1: BERTScore F1.

	LLaMA-3.2-11B				**PaLI-GEMMA-3B**			
Privacy	BLEU	RL	BP	F1	BLEU	RL	BP	F1
DP ($\varepsilon = 1$)	0.0081	0.0661	0.8145	0.8093	0.1140	0.1444	0.9049	0.8673
DP ($\varepsilon = 1000$)	0.0086	0.0709	0.8147	0.8099	0.1180	0.1489	0.9056	0.8680
Full	0.4980	0.5962	0.9368	0.9330	0.1160	0.1458	0.9054	0.8676

6.1 Results on Medical Tasks

A strong scale-dependent effect emerges. For LLaMA-3.2-11B (Table 1), DP-SGD leads to a collapse in lexical metrics: BLEU drops from 0.4980 (Full) to 0.0081 at $\varepsilon = 1$, and ROUGE-L from 0.5962 to 0.0661. However, semantic similarity degrades much less, with BERTScore F1 remaining 0.8093 at $\varepsilon = 1$ (vs. 0.9330 Full) and changing only marginally as privacy is relaxed. This lexical–semantic divergence suggests that DP noise disrupts token-level surface-form optimization

more severely than higher-level semantic representations, leading to degraded lexical overlap despite partially preserved semantic alignment.

By contrast, PaLI-GEMMA-3B is highly robust under DP (Table 2). Across all tested budgets, metrics remain nearly identical to the non-private baseline. For example, at $\varepsilon = 1$, BLEU is 0.1140 vs. 0.1160 (Full) and BERTScore F1 is 0.8673 vs. 0.8676; even at $\varepsilon = 1000$, performance remains effectively unchanged. As summarized in Table 3, although the larger model achieves higher absolute scores without privacy constraints, only the medium-sized model preserves its performance under DP across the entire privacy range.

6.2 Training Loss Curves Analysis

The loss curves mirror the metric-level findings. PaLI-GEMMA-3B shows stable training under both DP and non-DP settings (Figure 1), with nearly overlapping trajectories across privacy budgets, indicating that DP noise does not destabilize optimization at this scale. In contrast, LLaMA-3.2-11B exhibits high, flat loss under DP for all tested $\varepsilon \leq 1000$ (Figure 2), while the same LoRA configuration converges normally in the non-private run. These trends indicate that, under adapter-only DP-SGD, optimization enters a noise-dominated regime at larger model scales, where clipping and noise overwhelm useful gradient signal, and relaxing ε within this range does not restore stable training for the 11B model.

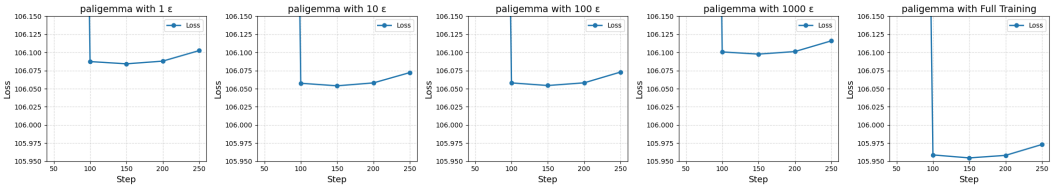


Figure 1: Training loss curves for PaLI-GEMMA-3B on VQA-Med 2019 under different privacy budgets ($\varepsilon = 1, 10, 100, 1000$) and full (non-private) training. The loss remains stable across all privacy levels, demonstrating the robustness of medium-sized models to differential privacy noise.

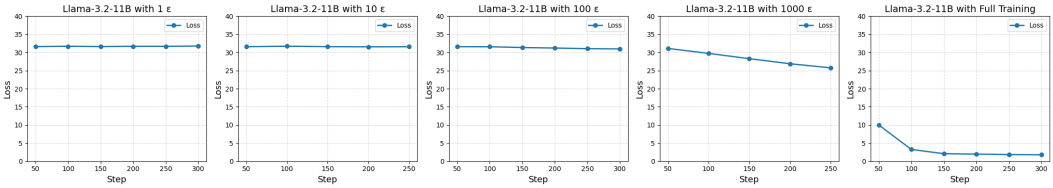


Figure 2: Training loss curves for LLaMA-3.2-11B-Vision-Instruct on VQA-Med 2019 under different privacy budgets ($\varepsilon = 1, 10, 100, 1000$) and full (non-private) training. The loss is high and flat under strong privacy, indicating limited learning, and only drops substantially in the non-private setting.

7 Conclusion

We studied differentially private adapter-only fine-tuning (LoRA + DP-SGD) for medical VLM tasks across privacy budgets ($\varepsilon \in \{1, 10, 100, 1000\}, \delta = 10^{-5}$). PaLI-GEMMA-3B maintains near-identical validation performance to non-private training and exhibits stable loss curves across all budgets. In contrast, LLaMA-3.2-11B shows severe degradation under DP: lexical metrics collapse while BERTScore remains relatively high, revealing a lexical–semantic mismatch, and loss curves indicate failed optimization for all tested ε values.

These results reveal a scale-dependent effect under differential privacy. While a 3B model remains trainable across privacy budgets, scaling the same DP-SGD+LoRA recipe to 11B shifts training into

an optimization-limited regime where useful gradient signal is dominated by clipping and noise. Relaxing the privacy budget over three orders of magnitude does not recover convergence, indicating that larger instruction-tuned models can be disproportionately fragile to DP-induced perturbations.

Limitations. A plausible contributing factor is that fixed low-rank adapters become a stronger capacity bottleneck as backbone scale increases, and DP noise further amplifies this underfitting effect. Understanding how scale, adapter capacity, and DP-induced noise jointly shape optimization stability is a key open problem for private fine-tuning at scale. We note that the two backbones compared here differ not only in scale but also in architecture and pretraining, so the observed behavior should be interpreted as a case study rather than a universal scaling law.

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*, pp. 308–318, 2016.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*. 9-12 September 2019, 2019.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pp. 265–284. Springer, 2006.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Meta AI. Llama 3.2-11b vision-instruct. `https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct`, 2024.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

A Pseudocode for DP-SGD Fine-Tuning with LoRA

Algorithm 1 DP-SGD Fine-Tuning with LoRA Adapters (Adapter-Only Updates)

Require: Dataset D, steps T, batch size B, learning rate γ, clipping norm C, noise scale σ, frozen backbone θ, LoRA params ϕ

for $t = 1$ to T **do**
  Sample minibatch $\{(x_i, y_i)\}_{i=1}^{B}$ from D
  for $i = 1$ to B **do**
    Compute per-example gradients $g_i = \nabla_\phi \mathcal{L}(x_i, y_i; \theta, \phi)$
    Clip $g_i \leftarrow g_i / \max(1, \|g_i\|_2 / C)$
  end for
  $\bar{g} \leftarrow \frac{1}{B} \sum_{i=1}^{B} g_i$
  $\tilde{g} \leftarrow \bar{g} + \mathcal{N}(0, \sigma^2 C^2 I)$
  Update adapters $\phi \leftarrow \phi - \gamma\, \tilde{g}$
end for
return ϕ

B Training Hyperparameters

Table 4 lists the training hyperparameters used across all experiments. All settings are kept fixed across privacy budgets *except* the DP-SGD noise scale σ, which is adjusted to obtain different target privacy levels. The reported privacy budgets (ε, δ) are computed using standard DP accounting based on the chosen noise scale, dataset size, batch size, and number of optimization steps.

Table 4: Training hyperparameters for DP-SGD fine-tuning with LoRA adapters. The noise scale σ is varied across experiments to achieve different privacy budgets.

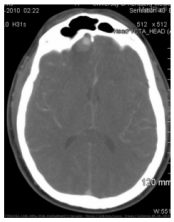
Hyperparameter	Value
Batch size B	8
Noise scale σ	Varied (per privacy budget)
Learning rate γ	5×10^{-4}
Clipping bound C	10
Number of training epochs	2
LoRA rank r	64
Optimizer	AdamW
Precision	bfloat16 (mixed)

C Qualitative Examples from the VQA-Med Benchmark

This appendix provides qualitative examples from the VQA-Med benchmark to complement the quantitative results in the main text. We show outputs from both LLaMA-3.2-11B-Vision-Instruct and PaLI-GEMMA-3B under a strong privacy setting ($\varepsilon = 1$) and the non-private (full training) setting. Each example includes the biomedical image, the user question, the ground-truth answer, and the corresponding model responses.

Table 5 presents a modality identification query for LLaMA-3.2-11B. The ground truth is "cta – ct angiography." Under DP ($\varepsilon = 1$), the model produces a longer, partially correct response that

Visual input example, Biomedical image:

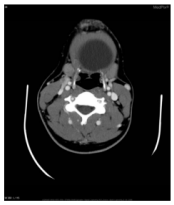


User	what modality is shown?
Ground Truth	cta - ct angiography
$\varepsilon = 1$	The image provided is a computed tomography (CT) scan of the head, as indicated by the presence of a cross-sectional view of the brain and the use of a CT scanner.
User	what modality is shown?
Ground Truth	cta - ct angiography
Full Training	ct noncontrast

Table 5: Example of medical vision-language question answering for LLaMA-3.2-11B at $\varepsilon = 1$ (DP) and full training.

identifies the scan as CT but does not specify angiography. Under full training, the output is concise but incorrect ("ct noncontrast"). This example illustrates how privacy-induced optimization noise can affect answer specificity and phrasing, and that more concise outputs are not necessarily more accurate.

Visual input example, Biomedical image:



User	is this a ct scan?
Ground Truth	yes
$\varepsilon = 1$	yes
User	is this a ct scan?
Ground Truth	yes
Full Training	yes

Table 6: Example of medical vision-language question answering for PaLI-GEMMA-3B at $\varepsilon = 1$ (DP) and full training.

Table 6 shows a binary classification query for PaLI-GEMMA-3B ("is this a ct scan?"). Both the DP ($\varepsilon = 1$) and full-training variants answer correctly ("yes"), consistent with the quantitative finding that PaLI-GEMMA-3B remains stable under strong privacy constraints. Overall, these examples

qualitatively reinforce the main conclusion: model scale and output style can interact with differential privacy in non-trivial ways, affecting answer specificity and surface form even when the core semantic intent is preserved.