A Resource-Constrained Evaluation of Lightweight Adaptation for Cross-Dataset Chest X-ray Report Generation

Minji Kim[1 [0009−0003−1514−5210]], Seun Kim[1 [0009−0009−4541−8105]], and Jang-Hwan Choi[1,2✉ [0000−0001−9273−034X]]

[1] Department of Artificial Intelligence, Ewha Womans University, Seoul, South Korea
{2277039, se0102s, choij}@ewha.ac.kr
[2] Department of Artificial Intelligence and Software, Ewha Womans University, Seoul, South Korea

Abstract. Reliable chest X-ray report generation requires models to remain robust when deployed outside the source dataset. This study examines a challenging IU X-Ray-to-CheXpert Plus transfer setting for an IU X-Ray-pretrained BiomedGPT model. Under constrained computation, we evaluate three lightweight adaptation families: supervised Low-Rank Adaptation (LoRA) variants, including vision-only, vision–decoder, knowledge-distillation/asymmetric-loss, and SWAD-inspired checkpoint averaging; router-only test-time adaptation with Mixture-of-Modality-Experts (MoME-TTA) using unlabeled target images; and Meta Variational Information Bottleneck (MetaVIB) latent regularization. A Mahalanobis analysis of BiomedGPT visual features showed pronounced source–target separation, with a source mean distance of 22.94, target mean distance of 165.86, and AUROC of 1.00, confirming a severe visual domain gap. However, report-level improvements were small: the best point estimates improved the source-only baseline by 0.0035 in ROUGE-1 and 0.0009 in BLEU-4, and BERTScore F1 varied by less than 0.002 across adapted models. Exploratory LLM-assisted error review further suggested that higher lexical overlap did not reliably correspond to better clinical specificity, fewer omissions, or fewer unsupported findings. These results indicate that feature-level or classification-oriented adaptation alone is insufficient for autoregressive radiology report generation, where target-domain report style, clinical terminology, and decoder behavior jointly affect output quality. Our study provides a practical diagnostic benchmark for developing domain-robust medical report generation methods in resource-constrained settings.

Keywords: Domain Shift Adaptation · Medical Vision-Language Model · Chest X-ray Report Generation · Parameter-Efficient Fine-Tuning · Test-Time Adaptation · Meta-learning

1 Introduction

Automated radiology report generation has progressed rapidly with large-scale medical vision-language models. Models such as BiomedGPT [16] can generate free-text reports from chest X-rays by combining visual encoders with autoregressive language generation. However, reliable clinical deployment requires these models to remain robust when applied outside the source dataset, since hospitals differ in imaging devices, acquisition protocols, patient populations, and report-writing conventions.

This work focuses on transfer from IU X-Ray (Indiana University Chest X-ray Collection) [4] to CheXpert Plus [3], a realistic and severe cross-dataset setting. Compared with IU X-Ray, CheXpert Plus reflects different visual characteristics and institutional reporting styles, introducing both visual and linguistic distribution shifts. Such shifts are particularly challenging for autoregressive report generation because even small representation changes may propagate through the decoder and affect the generated clinical narrative. Prior adaptation methods, including parameter-efficient fine-tuning, domain generalization, meta-learning, and test-time adaptation (TTA), were largely developed for classification or retrieval tasks, where adaptation can rely on class probabilities, feature invariance, or prediction entropy. Whether these objectives directly improve open-ended report generation remains unclear, particularly under resource constraints where full model retraining is impractical.

We investigate lightweight adaptation strategies for an IU X-Ray-pretrained BiomedGPT model transferred to CheXpert Plus. Rather than proposing a large new architecture, our goal is to provide a controlled diagnostic study of what currently fails and why. We compare supervised LoRA-based adaptation, unlabeled MoME-TTA, and MetaVIB-based latent regularization under limited computation. We evaluate the resulting reports using standard natural language generation metrics and complement them with an exploratory LLM-assisted error review focused on clinically relevant failure patterns.

The main contributions of this paper are as follows:

1. We benchmark lightweight adaptation strategies for cross-dataset chest X-ray report generation, including four LoRA-based variants, router-only MoME-TTA, and MetaVIB.
2. We quantify the IU X-Ray-to-CheXpert Plus visual domain gap using feature-space Mahalanobis distance and show strong separation in the BiomedGPT visual representation.
3. We adapt a classification-oriented TTA approach to an autoregressive report generation backbone and analyze the limitations of router-only inference-time adaptation.
4. We show that small improvements in lexical metrics do not necessarily correspond to clinically better reports, highlighting the need for evaluation protocols that capture omissions, unsupported findings, and target-domain report style.

Overall, our findings suggest that adaptation strategies developed for classification tasks cannot be assumed to improve autoregressive medical report generation under severe clinical domain shift. Effective adaptation requires joint consideration of visual robustness, target-domain language, and decoder-level clinical semantics.

2 Materials and Methods

2.1 Dataset Description

We evaluate cross-dataset adaptation from IU X-Ray (Indiana University Chest X-ray Collection) [4] to CheXpert Plus [3]. IU X-Ray contains 7,470 chest X-ray images paired with 3,955 reports. CheXpert Plus contains 223,462 image–report pairs across 187,711 studies from 64,725 patients. These datasets differ in scale, institution, imaging distribution, patient characteristics, and reporting conventions, making the transfer setting clinically realistic and challenging.

All experiments were conducted in a limited-resource Google Colab environment. For supervised training-time adaptation, 2,377 CheXpert Plus samples were used. For MoME-TTA, 897 unlabeled CheXpert Plus images were used for router optimization. For MetaVIB, 1,044 IU X-Ray images were used to construct internal pseudo-domains, and 1,000 samples were used for meta-learning experiments. Evaluation was performed on a randomly sampled CheXpert Plus subset of 201 images. Because the available evaluation set is small, the results should be interpreted as a resource-constrained diagnostic study rather than a definitive benchmark.

2.2 Base Model: BiomedGPT

We use BiomedGPT [16] as the backbone model for image-conditioned report generation. BiomedGPT is a multimodal generative model that combines a visual encoder with a language generation module. In this study, the source model is a BiomedGPT checkpoint pretrained on IU X-Ray. The source-only baseline applies this checkpoint directly to CheXpert Plus without target-domain adaptation. All evaluated adaptation methods retain the same backbone and update only lightweight components or adapter parameters.

2.3 Domain Shift Adaptation Methods

LoRA-based adaptation. For supervised target-domain adaptation, we apply Low-Rank Adaptation (LoRA) [7]. LoRA freezes the pretrained backbone and injects trainable low-rank adapter weights into selected modules, enabling parameter-efficient fine-tuning.

We evaluate four LoRA variants. First, *vision-only LoRA* updates selected visual encoder layers to adapt image features while leaving the language-generation path unchanged. Second, *vision–decoder LoRA* applies the same visual adapters

and additionally updates decoder cross-attention modules, allowing adaptation of the image-conditioned generation pathway. Third, *KD–ASL LoRA* combines cross-entropy training with token-level knowledge distillation from the frozen source model [6] and an auxiliary asymmetric loss [1] for pathology-label prediction using CheXbert-derived weak labels [12]. Fourth, *SWAD-inspired LoRA* averages selected low-validation-loss LoRA checkpoints following the robustness motivation of SWAD [2], while keeping the BiomedGPT backbone fixed.

MoME-TTA. We evaluate Mixture-of-Modality-Experts Test-Time Adaptation (MoME-TTA) [14] as an unlabeled target-domain adaptation method. MoME-TTA is attractive for clinical deployment because it adapts during inference using only target-domain images and updates a small routing component rather than the full model.

To apply MoME-TTA to BiomedGPT, we freeze the pretrained backbone, language generation modules, and parallel expert adapters. Only the entropy-based router is optimized using CheXpert Plus images. Because the original MoME design used MLP-style adapters, we replace them with convolutional bottleneck blocks that better preserve spatial representations in the ResNet-based BiomedGPT visual encoder. The modified adapters are inserted after the `conv2` outputs and within the final four blocks of the `conv3` stage. During adaptation, the trainable router accounts for approximately 0.009% of the total model parameters.

MetaVIB. MetaVIB [5] models classifier parameters probabilistically and uses a variational information bottleneck to encourage domain-robust latent representations. Although MetaVIB was developed for classification-oriented domain generalization, we use its latent regularization framework to test whether more domain-invariant visual features can improve autoregressive report generation.

Because only one labeled source domain is available, true multi-domain meta-learning is not possible in our setting. We therefore construct pseudo-domains within IU X-Ray: frontal-view images are treated as in-domain samples and lateral-view images as pseudo-OOD samples. BiomedGPT visual features are clustered into eight groups using k-means, and cluster assignments are used as pseudo-labels for meta-learning. This design provides a practical stress test, but the resulting pseudo-OOD split should not be interpreted as equivalent to a true external hospital-domain shift.

2.4 Implementation Details

To quantify the visual domain gap, we extract image features from the BiomedGPT visual encoder and compute Mahalanobis distances [9]. IU X-Ray features are used to estimate the source distribution, and distances are computed for both IU X-Ray and CheXpert Plus samples. AUROC is reported by treating CheXpert Plus as the out-of-domain target distribution.

Table 1. Mahalanobis distance evaluation between IU X-Ray and CheXpert Plus.

Metric	Value
Source mean (IU X-Ray)	22.9422
Target mean (CheXpert Plus)	165.8608
AUROC	1.0000

Table 2. OOD detection performance of MetaVIB.

Metric	Value
AUROC	0.5867
AUPR	0.5783
FPR95	0.9283
ID mean distance	2.5468
OOD mean distance	3.2550

Table 1 shows a large increase in mean Mahalanobis distance from IU X-Ray to CheXpert Plus and perfect separability under this feature-space OOD criterion. These results confirm that the target dataset lies far from the source distribution in the BiomedGPT visual representation. We therefore use this setting to test whether lightweight adaptation can translate feature-level alignment into improved report generation.

For LoRA adaptation, 2,377 CheXpert Plus samples are split into training and validation sets at a 9:1 ratio. In the vision-only setting, LoRA is applied to convolutional modules in the third stage of the visual encoder. In the vision–decoder setting, the same visual targets are retained and LoRA is additionally applied to cross-attention modules in the last four decoder layers. For KD–ASL LoRA, the loss weights are set to $\lambda_{\mathrm{CE}} = 1.0$, $\lambda_{\mathrm{KD}} = 0.2$, and $\lambda_{\mathrm{ASL}} = 0.05$, while uncertain and missing CheXbert labels are ignored. For SWAD-inspired LoRA, training is initialized from the best vision–decoder checkpoint, and adapters within 0.20 validation loss of the best checkpoint are averaged.

For MoME-TTA, an IU X-Ray feature bank is constructed using 594 source images, and 897 CheXpert Plus images are used for router optimization. For MetaVIB, 1,044 IU X-Ray images are used to construct pseudo-domains from frontal and lateral views. The clustering details are provided in Supplementary Table S1.

As shown in Table 2, MetaVIB increases the distance between pseudo-ID and pseudo-OOD groups, but AUROC and AUPR remain near random performance. This indicates substantial overlap between the pseudo-domains and suggests that view-based pseudo-OOD construction provides only weak supervision for learning domain-robust representations.

Table 3. Quantitative comparison of BiomedGPT models with different adaptation strategies on the CheXpert Plus dataset. Bold values denote the best point estimate for each metric.

Model	ROUGE-1	ROUGE-L	BLEU-1	BLEU-4	BERTScore F1
Source-only BiomedGPT	0.2213	0.1560	0.0993	0.0158	0.7972
Vision-only LoRA	0.2233	**0.1578**	0.0999	0.0149	0.7971
Vision–decoder LoRA	0.2022	0.1450	0.0859	0.0127	0.7984
KD–ASL LoRA	0.2094	0.1510	0.0917	0.0146	0.7973
SWAD-inspired LoRA	0.1995	0.1447	0.0832	0.0127	**0.7986**
MoME-TTA	0.2221	0.1567	0.1001	0.0166	0.7973
MetaVIB	**0.2248**	0.1567	**0.1029**	**0.0167**	0.7977

3 Results

3.1 Quantitative Evaluation

Generated reports are evaluated using ROUGE-1, ROUGE-L [8], BLEU-1, BLEU-4 [11], and BERTScore F1 [17]. These metrics are widely used for text generation, but they are imperfect proxies for clinical report quality because they do not directly measure missed findings, unsupported findings, abnormality localization, or normal-template bias [15, 10].

Table 3 shows that all adapted models remain close to the source-only baseline. MetaVIB obtains the best point estimates for ROUGE-1, BLEU-1, and BLEU-4, but its absolute improvements over the source-only model are only 0.0035, 0.0036, and 0.0009, respectively. Vision-only LoRA obtains the best ROUGE-L score, improving the baseline by 0.0018, while MoME-TTA slightly improves BLEU-4 by 0.0008. In contrast, decoder-side LoRA variants reduce lexical overlap scores despite maintaining similar or slightly higher BERTScore F1.

The BERTScore F1 values range only from 0.7971 to 0.7986 across all models, suggesting that sentence-level semantic similarity is largely unchanged by the evaluated adaptations. Taken together, these results indicate that the tested lightweight strategies provide at most marginal gains under the IU X-Ray-to-CheXpert Plus shift. The main empirical finding is therefore not that one method clearly solves the shift, but that feature-level, adapter-level, and classification-oriented regularization objectives are insufficient by themselves for robust cross-dataset report generation.

3.2 Exploratory Qualitative Evaluation

To complement lexical metrics, we conducted an exploratory LLM-assisted qualitative review. A fixed set of 50 CheXpert Plus cases was sampled from the evaluation subset, and outputs from all models were compared against the same

reference reports in randomized order. The LLM was instructed to identify recurring semantic error patterns, including clinically important omissions, unsupported findings, abnormality mismatch, normal-template bias, and lack of finding specificity. This review was used only as a supplementary analysis and was not intended to replace expert radiologist assessment.

The qualitative review was not fully aligned with the automatic metrics. Vision-only LoRA achieved favorable ROUGE scores but often produced incomplete or generic reports. Vision–decoder LoRA and SWAD-inspired LoRA produced lower lexical overlap scores, yet their outputs were sometimes more clinically specific. KD–ASL LoRA reduced some unsupported statements but did not consistently improve semantic alignment. MoME-TTA and MetaVIB showed limited qualitative differences from the source-only baseline.

These observations reinforce the limitations of standard NLG metrics in this setting. A report that shares common words with the reference may still omit clinically important findings, while a more specific report may receive a lower lexical score if it uses target-domain phrasing different from the reference. Evaluation of domain-shift adaptation for radiology report generation should therefore include clinically grounded metrics or expert review in addition to BLEU, ROUGE, and BERTScore.

The LLM-based evaluation details and prompt are provided in Supplementary Tables S2 and S3.

4 Discussion

4.1 Why Training-Time Adaptation Was Limited

The limited gains from LoRA-based adaptation suggest that updating a small subset of visual or cross-attention parameters does not guarantee clinically meaningful target-domain adaptation. Vision-only LoRA slightly improves lexical overlap but leaves the decoder biased toward source-domain reporting style. Vision–decoder LoRA and SWAD-inspired LoRA modify the image-conditioned generation pathway, but the small target-domain training set may be insufficient to stabilize autoregressive decoding, explaining why these variants preserve BERTScore while reducing ROUGE and BLEU.

KD–ASL LoRA introduces additional supervision through pathology-label prediction, but such labels provide only coarse image-level guidance and do not directly supervise report structure, uncertainty, severity, or localization. Consequently, they may be too weak to correct the linguistic and decoder-level mismatch between IU X-Ray and CheXpert Plus.

4.2 Why Test-Time Adaptation Was Limited

MoME-TTA offers practical advantages because it uses unlabeled target-domain images and updates only the router. However, most TTA objectives were developed for classification tasks with finite label spaces [13]. In radiology report

generation, router updates can refine visual representations but do not explicitly optimize clinical content selection or target-domain language.

Autoregressive decoding further amplifies this mismatch. Small changes in the visual representation may alter early tokens, which then condition all subsequent tokens. Under severe domain shift, such perturbations may improve lexical overlap in some cases but introduce omissions or unsupported findings in others. Our results therefore suggest that TTA for medical report generation should be redesigned around sequence-level or clinically grounded adaptation signals, rather than directly importing classification-oriented entropy minimization.

4.3 Domain Gap Beyond Visual Features

The Mahalanobis analysis demonstrates strong visual separation between IU X-Ray and CheXpert Plus, but the adaptation results show that visual alignment alone is insufficient. Cross-dataset report generation also requires linguistic adaptation because IU X-Ray reports are generally shorter and more structured, whereas CheXpert Plus reflects different institutional phrasing and clinical description patterns.

This explains the discrepancy between quantitative and qualitative evaluations. Lexical metrics reward word overlap with a single reference report, while clinical utility depends on whether abnormalities are correctly mentioned, denied, localized, and qualified. Stable BERTScore values across methods suggest that broad semantic similarity may be preserved, but they do not ensure clinically faithful reporting. Future methods should therefore combine visual feature robustness with decoder adaptation and target-domain clinical language alignment.

4.4 Limitations and Future Work

This study has several limitations. First, experiments were intentionally conducted under constrained computation using a small evaluation subset of 201 images, so the results should be interpreted as point estimates. In particular, the relatively small evaluation set limits the statistical power to determine whether the observed differences between adaptation methods and the baseline reflect consistent effects or sampling variability. Future work should include larger patient-level test sets, confidence intervals, and statistical testing.

Second, we did not include a full-model fine-tuning baseline or train a model from scratch on CheXpert Plus. These experiments were beyond the computational budget of the present study. Consequently, our results cannot determine whether the limited gains of parameter-efficient adaptation arise from an inherent limitation of lightweight adaptation or instead reflect insufficient target-domain training, or the characteristics of the report-generation task itself. Thus, the present findings should not be interpreted as evidence that parameter-efficient adaptation is intrinsically inferior to full-model fine-tuning. A controlled comparison using the same target-domain subset would be necessary to establish

the performance–resource trade-off between parameter-efficient and full-model adaptation.

Third, the qualitative review was LLM-assisted and exploratory; expert radiologist review or validated clinical metrics are required before drawing clinical conclusions. Third, the evaluated methods use different supervision regimes: LoRA uses target reports, MoME-TTA uses unlabeled target images, and MetaVIB relies on source-domain pseudo-domains.

Finally, the MetaVIB pseudo-domain design also uses view position as a proxy for domain shift, which is weaker than true multi-institutional meta-learning. Future work should explore clinically grounded objectives, such as report-label consistency, entity-level alignment, uncertainty calibration, and decoder-side adaptation strategies that preserve factuality while adapting to target-domain report style.

5 Conclusion

We evaluated lightweight adaptation strategies for cross-dataset chest X-ray report generation by transferring an IU X-Ray-pretrained BiomedGPT model to CheXpert Plus. Although Mahalanobis analysis revealed a severe visual domain gap, LoRA-based adaptation, MoME-TTA, and MetaVIB produced only limited improvements in standard NLG metrics. Qualitative review further suggested that higher lexical overlap does not necessarily imply better clinical reporting quality.

These findings indicate that lightweight feature-level or classification-oriented adaptation alone did not produce substantial gains under our experimental conditions. These results highlight the difficulty of obtaining substantial improvements under limited target-domain data and computational resources, and motivate future comparisons between parameter-efficient and full-model adaptation. Effective cross-domain adaptation likely requires joint consideration of visual robustness, target-domain language, and decoder-level clinical semantics. This study therefore provides a resource-constraint diagnostic baseline for future MedAGI research on domain-robust medical vision-language generation.

Acknowledgments. This work was supported by grants from the National Research Foundation of Korea (NRF) funded by the Korean government (MSIT) (Nos. RS-2025-02215813, RS-2025-00520578, and RS-2025-02634603) and by an NRF grant funded by the Korean government (MSIT and MOE) (No. RS-2025-16063688).

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

References

1. Baruch, E.B., Ridnik, T., Zamir, N., Noy, A., Friedman, I., Protter, M., Zelnik-Manor, L.: Asymmetric loss for multi-label classification. CoRR **abs/2009.14119** (2020), https://arxiv.org/abs/2009.14119

2. Cha, J., Cho, H., Lee, K., Park, S., Lee, Y., Park, S.: Domain generalization needs stochastic weight averaging for robustness on domain shifts. CoRR **abs/2102.08604** (2021), https://arxiv.org/abs/2102.08604
3. Chambon, P., Delbrouck, J.B., Sounack, T., Huang, S.C., Chen, Z., Varma, M., Truong, S.Q., Chuong, C.T., Langlotz, C.P.: Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats (2024), https://arxiv.org/abs/2405.19538
4. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2016)
5. Du, Y., Xu, J., Xiong, H., Qiu, Q., Zhen, X., Snoek, C.G.M., Shao, L.: Learning to learn with variational information bottleneck for domain generalization. CoRR **abs/2007.07645** (2020), https://arxiv.org/abs/2007.07645
6. Hinton, G.E., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. ArXiv **abs/1503.02531** (2015), https://api.semanticscholar.org/CorpusID:7200347
7. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Chen, W.: Lora: Low-rank adaptation of large language models. CoRR **abs/2106.09685** (2021), https://arxiv.org/abs/2106.09685
8. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), https://aclanthology.org/W04-1013/
9. Mclachlan, G.: Mahalanobis distance. Resonance **4**, 20–26 (06 1999). https://doi.org/10.1007/BF02834632
10. Ostmeier, S., Xu, J., Chen, Z., Varma, M., Blankemeier, L., Bluethgen, C., Md, A.E.M., Moseley, M., Langlotz, C., Chaudhari, A.S., Delbrouck, J.B.: Green: Generative radiology report evaluation and error notation. In: Findings of the Association for Computational Linguistics: EMNLP 2024. p. 374–390. Association for Computational Linguistics (2024). https://doi.org/10.18653/v1/2024.findings-emnlp.21, http://dx.doi.org/10.18653/v1/2024.findings-emnlp.21
11. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Isabelle, P., Charniak, E., Lin, D. (eds.) Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (Jul 2002). https://doi.org/10.3115/1073083.1073135, https://aclanthology.org/P02-1040/
12. Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A.Y., Lungren, M.P.: Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. CoRR **abs/2004.09167** (2020), https://arxiv.org/abs/2004.09167
13. Wang, D., Shelhamer, E., Liu, S., Olshausen, B.A., Darrell, T.: Fully test-time adaptation by entropy minimization (2020), https://arxiv.org/abs/2006.10726
14. Wang, H., Yu, Y., Zheng, H., Zhang, T.: Test-time adaptation of medical vision-language models with mixture of modality experts. In: Proceedings of the 33rd ACM International Conference on Multimedia. p. 4649–4658. MM '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3746027.3755543, https://doi.org/10.1145/3746027.3755543

15. Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E.P., Fonseca, E.K.U.N., Lee, H.M.H., Abad, Z.S.H., Ng, A.Y., Langlotz, C.P., Venugopal, V.K., Rajpurkar, P.: Evaluating progress in automatic chest x-ray radiology report generation. Patterns **4**(9), 100802 (2023). https://doi.org/https://doi.org/10.1016/j.patter.2023.100802, https://www.sciencedirect.com/science/article/pii/S2666389923001575
16. Zhang, K., Zhou, R., Adhikarla, E., Yan, Z., Liu, Y., Yu, J., Liu, Z., Chen, X., Davison, B.D., Ren, H., Huang, J., Chen, C., Zhou, Y., Fu, S., Liu, W., Liu, T., Li, X., Chen, Y., He, L., Zou, J., Li, Q., Liu, H., Sun, L.: A generalist vision–language foundation model for diverse biomedical tasks. Nature Medicine **30**(11), 3129–3141 (Aug 2024). https://doi.org/10.1038/s41591-024-03185-2, http://dx.doi.org/10.1038/s41591-024-03185-2
17. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with BERT (2019), http://arxiv.org/abs/1904.09675