Effect of Pretrained Initialisation and Component-Wise Tuning on Lightweight VLMs for Ultrasound Report Generation

Sharon Mlotshwa[1[0000−0001−7101−0618]], Pramit Saha[3[0009−0006−8090−1969]], Jamal Bentahar[1[0000−0002−3136−4849]], Muzammal Naseer[1[0000−0001−7663−7161]], and Mohammad Alsharid[1,2⋆[0000−0002−2271−0578]]

[1] Department of Computer Science, Khalifa University, Abu Dhabi, UAE
[2] Advanced Research and Innovation Center, Khalifa University, Abu Dhabi, UAE
[3] Department of Engineering Science, University of Oxford, Oxford, UK
mohammad.alsharid@ku.ac.ae

Abstract. Ultrasound report generation (US-RG) can ease sonographers' documentation burden, but it is a challenging task. Practical clinical deployment requires US-RG lightweight models, but progress is constrained by the scarcity of open-source ultrasound image-text datasets. Under these low-data and resource constrained conditions, we study how the choice of a lightweight US-RG model's pretrained initialisation and component-level (vision/language model) finetuning affects how it performs on multi-anatomy datasets. Our findings suggest that: (i) domain-specific ultrasound pretraining may not be necessary for lightweight US-RG as general biomedical initialisation and an initialisation from web-scale pretraining perform comparably; (ii) given the limited data, scaling the model yields little gains, whereas parameter-efficient adaptation of both the vision and language model is more promising; (iii) task-specific finetuning irrespective of the initialisation source weights outperforms zero-shot evaluation and (iv) lightweight models suitable for on-device deployment can be used for US-RG, but the gap between the semantic similarity and accurate identification of clinically relevant findings shows the need for methods that improve clinical factuality.

Keywords: Ultrasound · Lightweight models · Transfer learning.

1 Introduction

Ultrasound (US) is used in patient evaluation globally, but translating an exam into a report is a challenge. Studies [1,4,18] suggest that US report writing is an obstacle for sonographers, who frequently report difficulties in effectively communicating their findings. Automated report writing may reduce the burden of clinical documentation. Le et al. [8] shows that US-RG is one of the most difficult ultrasound understanding tasks when evaluating 23 closed, open, and medical-specific vision language models (VLMs), including GPT-4o, MedGemma-4B-it,

⋆ Corresponding author.

and Qwen-2.5-VL-72B-Instruct, under zero-shot settings. The best-performing model achieved less than 7.5% BLEU-4. They observed that performance does not consistently improve with increasing model size, as smaller architectures can at times outperform larger ones, suggesting that targeted training may be more beneficial than simply increasing scale, but this hypothesis was not evaluated experimentally. This finding leaves the operative question for model deployment unanswered: if scale is not the primary driver, then how capable can a small, domain-adapted VLM become at US-RG, and which factors most strongly determine its performance? *Does specificity help monotonically, and at what level of granularity does domain knowledge provide measurable benefit?* This question is addressed in this work, motivated by its practical relevance in resource constrained settings (RCS), where bedside point-of-care ultrasound (POCUS) systems commonly rely on smartphones or tablets with at least 4GB of RAM and typically up to 12GB [17], thereby requiring lightweight VLMs.

Motivated by the findings of Le et al. [8], we investigate how task-specific finetuning (targeted training) can be more effective than increasing model scale by isolating the effects of ultrasound-specific, general biomedical, and large-scale web knowledge priors. Specifically, we study the effect of initialisation source and component-level (vision/language model) tuning for US-RG in a sub-billion parameter VLM, evaluated on clinical metrics and natural language generation (NLG) metrics. Our findings suggest that (i) domain-specificity (ultrasound-specific pretraining) may not be necessary for model performance improvement in lightweight US-RG models; (ii) given the limited US dataset size, increasing model scale, does not guarantee performance gains, rather improving the parameter-efficient adaptation method on both the vision model and language model (LM) may be more promising; (iii) task-specific finetuning irrespective of the initialisation source weights outperforms zero-shot evaluation; and (iv) lightweight models are suitable for on-device deployment, but the gap between semantic similarity evaluation scores and accurate identification of clinically relevant findings shows the need for methods that improve clinical factuality.

2 Previous Work

Some early text generating methods for US proposed CNN/transformer encoder-decoders with knowledge distillation [9], CLIP-GPT [25], GAN (generator, discriminator, and clinical-style evaluator) [24], and curriculum learning [2]. Recent work scales to multimodal LLMs and mixtures-of-experts at 3-11B parameters [11,19,22]. However, they mostly do not study the direct relevance of edge-eployment. Prior studies have examined transfer learning and initialisation in the context of representation learning, classification, segmentation, and visual question answering [3,7,12,20,23]. Although these works show the importance of domain-adaptive pretraining and efficient adaptation strategies for discriminative tasks, our contribution is testing whether this survives in generative US-RG. Li et al. [10], surveys VLM adaptation and generalisation methods with the discussion largely centered on natural computer vision rather than medical imaging.

Ma et al. [15] proposes a compact foundational model for diagnostic classification and anatomical segmentation across 15 US organs, while Saeed et al. [21] developed a distillation framework to compress a fetal US foundation model for mobile deployment. However, both studies focus on classification, image-text similarity, and retrieval tasks rather than generating free-form diagnostic reports.

3 Methodology

Our objective is to investigate how a lightweight VLM's specialisation (its pretraining data), influences its performance on US-RG under RCS. We formulate the study as a controlled empirical evaluation over four factors: (i) initialisation source, (ii) component-wise finetuning strategy, (iii) model scale, and (iv) target anatomy. We denote the initialisation source by $I \in \mathcal{I}$, where $\mathcal{I} = \{U, G, W, R\}$, here U denotes ultrasound-specific pretraining, G denotes general biomedical pretraining, W denotes initialisation from large-scale web pretraining, and R denotes random initialisation. The component-wise finetuning strategy $S \in \mathcal{S}$ specifies whether the vision encoder and LM are fully finetuned, adapted using parameter-efficient finetuning (LoRA), or kept frozen, with $\mathcal{I}$, as the model's starting weights. Model scale is varied over 256M, 500M and 2.2B parameters, while the evaluation anatomies are based on $A \in \mathcal{A}$, where $\mathcal{A} = \{\text{liver}, \text{thyroid}, \text{breast}, \text{combined}\}$ and the combined setting contains data from all three anatomies. We focus our main experiments ($\mathcal{E}_{\text{main}}$) based on,

$$\mathcal{E}_{\text{main}} = \{M_1 \times \{U, G, W, R\},\ M_2 \times \{U, G, W\},\ M_3 \times \{U, G, W\}\}.$$

Here, M_1 corresponds to full finetuning of both vision and language components, M_2 to LoRA finetuning of both components, and M_3 to full finetuning of the vision encoder with LoRA adaptation of the LM.

Task Formulation. Given a US image $\mathbf{x}$ and a fixed instruction prompt p, the objective of US-RG is to generate a clinical report $\mathbf{y}$. We formulate this as a supervised finetuning problem, where the model is trained using the autoregressive next-token cross-entropy objective. To study the factors influencing model performance, each experimental configuration is defined by the tuple (I, S, A).

Lightweight Model. We use the SmolVLM [16] family of models for all experiments due to its emphasis on efficiency under RCS. SmolVLM is available in three model sizes (256M, 500M, and 2.2B parameters). Our primary experiments focus on the 256M variant, as its compact footprint makes it well suited for deployment on handheld POCUS systems, requiring less than 1GB of GPU memory during inference for the model and occupying under 600MB of disk storage. The larger 500M and 2.2B variants are included to assess the effect of model scale. Architecturally, the 256M model combines a 93M SigLIP-B/16 vision encoder with the 135M SmolLM2 LM, and the MLP projection.

Model Initialisation. We compare four model initialisations that differ in their pretraining data. The general biomedical initialisation (G) uses BIOMEDICA [13]'s SmolVLM-256M checkpoint[4], which is pretrained on more than 24

[4] https://huggingface.co/BIOMEDICA/BMC-smolvlm1-256M

million image-caption pairs extracted from 6 million articles in the PubMed Central Open Access subset. The dataset spans medical imaging modalities, including X-ray, CT, electrocardiography, microscopy, and others, making it a general-purpose biomedical initialisation. Our ultrasound-specific initialisation (U) is based on the US-365K dataset [6], which contains about 365 thousand ultrasound image–text pairs covering 52 anatomical regions. As we could not find any publicly available SmolVLM-256M checkpoint pretrained exclusively on US data, we performed this stage of pretraining using the US-365K dataset and refer to the resulting checkpoint as **SmolUS**. We publicly release this checkpoint for broader research use. The web-scale initialisation (W) corresponds to the original SmolVLM-256M [16] checkpoint, which is pretrained on diverse web-scale datasets spanning image captioning, document understanding, chart interpretation, reasoning, and mathematical tasks. Finally, the random initialisation (R) uses randomly initialized weights (without any pretraining).

Experimental Set-up. We conduct our study on the Chinese US Report Dataset [9], using about 3,513 breast, 2,458 thyroid, and 1,394 liver patient-level datasets. To the best of our knowledge, it is currently the only fully open sourced task-specific dataset for US-RG. Since the reports are originally written in Chinese, we use the English translation[5] by Lyu et al. [14]. We perform patient-level splitting using a 7:1:2 ratio for training, validation, and testing. We report BLEU 1–4, ROUGE-L, and BERTScore-F1 (for semantic similarity). To evaluate clinical performance, we use the clinical-efficacy (CE) metrics introduced by Li et al. [9], which assess the presence of clinically relevant findings rather than textual similarity. Each predefined clinical keyword is assigned a binary label indicating whether it appears in the report, this converts the evaluation to a multi-label classification task, enabling the calculation of accuracy, precision, recall, and F1-score. We use the same clinical keywords as in [9]; CE-accuracy measures exact clinical keyword set matching. All experiments are conducted on Google Colab using NVIDIA L4 GPU (24GB VRAM). Our SmolUS is the only exception, which is pretrained on Google Colab NVIDIA A100 GPU (40GB VRAM). Our code and and model checkpoints are available at this repository.

4 Results

We present our main US-RG performance results in Table 1, evaluated across all anatomies (breast, thyroid, liver, and the combined setting) and all model initialisation strategies *G, W, U* and *R*. Table 1 and Figure 1 show that on average *G, W* and *U* initialisations achieved comparable results across the NLG and the clinical metrics, and no single initialisation consistently performed best across all the anatomies. On average (Figure 1, left), BERTScore-F1, ROUGE-L, BLEU-4, CE-F1 across the initialisations *G, W* and *U* mostly differ within 1% difference, and by 2.6% on CE-accuracy; with *U* attaining 31.5. In contrast, in Table 1, the randomly initialized model (R) under M_1 consistently underperformed all pretrained counterparts. For example, on the breast dataset, its CE-accuracy

[5] https://github.com/TongxinWang0405/BMI702_project

Table 1: US-RG results when finetuning across all anatomies and initialisations *G, W, U* for M_1, M_2 and M_3, and additionally R for M_1. *Config.*= configuration.

		NLG metrics ↑						**CE metrics ↑**			
	Config.	BLEU–1	BLEU–2	BLEU–3	BLEU–4	ROUGE-L	BERTScore	Accuracy	Precision	Recall	F1-Score
Breast	M_1G	53.37	45.22	39.66	35.51	50.59	85.09	19.27	79.69	78.69	79.19
	M_1W	53.82	45.80	40.36	36.27	51.43	85.41	18.35	79.05	80.13	79.59
	M_1U	51.51	43.61	38.26	34.28	50.00	84.37	19.42	78.96	76.14	77.52
	M_1R	39.35	30.93	25.03	20.60	35.74	77.57	0.28	82.55	52.32	64.05
	M_2G	58.53	52.65	48.57	45.50	62.27	88.84	44.95	85.34	87.81	86.56
	M_2W	58.01	52.12	47.99	44.86	61.54	88.69	43.95	85.39	87.14	86.26
	M_2U	**59.37**	**53.16**	**48.83**	**45.54**	61.25	88.44	42.03	84.25	87.11	85.66
	M_3G	58.25	52.36	48.25	45.15	**62.51**	**89.09**	46.44	**86.26**	87.75	87.00
	M_3W	57.80	52.01	47.96	44.91	62.27	89.03	**47.30**	86.22	**88.28**	**87.24**
	M_3U	57.10	50.05	45.17	41.51	55.72	86.93	29.37	79.83	84.79	82.24
Thyroid	M_1G	57.00	49.66	43.92	39.44	57.59	87.27	11.91	82.30	85.22	83.74
	M_1W	58.63	51.14	45.28	40.69	58.10	87.48	10.59	82.05	85.92	83.94
	M_1U	58.23	50.64	44.66	40.00	57.53	87.35	9.47	81.17	86.41	83.71
	M_1R	49.49	40.75	33.92	28.57	44.97	81.30	1.02	77.91	76.30	77.10
	M_2G	64.13	55.65	49.40	44.49	59.50	87.53	22.00	82.17	91.30	86.49
	M_2W	**64.45**	**56.14**	**50.03**	**45.23**	**60.87**	87.77	**23.01**	**82.97**	90.67	**86.65**
	M_2U	61.92	54.12	48.27	43.66	59.55	87.64	16.09	81.79	89.15	85.32
	M_3G	60.09	52.60	46.84	42.30	58.92	**87.84**	14.56	81.29	89.09	85.01
	M_3W	59.54	51.56	45.73	41.10	56.87	86.03	17.11	80.59	90.67	85.34
	M_3U	58.48	50.29	44.37	39.73	56.94	85.61	19.04	79.83	**92.62**	85.75
Liver	M_1G	40.58	28.83	21.14	16.25	37.59	80.75	15.59	74.47	71.98	73.20
	M_1W	41.56	30.12	22.49	17.64	37.39	81.40	10.75	70.33	72.19	71.25
	M_1U	43.12	32.02	24.61	19.60	39.48	82.08	13.44	71.09	**77.38**	74.11
	M_1R	20.96	14.41	9.89	7.19	24.83	68.22	5.02	**99.11**	29.61	45.60
	M_2G	47.12	38.13	32.48	28.95	48.17	85.92	33.33	77.26	77.01	**77.14**
	M_2W	**55.60**	**47.52**	**42.48**	**39.37**	**55.39**	**87.52**	38.71	79.15	73.57	76.26
	M_2U	49.28	40.60	35.16	31.73	50.32	86.34	34.41	78.28	74.26	76.22
	M_3G	51.09	43.47	38.79	35.96	53.86	87.19	39.61	80.96	72.30	76.39
	M_3W	50.07	41.66	36.45	33.26	51.52	86.64	38.71	80.07	72.78	76.25
	M_3U	52.18	44.66	40.03	37.21	55.14	87.47	**40.14**	80.90	72.03	76.21
Combined	M_1G	48.95	41.28	35.86	31.85	49.36	84.40	32.79	71.02	68.49	69.73
	M_1W	52.36	44.97	39.71	35.79	53.57	86.18	42.57	73.69	70.43	72.02
	M_1U	47.32	40.30	35.34	31.65	51.27	85.30	34.52	69.33	66.12	67.69
	M_1R	36.77	27.36	20.94	16.39	27.30	74.50	7.94	54.35	22.36	31.69
	M_2G	58.16	52.21	47.92	44.68	63.92	89.55	60.15	83.89	77.80	80.73
	M_2W	**59.19**	**53.11**	**48.73**	**45.44**	64.03	89.51	61.47	84.56	**78.34**	81.34
	M_2U	58.90	52.74	48.29	44.94	63.21	89.18	58.83	82.38	78.02	80.14
	M_3G	58.14	52.33	48.12	44.95	**64.41**	**89.72**	**61.74**	**85.10**	78.31	**81.57**
	M_3W	57.03	50.43	45.75	42.23	60.43	88.47	56.55	80.88	76.25	78.50
	M_3U	57.91	52.00	47.69	44.45	63.47	89.37	60.66	84.72	78.04	81.24
Run-to-run variability — mean ± std over 3 seeds											
Thyroid	M_1G	57.66 $\pm$0.92	50.09 $\pm$0.66	44.22 $\pm$0.49	39.66 $\pm$0.37	57.36 $\pm$0.21	87.16 $\pm$0.10	10.93 $\pm$1.14	81.87 $\pm$0.39	85.54 $\pm$0.93	83.66 $\pm$0.37
	M_2G	63.51 $\pm$0.58	55.19 $\pm$0.45	49.03 $\pm$0.37	44.18 $\pm$0.30	59.41 $\pm$0.08	87.50 $\pm$0.10	20.06 $\pm$1.69	81.96 $\pm$0.23	90.48 $\pm$0.87	86.01 $\pm$0.45
	M_3G	60.75 $\pm$1.63	53.11 $\pm$1.06	47.39 $\pm$0.82	42.91 $\pm$0.72	59.99 $\pm$1.20	88.02 $\pm$0.49	19.48 $\pm$4.42	83.05 $\pm$1.60	88.91 $\pm$0.77	85.87 $\pm$0.75

was only 0.28, compared with 18.35-19.42 across the pretrained initialisations. Averaged across all anatomies, the LoRA-based strategies (M_2 and M_3) outperformed full finetuning (M_1), while performing similarly to each other (Figure 1, right). For example, the average CE-accuracy increased from 19.9 for M_1 to 39.9 and 39.3 for M_2 and M_3, respectively. In Table 1, CE-accuracy spanned the widest range across all datasets, from 0.28 on the breast dataset under M_1 to 61.74, on the combined anatomy dataset under M_3. Across anatomies, liver achieved the lowest NLG performance, thyroid the lowest CE-accuracy (23.01), while the combined setting achieved the highest CE-accuracy (61.74). To keep the experimental budget manageable, we ran a subset of the experiments 3 times on the thyroid dataset over M_1, M_2 and M_3 as shown in Table 1, and we observe that the run-to-run variability of the results is mostly less than 1.7%, only

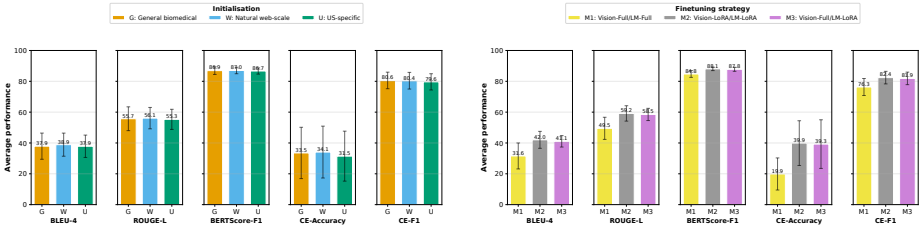

Fig. 1: *(Left)* Comparison of using initialisation G, W and U averaged over M_1, M_2 and M_3 and all anatomy settings. *(Right)* Comparison of finetuning strategy M_1, M_2 and M_3 averaged over G, W and U and all anatomy settings.

Table 2: Example of qualitative results with clinical keywords highlighted.

Liver Image	Reference	Generated (hit / miss)	Scores
	The shape of the liver was slightly plump, the capsule was smooth, the echo of the parenchyma was finely enhanced, the echo of the liver and kidneys was contrast-enhanced, and the portal vein system was not clearly displayed.	The liver was plump in shape, with a smooth capsule, dense and enhanced echoes in the parenchyma, unclear display of the portal venous system, and enhanced contrast in the echoes of the liver and kidneys. Missed: vein	**BLEU–4:** 9.69 **ROUGE-L:** 29.91 **BERTScore:** 83.68 **CE-Acc:** 0.0 **CE-F1:** 85.71

the CE-accuracy shows the highest variability of 4.42 for M_3G. A sample of the qualitative results is shown in Table 2. Table 3 shows the 256M model achieved the highest CE-accuracy across most anatomy settings, including a margin of over 5% and 8% over the 500M and 2.2B, respectively, on the thyroid dataset. Only in the combined anatomy setting did the 2.2B model perform better by just over 2%. For 256M, we measure the CPU inference latency on the Google Colab's free tier as a lightweight proxy. End-to-end generation ranges from 9s (44 tokens) to 15s (121 tokens), which is ∼5-8tokens/s. We report this only as a relative, platform-inclusive proxy.

5 Discussion

Effect of increasing domain specificity. Figure 1 shows that across the three finetuning methods M_1, M_2 and M_3, the choice of initialisation has relatively low

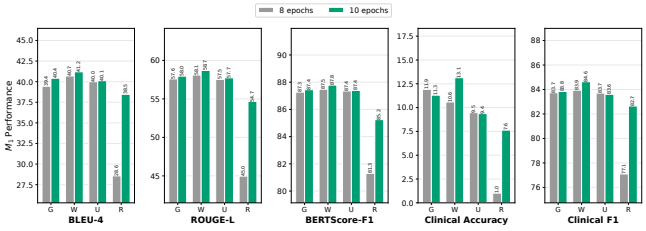

Fig. 2: **Ablation study**: Comparing model performance over G, W, U and R with extended training. Models use M_1 and are evaluated on the thyroid dataset.

Table 3: Model scale performance variation under M_2G. Top row per anatomy is the 256M result (M_2G in Table 1); subsequent rows give the change (Δ) at 500M and 2.2B. Negative red values mean scaling reduces performance.

Anatomy	Scale	BLEU-4	ROUGE-L	BERTScore	CE-Acc.	CE-F1
Breast	256M	45.50	62.27	88.84	44.95	86.56
	Δ500M	+0.33	−1.32	−0.38	−3.06	−0.58
	Δ2.2B	−0.03	+0.07	+0.12	−0.50	−0.02
Thyroid	256M	44.49	59.50	87.53	22.00	86.49
	Δ500M	−3.18	+0.23	+0.51	−5.40	−2.42
	Δ2.2B	−1.36	+0.08	+0.38	−8.55	−1.64
Liver	256M	28.95	48.17	85.92	33.33	77.14
	Δ500M	+2.19	+0.23	−0.03	−2.15	−0.80
	Δ2.2B	−5.29	−5.25	−1.59	−6.63	−2.74
Combined	256M	44.68	63.92	89.55	60.15	80.73
	Δ500M	+0.09	−0.27	−0.38	−0.27	+0.08
	Δ2.2B	−0.18	+0.64	+0.18	+2.07	+1.03

influence on US-RG performance. This could be likely due to ultrasound-specific pretraining mainly benefiting the visual feature learning, whereas the US-RG task is bottlenecked by the clinical text which is highly structured, repetitive and conforms to templated reports. Therefore, once the model is finetuned on in-domain clinical reports, much of this language-specific knowledge is learned regardless of the visual knowledge priors, and all models appear to converge to similar solutions. From a practical perspective, this suggests that readily available biomedical checkpoints or natural web-scale checkpoints may offer a competitive alternative when the ultrasound-specific pretrained models are unavailable. This may be beneficial in ultrasound research since open sourced data is very scarce. Random initialisation performs worse than the pretrained alternatives; however, Fig. 2, shows that additional training significantly improves the randomly initialized model, implying that much of the gap may arise from slower model convergence rather than insufficient learning capability. Therefore, given enough training steps, the model trained from scratch partly catches up, but pretrained initialisation converge much faster. This observation is consistent with He et al. [5]'s finding on CNNs when pretrained using ImageNet.

Effect of component-level adaptation strategies. Across M_1, M_2 and M_3, we observe that the adaptation strategy has a larger influence on the US-RG performance compared to the initialisation source. LoRA-based adaptations (M_2 and M_3) outperform full finetuning (M_1) across all the anatomies under study, possibly because updating fewer low-rank parameters helps preserve the useful pretrained cues, while reducing overfitting on the relatively small US datasets. A broader ablation comparison across M_1-M_9 (Fig. 3) when evaluated on the thyroid dataset shows that generally M_2, which LoRA finetunes both components performs the best on CE-accuracy.

Effect of targeted training. Table 4 shows that all finetuning methods across all the initialisation sources outperform the zero-shot evaluation by huge margins, indicating that task-specific adaptation is necessary for effective US-RG. The cross-task evaluation results in Table 5 reinforces this observation: for example, a model finetuned exclusively on the breast dataset and evaluated

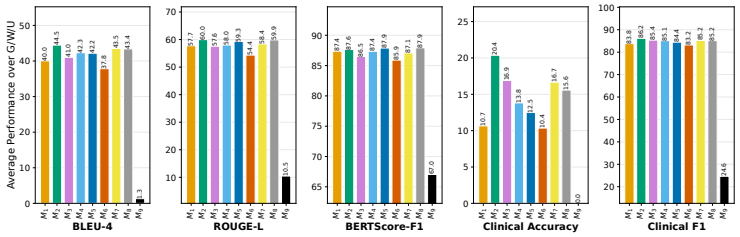

Fig. 3: **Ablation study**: Comparing finetuning strategies from M_1 to M_9, when evaluated on the thyroid dataset. The values are average performances over G, W and U initialisation. The definitions of M_4 to M_9 as pairs of vision encoder/LM finetuning methods are as follows: M_4: Frozen/LoRA, M_5: LoRA/Frozen, M_6: Full/Frozen, M_7: Frozen/Full, M_8: LoRA/Full and M_9: Frozen/Frozen.

on the thyroid dataset still performs considerably better than the frozen model evaluated on the thyroid dataset. This suggests that finetuning on the same task, learns transferrable report generation capabilities, although in general, the low clinical performance indicate that clinical factuality remains an open challenge.

Effect of model size. With only 1-3K patient-level single-anatomy training samples, the larger models seem undertrained relative to their capacity and overfit the small label distribution, thus the 256M model performs better on CE-accuracy. The benefits of the 2.2B model capacity on CE-accuracy arise only in the combined setting, where the data signal is increased (Table. 3).

Limitations. We focus on US-RG; therefore, the extent to which the observed initialisations and adaptation trends generalize to other US understanding tasks remains to be investigated. We use the English-translated dataset by Lyu et al. [14] while the original dataset was released in Chinese [9]. This allows standardized evaluation but may introduce subtle lexical variations. Extending the evaluation to the original Chinese reports represents a future direction.

6 Impact in RCS

Our findings suggest directions for future work in RCS. Comparable performance across initialisation sources suggests that, given the scarcity of open-sourced US data, the US-specific pretraining may not always be necessary for performance improvement in lightweight US-RG models. Increasing model scale does not guarantee performance gains rather improving the parameter-efficient adaptation method may be more rewarding. In-domain targeted training regardless of the initialisation source weights outperforms zero-shot evaluation. Lightweight US-RG models are practical for edge deployment, but further work is needed first to improve clinical-efficacy accuracy.

Table 4: Improvement over zero-shot: Δ = (finetuned) − (frozen M_9) for the same initialisation, averaged over M_1, M_2, M_3. *Comb.*=combined, *Init.*=initialisation.

		NLG ↑			CE ↑	
Dataset	Init.	BLEU-4	ROUGE-L	BERT-F1	Acc.	F1
Breast	G	+41.68	+45.82	+21.25	+36.89	+51.87
	W	+42.01	+50.96	+24.22	+36.53	+76.75
	U	+40.44	+47.34	+18.56	+30.27	+51.20
Thyroid	G	+38.77	+41.95	+19.65	+16.16	+48.16
	W	+41.77	+51.29	+23.07	+16.90	+59.07
	U	+41.06	+50.66	+17.67	+14.87	+74.14
Liver	G	+26.45	+30.08	+19.10	+27.18	+35.03
	W	+29.69	+31.96	+20.04	+29.03	+37.12
	U	+29.51	+37.51	+15.27	+28.43	+39.17
Comb.	G	+39.63	+46.38	+22.47	+51.29	+68.90
	W	+40.67	+47.90	+23.30	+53.36	+67.94
	U	+40.28	+50.70	+19.69	+50.21	+69.69

Table 5: Cross-task evaluation. How a model trained on one anatomy performs when evaluated on another. Thyroid→thyroid and Breast→breast imply on-task evaluation averaged over $M_1/M_2/M_3$, Zero shot uses M_9.

		NLG ↑			CE ↑	
Setting	Init.	BLEU–4	ROUGE-L	BERT-F1	Acc.	F1
Zero-shot→thyroid	G	3.30	16.73	67.90	0.0	36.92
	W	0.58	7.32	64.02	0.0	26.24
	U	0.07	7.34	69.19	0.0	10.79
Breast→thyroid	G	15.21	35.77	80.40	0.31	61.05
	W	15.18	35.90	80.38	0.41	62.64
	U	17.05	34.90	79.46	0.41	64.91
Thyroid→thyroid	G	42.07	58.67	87.55	16.16	85.08
	W	42.34	58.61	87.09	16.90	85.31
	U	41.13	58.01	86.87	14.88	84.92
Breast→breast	G	42.05	58.46	87.67	36.90	84.25
	W	42.01	58.41	87.71	36.53	84.36
	U	40.44	55.66	86.58	30.28	81.81

Disclosure of Interests. This work was funded by Khalifa University of Science and Technology through the Faculty Start-Up (FSU) Grant under Project ID: KU-INT-FSU-2025-014. The authors have no competing interests to declare.

References

1. Abuzaid, M.: Unveiling the landscape: Investigating education, skills, job description, and challenges in sonography professions and framework development. Radiography **30**(1), 125–131 (2024)
2. Alsharid, M., El-Bouri, R., Sharma, H., Drukker, L., Papageorghiou, A.T., Noble, J.A.: A course-focused dual curriculum for image captioning. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 716–720. IEEE (2021)
3. Ambsdorf, J., Munk, A., Llambias, S., Christensen, A.N., Mikolaj, K., Balestriero, R., Tolsgaard, M.G., Feragen, A., Nielsen, M.: General methods make great domain-specific foundation models: A case-study on fetal ultrasound. In: Gee, J.C., Alexander, D.C., Hong, J., Iglesias, J.E., Sudre, C.H., Venkataraman, A., Golland, P., Kim, J.H., Park, J. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. pp. 271–281. Springer Nature Switzerland, Cham (2026)
4. Ferreira, C.A., van Dyk, B., Mokoena, P.L.: The experiences of sonographers with regard to report writing and communicating their findings. Health SA Gesondheid **27**, 2066 (2022)
5. He, K., Girshick, R., Dollár, P.: Rethinking imagenet pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4918–4927 (2019)
6. Jin, J., Chai, H., Huang, X., Guo, X., Zheng, Z., Zhou, Z., Wang, J., Wang, X., Liu, J., Zhou, B.: Ultrasound-clip: Semantic-aware contrastive pre-training for ultrasound image-text understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6962–6971 (2026)

7. Khan, A.U., Ramasamy, G., Khan, M.D., Garrett, J., Bradshaw, T., Salkowski, L., Banerjee, I.: On the transfer learning behavior of domain-specific vision–language models in screening mammography. Journal of Biomedical Informatics p. 104986 (2026)
8. Le, A., Liu, H., Wang, Y., Liu, Z., Zhu, R., Weng, T., Yu, J., Wang, B., Wu, Y., Yan, K., et al.: U2-bench: Benchmarking large vision-language models on ultrasound understanding. arXiv preprint arXiv:2505.17779 (2025)
9. Li, J., Su, T., Zhao, B., Lv, F., Wang, Q., Navab, N., Hu, Y., Jiang, Z.: Ultrasound report generation with cross-modality feature alignment via unsupervised guidance. IEEE Transactions on Medical Imaging **44**(1), 19–30 (2024)
10. Li, X., Li, J., Li, F., Zhu, L., Yang, Y., Shen, H.T.: Generalizing vision-language models to novel domains: A comprehensive survey. ACM Transactions on Multimedia Computing, Communications and Applications (2025)
11. Li, Z., Li, M., Wang, W., Huang, Q.: Ultrasound report generation with fuzzy knowledge and multi-modal large language model. Expert Systems with Applications **292**, 128555 (2025)
12. Liu, J., Hu, T., Zhang, Y., Feng, Y., Hao, J., Lv, J., Liu, Z.: Parameter-efficient transfer learning for medical visual question answering. IEEE Transactions on Emerging Topics in Computational Intelligence **8**(4), 2816–2826 (2023)
13. Lozano, A., Sun, M.W., Burgess, J., Chen, L., Nirschl, J.J., Gu, J., Lopez, I., Aklilu, J., Rau, A., Katzer, A.W., et al.: Biomedica: An open biomedical image-caption archive, dataset, and vision-language models derived from scientific literature. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19724–19735 (2025)
14. Lyu, Z., Zhang, Y., Wang, T., Lan, R.: Ultrasound vision-language alignment via contrastive learning. arXiv preprint arXiv:2605.02126 (2026)
15. Ma, C., Jiao, J., Liang, S., Fu, J., Wang, Q., Li, Z., Wang, Y., Guo, Y.: Tinyusfm: Towards compact and efficient ultrasound foundation models. IEEE Journal of Biomedical and Health Informatics (2026)
16. Marafioti, A., Zohar, O., Farré, M., Noyan, M., Bakouch, E., Cuenca, P., Zakka, C., Allal, L.B., Lozhkov, A., Tazi, N., et al.: Smolvlm: Redefining small and efficient multimodal models. arXiv preprint arXiv:2504.05299 (2025)
17. Mariz, F.: Top 16 ultrasound for android in 2026: The definitive compatibility & performance guide. https://suresultmed.com/ultrasound-for-android/ (2026), accessed: 2026-07-13
18. Phillips, R., Alsop, S.: Developing preceptorship programmes by exploring the needs of newly qualified sonographers through the lens of experienced ultrasound preceptors. Ultrasound **33**(1), 4–11 (2025)
19. Pu, B., Yang, J., Lv, X., Xu, K., Li, K.: Unified mixture-of-experts framework for joint cardiac and vascular ultrasound analysis and report generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 8439–8447 (2026)
20. Raghu, M., Zhang, C., Kleinberg, J., Bengio, S.: Transfusion: Understanding transfer learning for medical imaging. Advances in neural information processing systems **32** (2019)
21. Saeed, N., Hanif, A., Maani, F.A., Alasmawi, H., Yaqub, M.: Dark: Diagonal-anchored repulsive knowledge distillation for vision-language models under extreme compression. arXiv preprint arXiv:2603.05421 (2026)
22. She, C., Lu, R., Chen, L., Wang, W., Huang, Q.: Echovlm: Dynamic mixture-of-experts vision-language model for universal ultrasound intelligence. In: Proceed-

ings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 10800–10822 (2026)
23. Taher, M.R.H., Haghighi, F., Gotway, M.B., Liang, J.: Large-scale benchmarking and boosting transfer learning for medical image analysis. Medical image analysis **102**, 103487 (2025)
24. Tang, Z., Liu, X., Bie, J., Xu, L., Xu, S.: Joint attention gan for medical report generation with clinical style preservation. Scientific Reports **15**(1), 44491 (2025)
25. Yan, L., Zhou, X., Wang, Y., Chang, X., Li, Q., Han, G.: Automated ultrasound diagnosis via clip-gpt synergy: A multimodal framework for image classification and report generation. IEEE Access (2025)