Clinical Reliability in Multilingual Chest X-ray Report Generation

Bilel Daoud[1], Ghazi ben Ghozlen[2], Yi Liu[1], Ali Khalfallah[2], and Shanli Ding[3]

[1] D3 Center, The University of Osaka, Osaka, Japan
bilel.daoud@is.ids.osaka-u.ac.jp
[2] National School of Electronics and Telecommunications of Sfax, University of Sfax, Sfax, Tunisia
[3] Graduate School of Biomedical Sciences, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

Abstract. Multilingual chest X-ray report generation requires more than fluent translation, because small changes in negation, uncertainty, anatomical location, or severity can alter clinical meaning. We study whether different multilingual generation pathways preserve clinical reliability for the same radiograph across languages. Three pathways are compared under a shared protocol: English report generation followed by translation, direct target-language generation, and LoRA-based adaptation of a medical vision–language model. The experiments use 2,000 chest X-ray studies and cover Arabic, Chinese, English, French, Hindi, Italian, Japanese, Korean, and Spanish. Assessment combines lexical metrics, ontology-normalized Clinical F1, reference-unsupported findings, cross-language consistency, and human clinical review. The adapted medical VLM achieves the highest Clinical F1 (0.901), lowest reference-unsupported rate (16.2%), and highest consistency score (0.886). Direct target-language generation is more reliable than English-first translation, but still shows language-dependent variation. These findings show that multilingual radiology reporting should be evaluated by both clinical correctness within each language and consistency of clinical content across languages.

Keywords: Chest X-ray report generation · Multilingual radiology · Medical vision-language models · Clinical reliability

1 Introduction

Chest X-ray is one of the most frequently performed imaging examinations, and its interpretation produces a large volume of radiology reports. Automatic report generation could help reduce this workload and support more consistent reporting across clinical settings. To be useful in practice, the generated report should accurately describe the image findings without changing clinically important details such as negation, uncertainty, anatomical location, laterality, or severity.

Most progress in chest X-ray report generation has been made in English. Public datasets such as MIMIC-CXR have supported model development [6], and evaluation tools such as CheXbert and RadGraph have helped move evaluation beyond word overlap [11,4]. Recent medical vision–language, multilingual generation, and translation models make multilingual report generation more feasible [1,12,8,10].

Generating reports in several languages adds another source of risk. A system may produce fluent text while changing the clinical meaning through translation choices, language-specific terminology, or differences in language-conditioned decoding. These failures can be difficult to see when evaluation is reported only as an average across languages, and they are especially important in radiology because small wording changes can alter the clinical interpretation [2].

In this work, we study how the choice of generation pathway affects chest X-ray reporting across languages. We compare English-first translation, direct target-language generation, and multilingual adaptation of CheXagent-8B across Arabic, Chinese, English, French, Hindi, Italian, Japanese, Korean, and Spanish. All systems are evaluated on the same patient-level splits, report format, target languages, and test cases, enabling a controlled empirical comparison of representative implementations of the three generation strategies. This study makes three research contributions. First, it formulates multilingual chest X-ray report generation as a clinical reliability problem, where the same image should lead to clinically consistent reports across languages. Second, it compares three representative generation pathways that connect image information to multilingual text in different ways. Third, it analyzes reliability beyond aggregate scores by measuring clinical correctness, reference-unsupported findings, harmful errors, and cross-language consistency.

2 Method

We evaluate three representative strategies for multilingual chest X-ray report generation under the same protocol. Given a chest radiograph I and a target language ℓ, each system generates one report $\hat{R}_\ell$. The target languages are Arabic (AR), Chinese (ZH), English (EN), French (FR), Hindi (HI), Italian (IT), Japanese (JA), Korean (KO), and Spanish (ES). All three systems use the same data partitions, image preprocessing, target languages, and report format. Language-balanced sampling is used during training. At inference time, each system receives the same chest X-ray and one language condition; the process is repeated for all nine languages. As illustrated in Fig. 1, the strategies differ in how image information is connected to multilingual text: (1) English report generation followed by translation, (2) direct generation with a radiology image encoder and multilingual decoder, and (3) parameter-efficient adaptation of an integrated medical vision–language model.

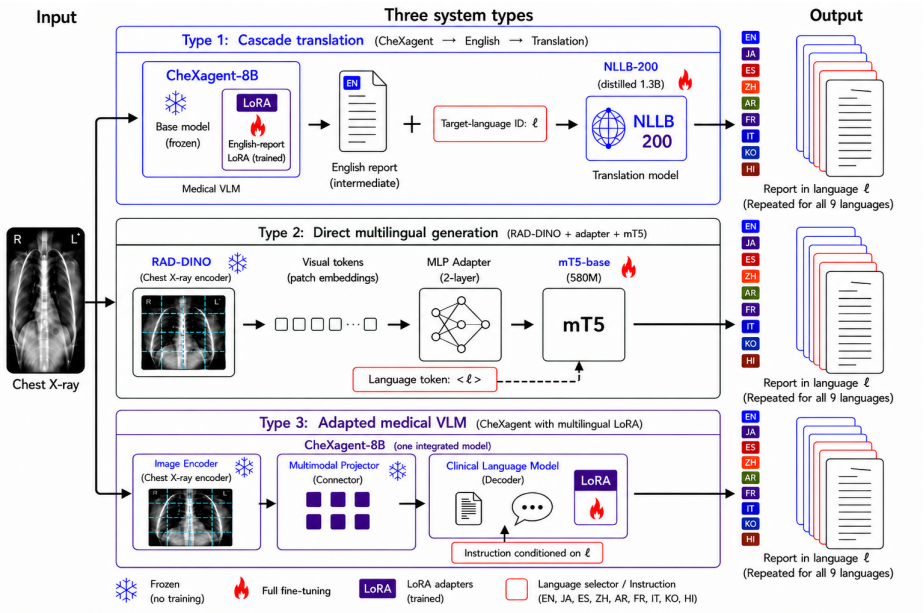

Fig. 1. Compared multilingual report-generation pathways. Type 1 generates an English report before translation; Type 2 directly decodes from frozen RAD-DINO features; Type 3 applies multilingual LoRA to the clinical language model of an otherwise frozen CheXagent-8B.

2.1 Type 1: Cascade Translation

The cascade system separates multilingual reporting into two steps: English report generation and translation. A LoRA-adapted CheXagent-8B [1,3] first generates an intermediate English report:

$$\hat{R}_{\mathrm{EN}} = G_{\mathrm{CXR}}(I).$$

The pretrained CheXagent parameters remain frozen, and LoRA modules inserted into selected language-model layers are optimized using English image–report pairs.

The generated English report is then translated by NLLB-200 distilled 1.3B [8]:

$$\hat{R}_{\ell} = T_{\mathrm{NLLB}}\left(\hat{R}_{\mathrm{EN}}, \ell\right),$$

where ℓ denotes the target-language identifier. NLLB-200 is fine-tuned with autoregressive cross-entropy on aligned medical report pairs to preserve clinically important expressions, including negation, severity, uncertainty, and anatomical location. Errors in the intermediate English report can propagate to all target languages.

2.2 Type 2: Direct Multilingual Generation

The second system generates the report in the requested language without producing an intermediate English report. RAD-DINO [10] first encodes the chest radiograph into patch-level visual representations:

$$F_I = E_{\text{RAD-DINO}}(I) = [f_1, \ldots, f_K].$$

RAD-DINO remains frozen during training. A two-layer MLP adapter maps the visual features to the hidden space of mT5-base [12]:

$$z_k = W_2\sigma\left(W_1 f_k + b_1\right) + b_2, \qquad Z_I = [z_1, \ldots, z_K].$$

The adapter is fully trainable.

A learned language token $< \ell >$ specifies the requested output language. This token and the adapted visual tokens are provided to mT5-base, which generates the report:

$$p\left(\hat{R}_\ell \mid I, \ell\right) = \prod_t p\left(r_{\ell,t} \mid r_{\ell,<t}, Z_I, < \ell >\right).$$

Both the MLP adapter and mT5-base are fine-tuned with token-level cross-entropy on multilingual image–report pairs. This avoids translation-induced errors, but the same image may still produce clinically inconsistent findings across languages.

2.3 Type 3: Multilingual Adaptation of a Medical VLM

The third system adapts CheXagent-8B for multilingual report generation. CheXagent is an integrated medical vision–language model composed of a chest X-ray vision encoder, a Q-Former-based multimodal bridge, a vision-to-language projection layer, and a Mistral-7B clinical language model [1,5]. The released CheXagent configuration uses Mistral-7B-v0.1 as its decoder-only language backbone and 128 learnable Q-Former query tokens for visual-language alignment [1].

Given a chest radiograph I, the vision encoder first extracts patch-level image representations:

$$F_I = E_{\text{img}}(I).$$

The Q-Former attends to these image features through a set of learned query tokens:

$$Q_I = Q_{\text{former}}\left(Q_0, F_I\right),$$

where Q_0 denotes the learnable query embeddings. A projection layer then maps the Q-Former outputs into the Mistral-7B embedding space:

$$Z_I = P_{\text{vl}}(Q_I).$$

To introduce multilingual generation while preserving the pretrained medical image–text alignment, the vision encoder, Q-Former, projection layer, and

Mistral-7B base parameters remain frozen. Multilingual LoRA adapters [3] are inserted into selected Mistral attention and projection layers and trained using the target-language reports prepared for training.

A language-conditioned instruction P_ℓ specifies the requested output language:

$$\hat{R}_\ell = D^{\mathrm{LoRA}}_{\mathrm{Mistral}}\left(Z_I, P_\ell\right).$$

Only the LoRA parameters are optimized using autoregressive token-level cross-entropy. This tests whether parameter-efficient adaptation can extend an English-oriented medical VLM while retaining pretrained visual-language alignment.

3 Experiments and Results

Experimental setting. We used 2,000 independent MIMIC-CXR studies [6], each containing a frontal chest radiograph and a human-authored English report. The data were split at the patient level into 1,600 training, 200 validation, and 200 test studies.

Initial target-language reference drafts were produced with GPT-4o-assisted translation, which was not used in any evaluated generation pipeline. Drafts were screened for finding, negation, uncertainty, severity, laterality, and location consistency, and inconsistent translations were corrected before training. Native-language reviewers checked 150 reports per language, and 50 challenging examples per language were medically reviewed from English-normalized findings. Linguistic ambiguities were adjudicated with the corresponding native-language reviewer.

All reported results were computed on test studies not used for training or validation. The evaluation involved 9 native-language reviewers and 5 medically qualified reviewers. Human evaluation used outputs for 25 test studies per language and strategy, resulting in 675 reports. Each report was first assessed by a native-language reviewer; clinical correctness, harmful errors, and substantial-edit decisions were then independently evaluated by two medically qualified reviewers using English-normalized findings. Clinical disagreements were resolved by consensus, and linguistic disagreements were adjudicated with the relevant native-language reviewer.

Implementation details. Images were resized to 224×224 pixels. Trainable modules were optimized with AdamW using a learning rate of 2×10^{-4}, weight decay of 10^{-4}, and a default batch size of 6. Training used mixed-precision computation on a single NVIDIA H200 141 GB GPU. Types 1–3 required approximately 14, 11, and 9 hours, respectively; Type 1 includes both CheXagent English-report LoRA adaptation and NLLB-200 fine-tuning. Checkpoints were selected by validation Clinical F1. CheXagent-based LoRA used rank 16 adapters, while NLLB-200 and mT5-base were fully fine-tuned. All systems used fixed seeds, deterministic split assignment, the same report prompt format, and beam search with beam size 4.

Table 1. Case-weighted performance across nine target languages. Best values are shown in bold.

Method	Clinical F1 ↑	Ref.-unsupported (%) ↓	Consistency ↑	BLEU ↑	ROUGE-L ↑
Cascade translation	0.858	21.7	0.824	0.247	0.401
Direct multilingual	0.881	18.6	0.857	0.262	0.419
Adapted medical VLM	**0.901**	**16.2**	**0.886**	**0.276**	**0.436**

Evaluation measures. Automatic evaluation combined lexical and clinical measures. BLEU and ROUGE-L measured text overlap [9,7]. For Clinical F1, reports were mapped to a shared English-normalized concept inventory using language-specific terminology dictionaries and deterministic negation and uncertainty rules, then evaluated with the same frozen CheXbert finding labeler for all languages [11].

Reference-unsupported findings were counted when a positive or uncertain generated finding was absent from the corresponding normalized reference report. We report this reference-based measure as the proportion of generated findings that were unsupported.

Cross-language consistency measured whether the same image produced similar normalized findings across target languages. Higher consistency indicates more stable clinical content across languages.

Human reviewers assessed linguistic acceptability, preservation of clinical meaning, clinical acceptability, harmful errors, and the need for substantial clinical editing. Confidence intervals and paired comparisons were computed with case-level bootstrap resampling.

Overall multilingual performance. Table 1 summarizes case-weighted results across the nine languages. The adapted medical VLM achieved the highest Clinical F1 (0.901), lowest reference-unsupported rate (16.2%), and highest consistency score (0.886). Direct multilingual generation improved over cascade translation in clinical correctness and consistency, suggesting that removing the intermediate translation stage reduced one source of clinical error.

Compared with direct multilingual generation, the adapted medical VLM improved Clinical F1 by 0.020 (95% CI: [0.006, 0.033], $p = 0.006$) and reduced reference-unsupported findings by 2.4 percentage points (95% CI: [0.7, 4.0], $p = 0.008$). Compared with cascade translation, it improved Clinical F1 by 0.043 (95% CI: [0.027, 0.058], $p < 0.001$) and reduced reference-unsupported findings by 5.5 percentage points (95% CI: [3.6, 7.4], $p < 0.001$).

Language-wise analysis. Table 2 reports Clinical F1 and reference-unsupported rate separately for each language. English yielded the strongest results for all three strategies, whereas Korean and Hindi were the most challenging. This pattern shows that aggregate scores alone do not fully describe multilingual behavior.

Cascade translation showed the largest language-dependent variation, with Clinical F1 ranging from 0.889 in English to 0.817 in Hindi. Direct multilingual generation improved both Clinical F1 and reference-unsupported rate in every language, with the largest gains in Arabic, Korean, and Hindi. The adapted

Table 2. Language-wise Clinical F1 / reference-unsupported rate (%).

Language	Cascade	Direct	Adapted VLM
EN	0.889 / 18.2	0.908 / 15.6	0.921 / 13.8
JA	0.861 / 21.0	0.885 / 18.1	0.904 / 15.8
ES	0.875 / 19.8	0.897 / 16.9	0.913 / 14.7
ZH	0.852 / 22.4	0.878 / 19.1	0.899 / 16.6
AR	0.838 / 24.0	0.868 / 20.6	0.891 / 17.8
FR	0.872 / 20.3	0.895 / 17.3	0.911 / 15.1
IT	0.869 / 20.6	0.892 / 17.5	0.908 / 15.3
KO	0.829 / 24.9	0.861 / 21.2	0.886 / 18.5
HI	0.817 / 26.1	0.846 / 22.8	0.876 / 19.8

Table 3. Human review on 25 test studies per language and strategy. Best values are shown in bold.

Method	Clinical accept. (%) ↑	Harmful error (%) ↓	Substantial edit (%) ↓
Cascade translation	73.5	10.8	27.1
Direct multilingual	79.4	7.6	21.3
Adapted medical VLM	**84.2**	**5.4**	**16.7**

medical VLM achieved the best result in every language and showed the smallest spread across languages.

Human evaluation and error analysis. Human review followed the automatic clinical metrics (Table 3). The adapted medical VLM achieved the highest clinical acceptability and the lowest harmful-error and substantial-edit rates. Cascade outputs were most often affected by terminology shifts and changes in negation or uncertainty. Direct multilingual generation reduced these failures but showed occasional omissions and language-dependent severity differences.

4 Discussion and Conclusion

This study shows that the pathway used for multilingual chest X-ray report generation affects clinical reliability. Cascade translation produced fluent text, but errors in the English report could be carried into all translated reports, and translation could further change negation, uncertainty, severity, or anatomical terminology. Direct multilingual generation reduced these translation-related errors, but language conditioning could still change the clinical content for the same image. Type 2 should therefore be interpreted as a modular encoder–decoder baseline, whereas Type 3 evaluates whether preserving an integrated medical vision–language alignment provides an advantage.

The results also show why aggregate multilingual scores are insufficient. They can hide weaker language-specific performance and do not reveal whether the same image receives consistent clinical descriptions. Reference-unsupported findings and cross-language consistency therefore provide complementary evidence that is not captured by lexical similarity alone.

Several limitations remain. The study used 2,000 MIMIC-CXR studies, which limits rare-finding coverage and language-specific precision. Non-English references were created through GPT-4o-assisted translation and clinical-consistency

screening, with expert review on subsets rather than the full corpus. The experiments were retrospective and used public data only. Reference-unsupported findings were also defined against the reference report, so image-supported findings absent from the reference could be counted as unsupported. Future work should extend this evaluation to larger external cohorts, additional report styles, and more complete native-language and medical review processes. In conclusion, direct multilingual generation was more clinically reliable than English-first translation, and multilingual adaptation of a medical VLM performed best across the nine languages.

Acknowledgments

This work was supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan, through the AI for Science Pioneering Research Creation Project (SPReAD), grant number 26263070.

References

1. Chen, Z., Varma, M., Xu, J.B., et al.: A vision-language foundation model to enhance efficiency of chest x-ray interpretation. arXiv preprint arXiv:2401.12208 (2024)
2. Dew, K., Turner, A.M., Choi, Y.R., Bosold, A., Kirchhoff, K.: Development of machine translation technology for assisting health communication: A systematic review. Journal of Biomedical Informatics **85**, 56–67 (2018)
3. 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)
4. Jain, S., Agrawal, A., Saporta, A., et al.: Radgraph: Extracting clinical entities and relations from radiology reports. In: NeurIPS Datasets Track (2021)
5. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., et al.: Mistral 7b. arXiv preprint arXiv:2310.06825 (2023)
6. Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., et al.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data **6**, 317 (2019)
7. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out (2004)
8. NLLB Team, Costa-jussà, M.R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., et al.: No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672 (2022)
9. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: A method for automatic evaluation of machine translation. In: ACL (2002)
10. Pérez-García, F., Sharma, H., Bond-Taylor, S., Bouzid, K., Salvatelli, V., Ilse, M., et al.: Exploring scalable medical image encoders beyond text supervision. arXiv preprint arXiv:2401.10815 (2024)
11. Smit, A., Jain, S., Rajpurkar, P., et al.: Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling. In: Findings of EMNLP (2020)
12. Xue, L., Constant, N., Roberts, A., et al.: mt5: A massively multilingual pre-trained text-to-text transformer. In: NAACL (2021)