CONrep: Improving Reliability of Vision-language Report Generation via Conformal Prediction

Shahriar Faghani, MD, Radiology Resident, University of Pennsylvania
Danial Elyassirad; Benyamin Gheiji; Mahsa Vatanparaast; Seyed Amir Asef Agah; Amir Mahmoud Ahmadzadeh, MD; Mana Moassefi, MD

Introduction

Automated report generation is one of the main applications of AI in medical imaging, driven by recent advances in vision-language models (VLMs). While these models demonstrate strong performance in generating reports, they typically lack reliable uncertainty estimates, which limits their safe and trustworthy deployment. To address this limitation, we propose CONRep, a conformal prediction (CP)-based framework that enhances the reliability of VLM-generated reports (VLMGRs) by providing uncertainty estimates.

Hypothesis

Integrating CP into VLMGRs improves their reliability by enabling uncertainty quantification and the identification of potentially unreliable outputs.

Methods

We applied CONRep using the Open-I chest x-ray dataset. The dataset was split into 30% calibration and 70% test sets. MedGemma was used to generate impression-level reports from images. Image-generated text similarity was quantified using cosine similarity, defined as the normalized dot product between image and text embeddings, computed with a contrastively trained VLM (BiomedCLIP). CP threshold was derived from calibration nonconformity scores using a hinge loss (1 – cosine similarity) with $\alpha = 0.05$, and subsequently applied to the test set to flag uncertain reports (cases with cosine similarity below the CP threshold). Performance was evaluated by assessing the correlation between image-VLMGR and ground truth (GT)-VLMGR cosines. In addition, GT-VLMGR similarity across certain, uncertain, and highly uncertain test subsets was compared statistically.

Results

From 2562 test cases, 1736 cases VLMGR flagged as certain, 547 as uncertain, and the remaining as highly uncertain. The cosine similarity of image-VLMGR has a significant correlation with the cosine of GT-VLMGR in the test set (p-value < 0.001). Furthermore, similarity between the GT and generated impression was significantly higher in certain cases (p-value < 0.001).

Conclusion

CONRep provides an effective, posthoc and model-agnostic approach for quantifying uncertainty in VLMGRs, supporting more reliable and trustworthy deployment of automated reporting systems.

Figure(s)

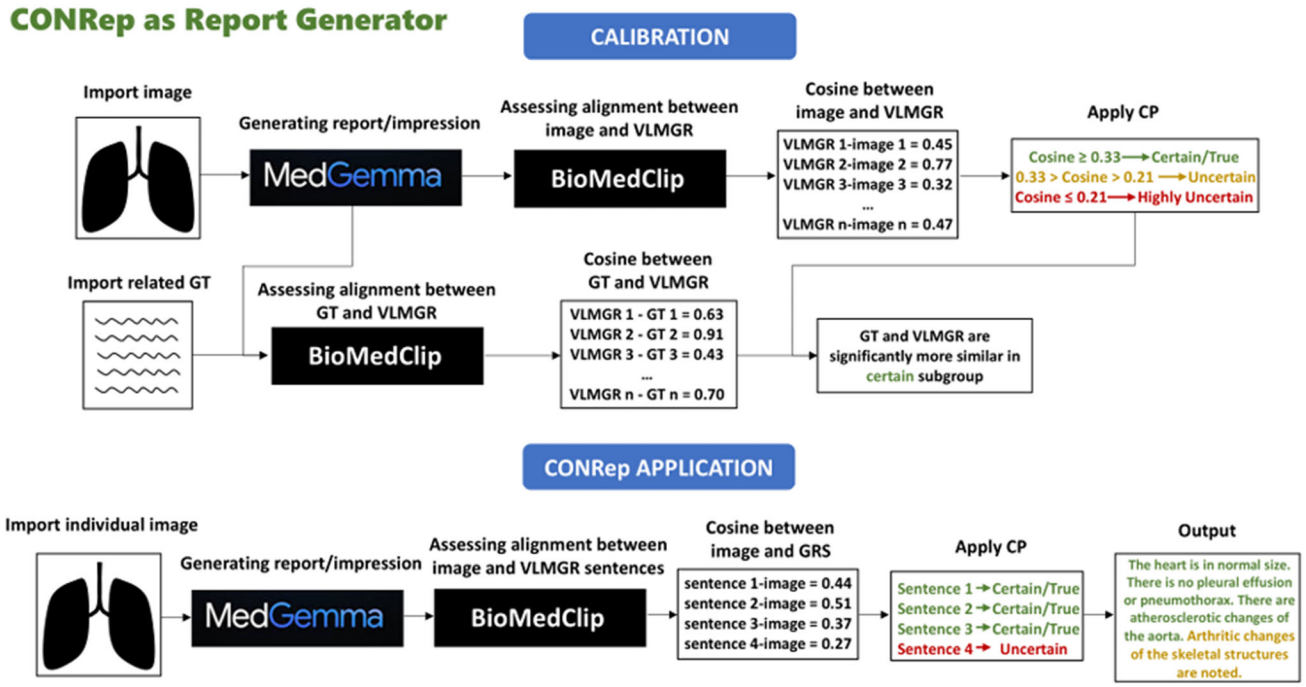
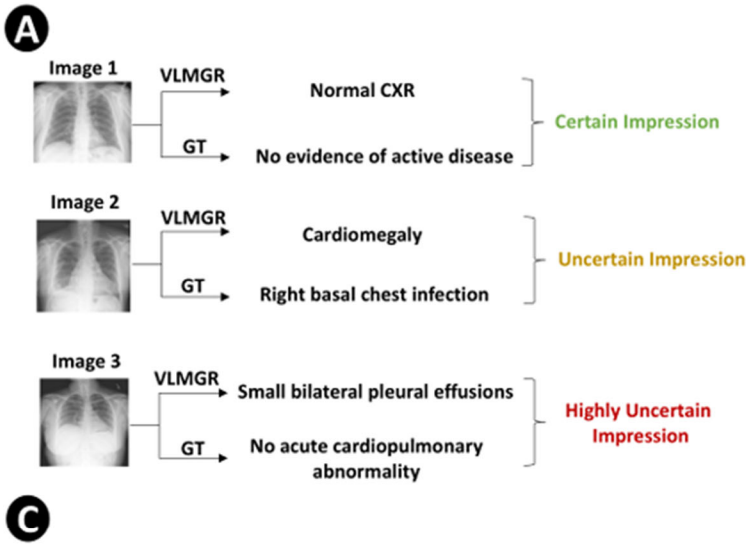
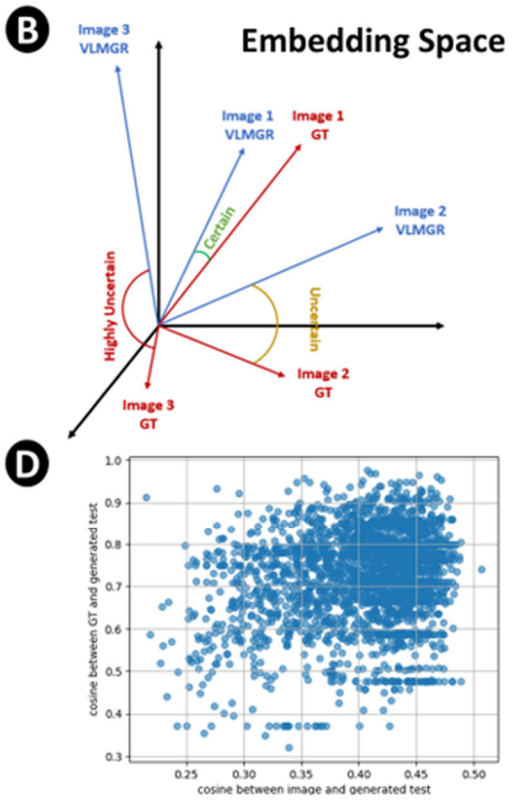


Figure 1. Study workflow. Images are input to a VLM for report generation. Image-VLMGR alignment is measured via cosine similarity using a contrastively trained VLM. CP thresholds derived from calibration data are applied to test samples to identify certain, uncertain, and highly uncertain reports.





C

	Certain	Uncertain	Highly Uncertain
N cases	1736	547	279
VLMGR-GT similarity	0.74	0.71	0.66

Figure 2. Qualitative and quantitative evaluation of CONRep. A) Representative examples of certain, uncertain, and highly uncertain cases, showing the input image, GT, and the VLMGR. B) Schematic illustration of image-text similarity computation using cosine similarity by a contrastively trained VLM. C) Comparison of similarity between GT and generated reports (VLMGR) across test-set subgroups (certain, uncertain, and highly uncertain), demonstrating significantly higher agreement in certain cases. D) Correlation analysis between cosine similarity of GT-VLMGR text pairs and cosine similarity of image-VLMGR pairs in the test set (correlation = 0.19, P value <0.001).

Keywords

Applications; Artificial Intelligence; Emerging Technologies; Imaging Research; Machine Learning