Grounding Before Reporting: Local-to-Global Decoupled Learning for Medical Vision-Language Models

Wenzheng Huang [1[0009-0004-0969-7331]], Jiachen Yuan[1[0009-0000-8763-0440]], Qiujin Wang[1[0009-0001-4682-0392]], Zihan Ye[1[0009-0002-8529-285X]], Camilla Fazi[2,3[0000-0003-3419-4600]], Federica Fontanella[4[0000-0001-7003-4984]] and Netzahualcoyotl Hernandez-Cruz[1(✉)[0000-0002-3603-4806]]

[1] School of AI and Advanced Computing, XJTLU Entrepreneur College (Taicang), Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, China
[2] Department of Neurosciences, Psychology, Drug Research and Child Health, University of Florence, Florence, Italy
[3] Division of Neonatology, Careggi University Hospital of Florence, Florence, Italy
[4]Department of Obstetrics and Gynecology, University Medical Center Groningen, the Netherlands
N.HernandezCruz@xjtlu.edu.cn

Abstract. Vision-Language Models (VLMs) show promise for automated medical report generation, but conventional joint training often couples fine-grained visual grounding and holistic report generation within a single trainable parameter space. This causes optimization conflicts, where models over-rely on dominant global priors and produce generic reports with degraded anatomical grounding. To address this, we propose Local-to-Global Decoupled Learning (LGDL), a staged framework that preserves local anatomical knowledge during global adaptation. LGDL comprises three components: Local Expert Initialization for organ-level grounding, Decoupled Sequential LoRA Adaptation for separate optimization stages, and Local Knowledge Retention for lightweight replay. Evaluated on two grounded benchmarks: FOCUS (fetal ultrasound) and PadChest-GR (chest X-rays), LGDL improves Global and Local F1-scores by +0.1237 and +0.0167 on FOCUS, and +0.0267 and +0.0249 on PadChest-GR. These results demonstrate that explicit decoupling effectively mitigates supervision conflicts, yielding more accurate and anatomically grounded reports.

Keywords: Vision-Language Models · Medical Report Generation · Decoupled Learning · Parameter-Efficient Fine-Tuning.

1 Introduction

Vision-language models (VLMs) are emerging as powerful tools for automated medical report generation, yet extending this capability to fine-grained, multi-structure medical imaging remains challenging [1,2,3,4,5,6]. Existing methods often rely on joint training, in which fine-grained local descriptions and holistic global reports are optimized simultaneously within a single trainable parameter space [4,5,7,8]. Although this strat-

egy is simple and widely adopted, it may introduce interference with optimization between local anatomical grounding and global report generation. In particular, models can over-rely on dominant global priors and produce generic reports that appear plausible but are weakly grounded in specific visual evidence.

This issue is especially pronounced in screening-oriented medical settings such as fetal ultrasound, where normal cases dominate the data distribution [9,10,11]. In such scenarios, end-to-end models may learn frequent normal-case patterns rather than perform evidence-based reasoning across individual anatomical structures. Local annotations are therefore important because they provide explicit supervision for organ-level visual-semantic alignment, helping the model preserve fine-grained anatomical evidence before producing a holistic diagnostic summary.

Medical Vision-Language Models. Automatic medical report generation has been widely studied as a vision-to-language task that translates medical images into clinically meaningful reports. Recent VLMs and LLM-based frameworks have improved linguistic fluency, instruction following, and cross-modal generalization, including LLaVA-Med [1], R2GenGPT [4], and XrayGPT [5]. Related evaluations and surveys also reflect the rapid progress of this direction [7,8,12]. However, most existing frameworks are still optimized primarily with image-level or report-level supervision, providing limited explicit constraints on whether fine-grained anatomical evidence is faithfully captured prior to global summarization.

Global-Local Grounding. To improve interpretability and visual grounding, recent studies have explored localized or anatomy-aware supervision in medical vision-language learning. GLoRIA [13] learns global-local representations from radiology reports, while MedKLIP [14] introduces knowledge-enhanced alignment between chest X-ray images and pathology-aware concepts. MedCLIP [15] and BiomedCLIP [16] further show the value of large-scale contrastive pretraining in biomedical domains. Nevertheless, local signals are often treated as auxiliary supervision and optimized jointly with global objectives in a shared parameter space, which may cause competition between fine-grained anatomical discrimination and high-level clinical abstraction. Unlike conventional region-supervised methods, our setting uses dense organ-level free-text descriptions, requiring the model to preserve local linguistic grounding while learning to generate global reports.

Parameter-Efficient and Staged Adaptation. Parameter-efficient fine-tuning (PEFT) has become a practical strategy for adapting large VLMs under limited resources. LoRA [17] is particularly attractive due to its strong performance and low adaptation overhead, and has been widely used in medical multimodal modeling [1,2,5]. QLoRA [18] and DoRA [19] further improve adaptation efficiency. Beyond efficiency, PEFT modules provide a natural mechanism for separating heterogeneous objectives. This motivates us to use LoRA not only as a lightweight tuning tool but also as a structural

decoupling mechanism to isolate local grounding from global report generation. Inspired by continual learning [20,21], we further introduce lightweight replay to reduce forgetting during the transition from local to global supervision.

To address the above limitations, we propose Local-to-Global Decoupled Learning (LGDL), a staged training framework for grounded medical report generation. LGDL consists of three components. First, Local Expert Initialization trains a local LoRA adapter exclusively on organ-level descriptions, establishing fine-grained visual-semantic grounding. The learned local knowledge is then merged into the backbone before global adaptation. Second, Decoupled Sequential LoRA Adaptation (DSLA) initializes a new global LoRA adapter and optimizes it solely for report-level generation, thereby separating global reasoning from local grounding at the parameter level. Third, Local Knowledge Retention (LKR) incorporates a small portion of local samples during global adaptation to preserve anatomical grounding and stabilize the transition from local to global learning. We evaluate LGDL on two datasets: FOCUS, a fetal ultrasound dataset with organ-level descriptions and global diagnostic summaries, and PadChest-GR, a grounded chest X-ray dataset. On FOCUS, LGDL improves the Global F1-score from 0.6198 to 0.7435 and the Local F1-score from 0.4574 to 0.4741 over conventional joint training. On PadChest-GR, it also achieves consistent gains in both global and local metrics, demonstrating that the proposed local-to-global decoupling strategy generalizes beyond a single imaging modality. The contributions are:

– **A local-to-global decoupled training framework for grounded medical report generation.** We propose LGDL, which explicitly separates local anatomical grounding from global report generation to mitigate interference from optimization under multi-granularity supervision.

– **A parameter-decoupled adaptation strategy with replay-based retention.** We show that parameter-efficient fine-tuning can serve not only as an efficient tool but also as a structural decoupling mechanism. The proposed DSLA isolates global adaptation from previously learned local knowledge, while LKR helps preserve anatomical grounding during global adaptation.

– **Cross-dataset validation under different supervision structures.** We evaluate LGDL on FOCUS and PadChest-GR, which differ in imaging modality, annotation granularity, and reporting style. The consistent improvements demonstrate that the proposed local-to-global decoupling strategy is applicable to different grounded medical report generation settings.

2 Method

We propose Local-to-Global Decoupled Learning (LGDL), a staged training framework for anatomically grounded medical report generation under multi-granularity supervision. As illustrated in **Fig. 1**, LGDL first learns anatomy-aware local grounding from fine-grained descriptions, then performs global report adaptation in a decoupled LoRA subspace, and finally uses lightweight replay to retain local knowledge. Accordingly, LGDL code coupling comprises three components: Local Expert Initialization for visual-semantic grounding, Decoupled Sequential LoRA Adaptation (DSLA) for

separating local and global optimization, and Local Knowledge Retention (LKR) for preserving anatomical grounding during global adaptation.

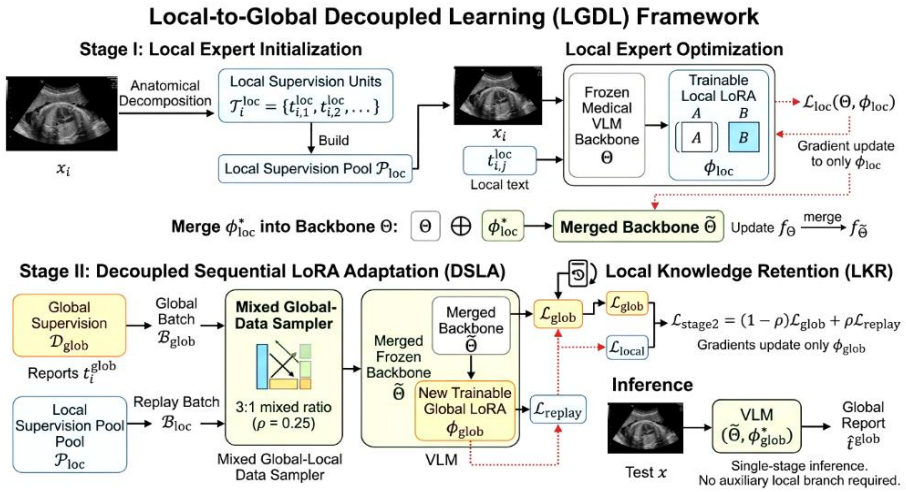


Fig. 1. The proposed LGDL framework first learns organ-level visual-semantic grounding through Local Expert Initialization, then performs global report adaptation with a re-initialized LoRA adapter while retaining local knowledge through lightweight replay.

2.1 Problem Formulation

Given an input medical image $x_i \in \mathcal{X}$, we consider two levels of supervision. Local supervision is denoted by $\mathcal{T}_i^{\text{loc}} = \{t_{i,1}^{\text{loc}}, \ldots, t_{i,K_i}^{\text{loc}}\}$, where each $t_{i,k}^{\text{loc}}$ describes a specific anatomical structure or localized finding. The global supervision is a holistic report t_i^{glob} describing the overall clinical interpretation. Accordingly, the local and global datasets are defined as:

$$D_{loc} = \{(x_i, T_i^{loc})\}_{i=1}^{N}, \quad D_{glob} = \{(x_i, t_i^{glob})\}_{i=1}^{N}. \tag{1}$$

A conventional joint-training paradigm optimizes both objectives simultaneously:

$$\theta^* = \arg\min_{\theta} \left(L_{loc}(\theta) + \lambda L_{glob}(\theta) \right), \tag{2}$$

where λ balances the local and global losses. However, local supervision encourages fine-grained anatomical discrimination, whereas global supervision emphasizes report-level semantic abstraction. Optimizing both in a shared parameter space may therefore introduce competing gradient signals, motivating the decoupled design of LGDL.

2.2 Overall Framework

Let f_Θ denote a pretrained medical VLM with frozen backbone parameters Θ. We adopt LoRA for parameter-efficient adaptation. For a projection matrix $W \in R^{d\times k}$, LoRA updates it as:

$$W' = W + \Delta W, \quad \Delta W = BA, \tag{3}$$

where $A \in R^{r\times k}$ and $B \in R^{d\times r}$ are low-rank trainable matrices.
LGDL introduces two sequential adapters: a local adapter ϕ_{loc} for anatomical grounding and a global adapter ϕ_{glob} for report generation. The overall training pipeline is formulated as:

$$f_\Theta \to f_{\Theta+\phi_{loc}} \to f_{\tilde{\Theta}} \to f_{\tilde{\Theta}+\phi_{glob}}. \tag{4}$$

2.3 Stage I: Local Expert Initialization

Stage I aims to establish fine-grained visual-semantic grounding before global report adaptation. For each image x_i, the anatomy-specific local descriptions are flattened into a unified local supervision pool:

$$P_{loc} = \{(x_i, t_{i,j}^{loc}) \mid i = 1 \ldots N,\ \mathrm{j} = 1 \ldots K_i\}. \tag{5}$$

With the pretrained backbone Θ frozen, only the local adapter ϕ_{loc} is optimized using the autoregressive local generation loss:

$$L_{loc} = -\sum_{(x_i, t_{i,j}^{loc})\in P_{loc}} \sum_{m=1}^{|t_{i,j}^{loc}|} \log p\left(t_{i,j,m}^{loc} \mid x_i, t_{i,j,<m}^{loc}; \Theta, \phi_{loc}\right). \tag{6}$$

After convergence, the learned local adapter is merged into the backbone:

$$\tilde{\Theta} = \Theta \oplus \phi_{loc}^{*}. \tag{7}$$

This produces an anatomy-aware backbone $\tilde{\Theta}$, which serves as the foundation for subsequent global report learning.

2.4 Stage II: Decoupled Sequential LoRA Adaptation

Stage II performs global report adaptation on a newly initialized LoRA subspace. Unlike joint training, where local and global objectives are optimized simultaneously within the same trainable parameters, LGDL first consolidates local anatomical knowledge into $\tilde{\Theta}$ and then updates only the global adapter ϕ_{glob}.

This design reduces direct competition between fine-grained grounding and high-level report abstraction. Formally, the global report generation objective is:

$$L_{glob} = -\sum_{\left(x_i, t_i^{glob}\right)\in D_{glob}} \sum_{m=1}^{\left|t_i^{glob}\right|} \log p\left(t_{i,m}^{glob} \mid x_i, t_{i,<m}^{glob}; \widetilde{\Theta}, \phi_{glob}\right). \quad (8)$$

Here, $\widetilde{\Theta}$ remains frozen during Stage II, and only ϕ_{glob} is updated. Therefore, the global adapter can focus on report-level abstraction without directly overwriting the local representations acquired in Stage I.

2.5 Local Knowledge Retention

Although $\widetilde{\Theta}$ encodes local anatomical knowledge, global adaptation may still bias the model toward dominant report-level patterns. We therefore introduce Local Knowledge Retention (LKR), a lightweight replay mechanism for Stage II.

Specifically, each Stage-II update uses a global mini-batch $B_{glob} \sim D_{glob}$ and a small local replay mini-batch $B_{loc} \sim P_{loc}$. The optimization objective is:

$$L_{stage2} = (1-\rho)L_{glob} + \rho\, E_{(x,t)\sim B_{loc}}\left[l_{NLL}\left(t \mid x; \widetilde{\Theta}, \phi_{glob}\right)\right], \quad (9)$$

where $l_{\text{NLL}}(t \mid x;\cdot) = -\sum_{m=1}^{|t|} \log p\left(t_m \mid x, t_{<m};\cdot\right)$ denotes the autoregressive negative log-likelihood on replayed local descriptions. This expectation-based form emphasizes that LKR serves as a sampled retention regularizer rather than a second full local-training stage.

We adopt a 3:1 global-to-local sampling ratio and set $\rho = 0.25$, which balances global report quality and local anatomical grounding; this choice is examined in Sec. 3.4. The final global adapter is obtained by:

$$\phi_{glob}^{*} = \arg\min_{\phi_{glob}} L_{stage2}\left(\widetilde{\Theta}, \phi_{glob}\right). \quad (10)$$

2.6 Inference

At test time, LGDL directly generates the global report through a single end-to-end autoregressive decoding process:

$$\widehat{t^{glob}} = \arg\max_{t} p\left(t \mid x; \widetilde{\Theta}, \phi_{glob}^{*}\right). \quad (11)$$

As shown in **Fig. 1**, inference requires only the merged backbone $\widetilde{\Theta}$ and the learned global adapter ϕ_{glob}^{*}. After Stage I, the local adapter is merged into the backbone, so inference uses only the final model without an auxiliary branch, extra forward pass, or multi-stage decoding. LGDL therefore adds no inference overhead over joint training; its extra cost is limited to training, where only lightweight LoRA adapters are updated.

3 Experiments and Results

3.1 Experimental Setup

We evaluate LGDL on two datasets: FOCUS for fetal cardiac ultrasound and PadChest-GR for chest X-ray reporting. FOCUS [22] contains 300 four-chamber view (4CHV) fetal ultrasound images from 217 pregnant participants. To support local-to-global report generation, two fetal sonographers annotated the data at two levels of detail: local and global. At both levels, annotations included bounding boxes and organ-specific textual descriptions. Locally, these annotations were provided for 16 anatomical structures, and the heart as a whole. The annotations provide dense anatomical supervision, while the global descriptions serve as holistic report-generation targets. The annotations were conducted and managed on Label Studio [23]. PadChest-GR [24] consists of chest X-ray images with region-level findings as local supervision and impression-level summaries as global targets, enabling evaluation of whether local-to-global decoupling remains effective beyond dense fetal ultrasound annotation.

Evaluation Metrics. We report SBERT, F1-score, and ROUGE-L for both global reports and local grounded descriptions. Global F1 is treated as a report-level overlap indicator, while Local F1 reflects anatomical grounding.

Implementation Details. Both datasets are split into training, validation, and test sets using an 8:1:1 ratio. For FOCUS, splitting is conducted at the participant level to avoid data leakage across images from the same subject. All reported results are evaluated on the held-out test sets.

All experiments are implemented using VILA-M3-8B and trained on four NVIDIA A800 GPUs for 15 epochs. We use a per-device batch size of 4, learning rate 2×10^{-4}, LoRA rank 64, LoRA alpha 128, maximum sequence length 2048, and generation decoding length 256. The joint-training baseline and LGDL share the same backbone, splits, preprocessing, decoding strategy, and LoRA settings, so the comparison isolates the proposed decoupling strategy. For LGDL, the local replay ratio is set to $\rho = 0.25$ unless otherwise specified.

3.2 Main Quantitative Results

Table 1 reports the quantitative results on FOCUS and PadChest-GR. On FOCUS, LGDL improves Global F1 from 0.6198 to 0.7435 and Local F1 from 0.4574 to 0.4741. Since both methods use the same backbone and tuning settings, the gains are mainly attributable to the proposed local-to-global decoupling.

The Stage I local expert achieves the highest local scores on FOCUS, with Local F1 of 0.5091 and ROUGE-L of 0.4607. This is expected because Stage I is optimized exclusively for local anatomical descriptions. After global adaptation, the final LGDL

model slightly reduces local-only performance but substantially improves global report-level performance, showing that the proposed staged framework maintains a favorable balance between local grounding and holistic summarization.

Table 1. Quantitative results across datasets. Bold indicates the best result in each category.

Dataset	Strategy	Category	SBERT ↑	F1 ↑	ROUGE-L ↑
FOCUS	Baseline (Joint Train)	Global	0.8412	0.6198	0.5528
		Local	0.7978	0.4574	0.4122
	Local Expert	Local	**0.8246**	**0.5091**	**0.4607**
	LGDL (Ours)	Global	**0.8981**	**0.7435**	**0.7066**
		Local	0.8136	0.4741	0.4187
PadChest-GR	Baseline (Joint Train)	Global	0.3030	0.2151	0.1930
		Local	0.5911	0.3966	0.3847
	LGDL (Ours)	Global	**0.3663**	**0.2418**	**0.2154**
		Local	**0.6024**	**0.4215**	**0.4109**

On PadChest-GR, LGDL consistently improves over joint training across all reported metrics, increasing Global F1 from 0.2151 to 0.2418 and Local F1 from 0.3966 to 0.4215. Although the gain is smaller than on FOCUS, this result is still meaningful because PadChest-GR provides shorter region-level phrases rather than dense organ-specific descriptions. The consistent improvement suggests that explicit parameter decoupling is beneficial beyond the dense-captioning setting, indicating cross-modality robustness from fetal ultrasound to chest X-ray reporting.

3.3 Ablation on Framework Components

We conduct component ablations on FOCUS to examine the effects of Local Expert Initialization (LEI), Decoupled Sequential LoRA Adaptation (DSLA), and Local Knowledge Retention (LKR). The results are reported in **Table 2**.

Variant (B) performs LEI and then continues global training on the same LoRA adapter without DSLA. Although Global F1 increases from 0.6198 to 0.6568, Local F1 drops from 0.4574 to 0.4346, showing that sequential training alone cannot prevent global updates from overwriting fine-grained anatomical knowledge.

Table 2. Ablation study of LGDL components on FOCUS. The highest Global F1 is underlined, and the highest Local F1 is bolded.

Strategy	LEI	DSLA	LKR	Global F1 ↑	Local F1 ↑
(A) Joint Training				0.6198	0.4574
(B) w/o DSLA	✓			0.6568	0.4346
(C) w/o LKR	✓	✓		<u>0.7624</u>	0.4466
(D) Full LGDL	✓	✓	✓	0.7435	**0.4741**

Variant (C) uses LEI and DSLA but removes LKR by setting $\rho=0$. It achieves the highest Global F1 of 0.7624, yet Local F1 decreases to 0.4466, below the joint-training

baseline. This suggests that strong report-level overlap may still rely on dominant report patterns rather than grounded local evidence.

The full LGDL model restores local supervision through LKR during Stage II. It achieves the best Local F1 of 0.4741 while still maintaining a large Global F1 improvement over the baseline (0.7435 vs. 0.6198). We therefore select the full model for its stronger grounding-reporting balance rather than the highest global score alone.

3.4 Ablation on Local Knowledge Retention Ratio

We further study the effect of the local replay ratio $\rho \in [0,0.5]$ on FOCUS, as shown in **Fig. 2**. When $\rho = 0$, the model performs pure global adaptation and obtains high Global F1, but Local F1 becomes the weakest among the variants. As ρ increases, local grounding is gradually recovered, while overly large replay ratios introduce stronger competition with the global report-generation objective.

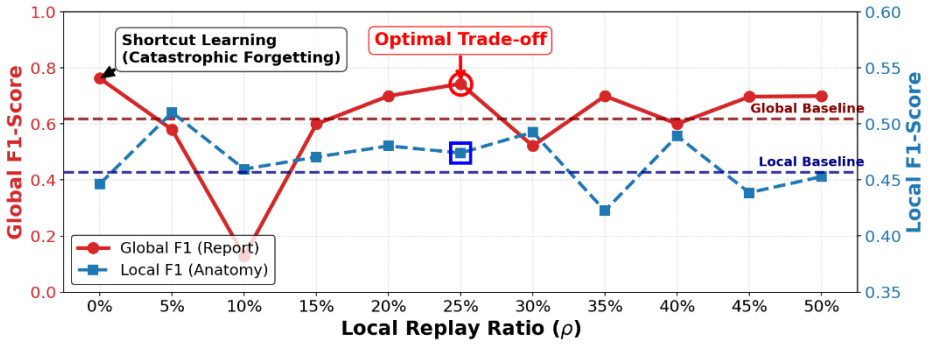


Fig. 2. Effect of the replay ratio ρ on FOCUS. A moderate replay ratio provides the best local–global balance, while larger values tend to favor local retention over global report generation. Dashed lines indicate joint-training baselines.

The trend shows moderate sensitivity to ρ: too little replay weakens local retention, while too much replay biases Stage II toward local supervision. We set $\rho = 0.25$ as the default because it gives the best observed local-global balance, with Global F1 = 0.7435 and Local F1 = 0.4741.

These results clarify the role of LKR in balancing report generation and anatomical grounding. Although ρ=0.25 does not yield the highest Global F1, it achieves the strongest Local F1 while maintaining a clear global improvement over joint training. This supports our design choice that grounded report generation should not be judged by global text similarity alone.

3.5 Qualitative Visualization

Fig. 3 shows an attention-based visual attribution example for the right atrium in a 4CHV from the FOCUS dataset. The maps are obtained from cross-attention weights

in the final decoding layer and upsampled to the original image resolution. Compared with joint training, LGDL focuses more precisely on clinically relevant structures, including the right atrium wall, right ventricular wall, and moderator band. As the moderator band is an important cue for identifying the morphological right ventricle, its highlighted response also supports more reliable right atrial grounding. In contrast, the baseline attribution is more diffuse and less specifically tied to the target anatomy.

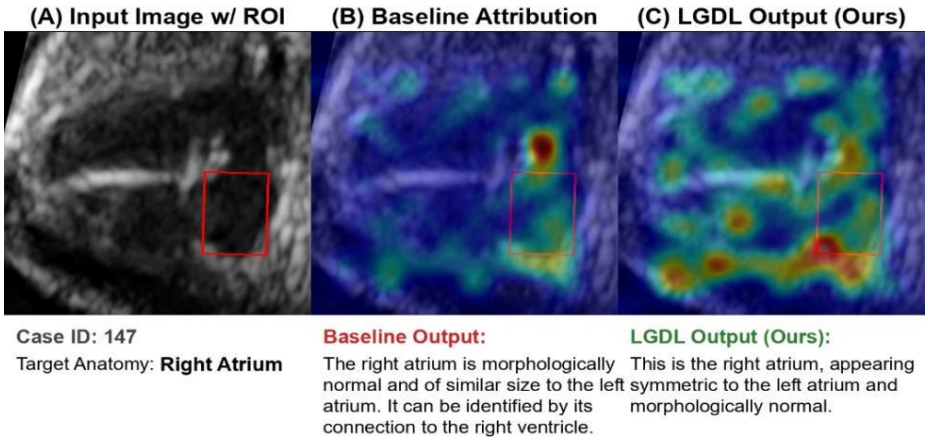


Fig. 3. Qualitative comparison of right-atrium report generation from a 4CHV from the FOCUS dataset. A–C images show the ground truth, joint-training baseline, and LGDL prediction, respectively. LGDL better preserves local anatomical evidence for fine-grained report generation.

4 Discussion and Conclusion

4.1 Discussion

The results show that combining local and global supervision improves both anatomical grounding and report generation. Stage I learns fine-grained anatomical representations, and the final LGDL model preserves this information more effectively than joint training while improving global report generation.

These findings indicate that joint optimization can entangle anatomical discrimination with report abstraction. LGDL mitigates this by consolidating local knowledge before global adaptation, while LKR reduces reliance on dominant report patterns through lightweight local replay.

4.2 Limitations and Future Work

This study has three main limitations. First, FOCUS is densely annotated but small, so future work should evaluate larger multi-center cohorts. Second, the current framework uses static 2D images, whereas fetal echocardiography is dynamic; extending LGDL to

ultrasound videos is a natural direction. Third, ρ is fixed, and adaptive retention based on forgetting behavior may further improve the local-global balance.

4.3 Conclusion

We proposed LGDL for anatomically grounded medical report generation. By separating local grounding from global report adaptation via sequential LoRA, LGDL reduces interference between fine-grained perception and holistic summarization during optimization. Experiments on FOCUS and PadChest-GR show consistent improvements over joint training, demonstrating the value of explicitly separating grounding and reporting in medical VLMs.

Acknowledgments. This work was supported by the Research Development Fund Grant No. RDF-25-01-076.

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

References

1. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. arXiv preprint arXiv:2306.00890 (2023)
2. Moor, M., Huang, Q., Wu, S., Yasunaga, M., Dalmia, Y., Leskovec, J., Zakka, C., Reis, E.P., Rajpurkar, P.: Med-Flamingo: a Multimodal Medical Few-shot Learner. In: Proceedings of the 3rd Machine Learning for Health Symposium, Proceedings of Machine Learning Research, vol. 225, pp. 353–367 (2023)
3. Chen, Z., Varma, M., Delbrouck, J.-B., Paschali, M., Blankemeier, L., Van Veen, D., Valanarasu, J.M.J., Youssef, A., Cohen, J.P., Reis, E.P., Tsai, E.B., et al.: CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation. arXiv preprint:2401.12208 (2024)
4. Wang, Z., Liu, L., Wang, L., Zhou, L.: R2GenGPT: Radiology Report Generation with Frozen Large Language Models. Medical Image Analysis 91, 103033 (2024)
5. Thawakar, O.C., Shaker, A.M., Mullappilly, S.S., Cholakkal, H., Anwer, R.M., Khan, S.H., Laaksonen, J., Khan, F.S.: XrayGPT: Chest Radiographs Summarization Using Large Medical Vision-Language Models. In: Proceedings of the 23rd Workshop on Biomedical Natural Language Processing, pp. 440–448 (2024)
6. Tu, T., Azizi, S., Driess, D., et al.: Towards Generalist Biomedical AI. NEJM AI 1(3), AIoa2300138 (2024)
7. Wang, X., Figueredo, G., Li, R., Zhang, W.E., Chen, W., Chen, X.: A Survey of Deep-Learning-Based Radiology Report Generation Using Multimodal Inputs. Medical Image Analysis 103, 103627 (2025)
8. Salhi, M., Akhloufi, M.A.: Recent Progress in Deep Learning for Chest X-Ray Report Generation. BioMedInformatics 6(1), 3 (2026)
9. Hernandez-Cruz, N., Patey, O., Salovic, B., Papageorghiou, A., and Noble, J.A.: Machine learning-based method for the detection of dextrocardia in ultrasound video clips. In: Computer Methods and Programs in Biomedicine, vol. 271, pp. 109023 (2025)

10. Hernandez-Cruz, N., Patey, O., Teng, C., Papageorghiou, A.T., and Noble, J.A.: A comprehensive scoping review on machine learning-based fetal echocardiography analysis. In: Computers in Biology and Medicine, vol. 186, pp. 109666 (2025)
11. Hernandez-Cruz, N., Patey, O., Salovic, B., Mishra, D., Sarker, M.M.K., Papageorghiou, A., and Noble, J.A.: Detection of fetal congenital heart defects on three-vessel view ultrasound videos. In: WFUMB Ultrasound Open, vol. 2, no. 2, pp. 100075 (2024)
12. Ziegelmayer, S., Marka, A.W., Calandriello, L., et al.: Evaluation of GPT-4's Chest X-Ray Impression Generation: A Reader Study on Performance and Perception. Journal of Medical Internet Research 25, e50865 (2023)
13. Huang, S.-C., Shen, L., Lungren, M.P., Yeung, S.: GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-Efficient Medical Image Recognition. In: Proceedings of ICCV, pp. 3942–3951 (2021)
14. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: MedKLIP: Medical Knowledge Enhanced Language-Image Pre-Training for X-Ray Diagnosis. In: Proceedings of ICCV, pp. 21372–21383 (2023)
15. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. In: Proceedings of EMNLP, pp. 3876–3887 (2022)
16. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, J.S., Rao, R., Wei, M., Valluri, N., Wong, C., et al.: BiomedCLIP: A Multimodal Biomedical Foundation Model Pretrained from Fifteen Million Scientific Image-Text Pairs. arXiv preprint arXiv:2303.00915 (2023)
17. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-Rank Adaptation of Large Language Models. In: International Conference on Learning Representations (2022)
18. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: QLoRA: Efficient Finetuning of Quantized LLMs. In: Advances in Neural Information Processing Systems 36 (2023)
19. Liu, S.-Y., Wang, C.-Y., Yin, H., Molchanov, P., Wang, Y.-C.F., Cheng, K.-T., Chen, M.-H.: DoRA: Weight-Decomposed Low-Rank Adaptation. In: Proceedings of the 41st International Conference on Machine Learning, Proceedings of Machine Learning Research, vol. 235, pp. 32100–32121 (2024)
20. Rebuffi, S.-A., Kolesnikov, A., Sperl, G., Lampert, C.H.: iCaRL: Incremental Classifier and Representation Learning. In: Proceedings of CVPR, pp. 2001–2010 (2017)
21. Shi, H., Xu, Z., Wang, H., Qin, W., Wang, W., Wang, Y., Wang, Z., Ebrahimi, S., Wang, H.: Continual Learning of Large Language Models: A Comprehensive Survey. ACM Computing Surveys 58(5) (2025)
22. Wu, S., Zhang, H., Ye, T., Xie, H., Zeng, P., Sun, Q., Wang, P., Huang, B., Du, L., Wu, G.: FOCUS: Four-chamber Ultrasound Image Dataset for Fetal Cardiac Biometric Measurement (1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.14597550 (2025)
23. Tkachenko, M., Malyuk, M., Holmanyuk, A., Liubimov, N.: Label Studio: Data labeling software. HumanSignal, open-source software. https://github.com/HumanSignal/label-studio (2020–2025)
24. de Castro, D.C., Bustos, A., Bannur, S., Hyland, S.L., Bouzid, K., Wetscherek, M.T., Sánchez-Valverde, M.D., Jaques-Pérez, L., Pérez-Rodríguez, L., Takeda, K., Salinas-Serrano, J.M., Alvarez-Valle, J., Galant-Herrero, J., Pertusa, A.: PadChest-GR: A Bilingual Chest X-Ray Dataset for Grounded Radiology Report Generation. NEJM AI 2(7), AIdbp2401120 (2025)