Indication Driven Autoregressive Report Generation for Cardiac Magnetic Resonance Imaging

Makiya Nakashima NAKASHM2@CCF.ORG
Cardiovascular Innovation Research Center, Cleveland Clinic

Po-Hao Chen CHENP2@CCF.ORG
Diagnostics Institute, Cleveland Clinic

Michael Bolen BOLENM@CCF.ORG
Diagnostics Institute, Cleveland Clinic

Christopher Nguyen NGUYENC6@CCF.ORG
Cardiovascular Innovation Research Center, Cleveland Clinic

W. H. Wilson Tang TANGW@CCF.ORG
Heart Vascular and Thoracic Institute, Cleveland Clinic

Richard Grimm GRIMMR@CCF.ORG
Heart Vascular and Thoracic Institute, Cleveland Clinic

Deborah Kwon KWOND@CCF.ORG
Heart Vascular and Thoracic Institute, Cleveland Clinic

David Chen CHEND3@CCF.ORG
Cardiovascular Innovation Research Center, Cleveland Clinic

Abstract

Interpreting and documenting findings from cardiac imaging studies is increasingly burdensome to readers in part due to the increasing amount of advanced cardiac imaging studies which capture multi-parametric data. This is particularly true of cardiac magnetic resonance imaging (CMR) studies which encode features of morphology, function, flow, parametric mapping, and myocardial viability in multiple 2D planes, but require a substantial amount of time to analyze, document, and integrate the numerous complex imaging features into a comprehensive report. Additionally, clearly communicating complex CMR findings and diagnoses to referring physicians with varying CMR knowledge and the ability to clinically correlated complex CMR findings is highly variable. Automatic interpretation and generation of the report have great potential to reduce the burden on readers and improve access through higher patient throughput. As such, there has been significant work in this area, although much of it has been focused on more simplistic chest X-ray and single view echocardiography. These data sources are represented by only a single view or have only a single source of contrast, greatly reducing the necessary complexity of the latent visual space. Furthermore, we recognize that clinical histories are important for accurate reporting. In this work, we propose to treat the CMR study as a multi-scene video and generate the corresponding report in an autoregressive manner. We further warm-start the generated report with the indications for the exam to improve the relevance of the generated report. We validate our model on two closed CMR datasets from two different sites and demonstrate that our model offers significant improvements on both language generation metrics and human reader preference.

Keywords: report generation, video captioning, clinical imaging, multimodal learning

Data and Code Availability The data used to support this research is not openly available due to privacy restrictions on healthcare data. The data can be made available to interested researchers through an established Data Use Agreement with our institution.
Code developed for this work will be made publicly available at https://github.com/Makiya11/

CMR-TAGET. The code is available for review as supplemental material at the time of submission.

Institutional Review Board (IRB) This study was approved by our institution's IRB with the study number 22-1213.

1. Introduction

Conveying information between radiologists and clinicians is integral to successfully managing of patients. Although it has been said that "a picture is worth a thousand words", the report must clearly address the clinical question raised by the treating clinician (Wallis and McCoubrie, 2011). Comprehensively detailing all visualized abnormalities without providing clear decision support can lead to miscommunications and mistakes in interpreting the study results. This is particularly true for cardiac magnetic resonance imaging (CMR) given it's complexity and information density. CMR contains multi-dimensional information of cardiac function, morphology, viability, flow, perfusion, and tissue characterization(Kramer et al., 2020). Although there have been significant efforts to provide detail guidelines or produce structured reports to enforce reporting standards, the human nature of interpreting medical images and recording them in text introduces several areas for potential inconsistencies and miscommunications. Furthermore, producing these reports is time-consuming (Ehrenfeld and Wanderer, 2018), a primary cause of radiologist burnout (Pourvaziri et al., 2022), and a factor that limits patient access to advanced diagnostic services.

Automated report generation can play a role in reducing the clerical burden on radiologists and improving patient care. Several works have been shown for chest X-ray Liu et al. (2019); Boag et al. (2020); Kong et al. (2022); Zhou et al. (2021). However, chest X-ray is a relatively simple imaging modality with 1 or 2 views of the heart, akin to image caption generation. In contrast, the typical CMR exam contains 100s if not over 1000 images. Interpreting CMR is both challenging and time consuming given the need to extract numerous quantitative measurements and synthesize information from multiple different image types and views of the heart into a single diagnosis. Therefore, not only do the images provide a wide variety of diagnostic and prognostic information for several cardiac diseases, but the associated reports are also similarly more complex.

To address these issues, we propose CMR-Transformer-based AutoRegressive GEnerator for Text (CMR-TARGET) combined with indication prompting. The proposed model leverages CMR-former, a video-text multimodal transformer model based on the TimeSformer (Bertasius et al., 2021) architecture to co-train the image and text encoders (Qiu et al., 2023; Nakashima et al., 2023), which incorporates the most common CMR views (short axis, 4 chamber long axis, and 2 chamber long axis) and image types (Cine and Late Gadolinium Enhancement (LGE)) in a single embedding. The model further uniquely incorporates clinical history by including the indication for the study as an input, thereby warm-starting the report.

1.1. Related Works

X-Ray Report Generation Report generation is often cast as an image captioning task (Yuan et al., 2019; Xue et al., 2018; Jing et al., 2017) with encoder-decoder architectures. For instance, Yuan et al. (2019) used a CNN encoded with a LSTM decoder for report generation. More recent works have leveraged a unified Transformer architecture with the popularization of Transformer based models. Chen et al. (2020) proposed an end-to-end Transformer architecture augmented with a memory module to encode prior generated sentences to minimize repeated concepts. Wang et al. (2022b) introduced a memory module to inform more semantically relevant generation through a lexigraphic dictionary. In contrast, TranSQ proposed by Kong et al. (2022) treats the report as a retrieval task, selecting relevant sentences from a dictionary. Such works have achieved moderate success in chest X-ray applications where the input is primarily limited to one or two images.

Video Captioning However, CMR report generation is more akin to video captioning where the input is a sequence of spatial-temporally related images. The advent of deep image embeddings was quickly applied to videos through recurrent temporal layers (Donahue et al., 2017; Li et al., 2015). Transformers were once again applied to video captioning in order to provide feature localization between two independently trained image and text embeddings (Zanfir et al., 2017; Yan et al., 2021). Current video captioning now takes advantage of large public video and text data sources such as WebVid-10M (Bain et al., 2021) to jointly train the image and text embeddings (Seo et al., 2022). Generative text decoders achieve

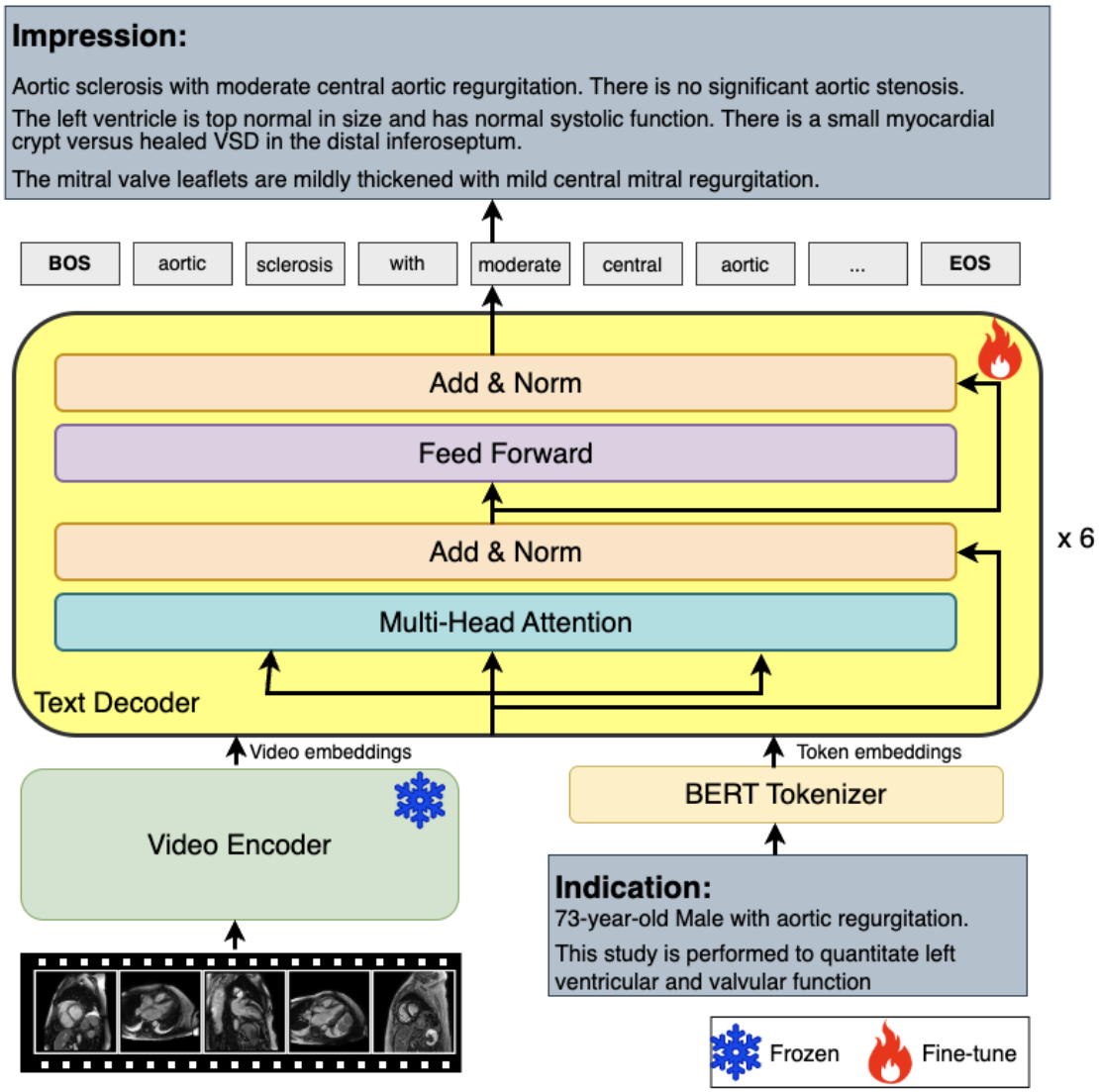

Figure 1: CMR-TARGET approach. The 2D CMR images are concatenated together into a video format and passed through a pre-trained CMR domain specific space-time vision transformer. The resulting embeddings are concatenated with the tokens from the indications and passed through the text generator, which produces the impressions of the report autoregressively.

more human-like descriptions compared to early template based models (Yan et al., 2021). However, many of these tasks require large amounts of open domain data. Furthermore, these frameworks have not been applied to healthcare tasks.

Our work extends upon both these fields, combining the video capabilities in video captioning with domain specific learning implicit of X-Ray report generation tasks.

2. Methods

2.1. Model

We denote the CMR studies and the associated impressions as (X_n, Y_n) where the impressions are a sequence of tokens $\mathbf{Y} = (y_1, ..., y_k)$, where k is the number of tokens in an impression. The images of the CMR studies are processed through a video-transformer V (detailed in Qiu et al. (2023)) result-

ing in an embedding v_l. The embeddings are then decoded through our text decoder T. Specific details the encoders V and T are described below.

Visual encoder CMR studies are a sequence of 2D images representing 3D, 4D, or sometimes even 5D data. Instead of developing separate models for each image type, we concatenate the images together in the sequence $X = (x_1, ..., x_i)$ where i is the total number of images in that study. This formulation takes inspiration from movies which are sequences from frames grouped into "scenes". The CMR study is passed into the visual encoder (V) which is comprised of a visual tokenizer and space-time transformer. Here, each image of the study is broken down into patches as typical of visual transformers (Dosovitskiy et al., 2020). Each patch is also encoded with a learned spatial and temporal embedding. The spatial and temporal relationships are finally learned using space-time transformer blocks (Bertasius et al., 2021). The visual embedding of the study is derived from the final ['CLS'] token of the final block, similar to language transformer models(Devlin et al., 2018). This visual encoder is pretrained using contrastive image-language training on CMR data as detailed in Qiu et al. (2023).

Text generator For the text generator, we take inspiration from image captioning work (Wang et al., 2022a). The Generative Image-to-text Transformer (GIT) model uses 6 transformer blocks, with each block consisting of a self-attention layer and a feed-forward layer. The text is tokenized using the standard BERT tokenizer(Devlin et al., 2018). The input into the generator starts with the video embeddings from the last layers of the video encoder V. Here, CMR-TARGET differs where instead of just the video embeddings, we also concatenate the entire sequence of the tokens of the indications to the video embeddings. The generated text then starts with the ['BOS'] (beginning of sentence) token, and continues to be decoded in an auto-regressive manner until either the ['EOS'] (end of sentence) token is encountered or the maximum number of steps is reached. A seq2seq attention mask (Bahdanau et al., 2014) is used to allow the output to attend to the image embedding, the indication tokens, and previously generated text tokens.

2.2. Training

The model is trained using cross-entropy loss (CE) for next-word prediction (1). Z represents the image embeddings and indication text tokens. y_j denotes the tokens in the sequence that have already been generated. The training process includes special tokens y_0 and y_{N+1}, which correspond to the ['BOS'] and ['EOS'] markers, respectively.

$$\mathcal{L} = \frac{1}{N+1} \sum_{k=1}^{N+1} \mathrm{CE}\left(y_k, p(y_k \mid Z, y_j)\right), j = 0, \ldots, k-1 \tag{1}$$

The visual encoder was pre-trained through contrastive language video pretraining (Bain et al., 2021; Qiu et al., 2023) in the training dataset and its weights were frozen. The text generator was fine-tuned for the generation task.

2.3. Autoregressive Generation

The autoregressive beam search algorithm is utilized to generate report sentences in a sequential manner, where each word is predicted based on the indication words, image embeddings, and all previously generated words. Beam search explores multiple potential continuations at each step to select the most probable sequence. For this setup, the beam size is set to 1, and the maximum generation length is set to 100 steps.

3. Experiments

3.1. CMR Dataset

The dataset contains CMR studies from a single health system in both inpatient and outpatient settings between the 2008 and 2022. The total size of the initial cohort was 63,637 CMR studies from 48,976 unique patients. This dataset includes a wide range of protocols and vendors. We chose to pare-down this dataset by identifying studies that are complete for left ventricular views (short axis (SAX), 4 chamber (4CH), 2 chamber long axis (2CH), and 3 chamber long axis (3CH)), and both cine and late gadolinium enhancement (LGE) image types. The dataset was then divided at the patient level into training, validation, and testing subsets, with 70% allocated for training, 15% for validation, and 15% for testing. The resulting dataset includes 9,705 training studies, 2,100 validation studies, and 2,094 testing stud-

ies. We also evaluated our model on a dataset of 411 studies from satellite hospitals which had different patient populations, vendors of scanners, imaging protocols, and reporting standards. Of note, the imaging protocol in our training dataset used post-contrast cine images whereas the satellite hospitals used a pre-contrast cine protocol, which does impact image characteristics (Tang et al., 2023).

The associated report was pulled from our electronic health records. The reports are parsed into 3 main sections (indications, findings, and impressions) using a rule-based parser. We used the impressions and indications section for this work as findings often a significant amount of tabular data and impressions represented a more clinically relevant synthesis of the information contained in the exam.

All images were resized and normalized to a matrix size of 224x224 and spatial resolution of $1.5mm^2$. Non-square images were zero-padded to achieve a square matrix. The pixel values were normalized to a range of 0-1. For the impressions text, we removed all text which were under 25 words long. All numerals and stop words were removed to reduce redundancy and focus the model on the most clinically relevant terms.

3.2. Evaluation

We compare the text generated by our proposed model with others both quantitatively and qualitatively. First, we use 5 common quantitative metrics of evaluate the generated text: BiLingual Evaluation Understudy (BLEU) (Papineni et al., 2002), automated metric for machine translation evaluation (METEOR) (Banerjee and Lavie, 2005), Recall-Oriented Understandy for Gisting Evaluation-Longest common sequence (ROUGE-L) (Lin, 2004), and Consensus-based Image Description Evaluation (CIDEr) (Vedantam et al., 2015), and BERTscore (Zhang et al., 2019). The BLEU metric is the geometric mean of n-grams which appear in the reference and the generated text, reflecting content and short context. The METEOR metric is similar to the BLEU metric but with higher weighting for recall and penalty for poor alignment. The ROUGE-L metric reflects sentence-level similarity. CIDEr measures the cosine similarity of Term Frequency-Inverse Document Frequency n-grams, increasing reward for infrequency terms. Finally, BERTscore compares sentence-level and system-level similarity by comparing the BERT embeddings of the generated report and the ground truth.

Next, we also identify the clinical completeness of our generated text through named entity recognition (NER). We utilized the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) (Donnelly et al., 2006) to identify differential cardiac diagnoses found in the generated and reference impression. We focus on a few cardiovascular entities important for CMR: aortic regurgitation, cardiomyopathy, hypertrophic cardiomyopathy, mitral regurgitation, pericardial effusion, and tricuspid regurgitation. The accuracy of these entities was evaluated by the F1 score.

Finally, we asked a level 3 CMR physician to provide human feedback on the preferred generated text. We select 20 random impressions in the test set and have our reader select their preference compared to the reference impression. The reader could also select "no preference" if none of the generated impressions were superior to any other.

3.3. Quantitative Results

Automated comparison of generated reports The results of statistical measures of our proposed report generation model are presented in Table 1. Our approach outperforms multiple previously described methods across all metrics except for CIDEr. The base CMR-TARGET model outperform the next best model by 16.1% averaged across all metrics. The BLEU-n and ROUGE-L metrics, which are statistical correlations of word distributions, show CMR-TARGET generates more of the words from the original report compared to other report generation options. CMR-TARGET is also better in BERTscore which reflects that the overall semantic content of the generated reports are more similar to the ground truth than just a few specific entities. The poor performance of the Retrieval method would also suggest that the decoder has learned more than just to retrieval reports which are similar to the queried study.

Factual comparison of generated reports Furthermore, the proposed method is more factually accurate compared to other baselines, as shown in Table 2. The base model was able to achieve the highest F1 scores in 3 of the 5 clinical entities with the Aortic regurgitation being a close 2nd. M^2-Trans model's aortic regurgitation, tricuspid regurgitation, or pericardial effusion resulted in an F1 score of 0. This outcome is due to the model generated the most common

Table 1: Metrics comparing the generated impression sentences to the ground truth clinical impression. CMR-TARGET exceeds other state-of-the-art models across multiple metrics. The model also generalizes better to unseen sites, although there is still significant degradation in performance.

Dataset	Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDERr	BERTscore
Internal	Retrieval	0.2966	0.1682	0.1063	0.0739	0.1219	0.2219	0.0632	0.6483
	M^2 Trans	0.3000	0.2197	0.1738	0.1439	0.1603	0.3490	0.2924	0.7050
	CvT-21 DistillGPT2	0.3372	0.2427	0.1895	0.1559	0.1683	0.3501	0.3446	0.7141
	CMR-TARGET	0.3754	0.2679	0.2057	0.1655	0.1815	0.3640	0.3248	0.7226
	CMR-TARGET-Ind	**0.4102**	**0.3050**	**0.2417**	**0.1997**	**0.2015**	**0.3978**	**0.4912**	**0.7435**
External	Retrieval	0.1948	0.0847	0.0399	0.0210	0.0913	0.1658	0.0386	0.6145
	M^2 Trans	0.2289	0.1168	0.0676	0.0448	0.0906	0.2065	0.1352	0.6293
	CvT-21 DistillGPT2	0.2216	0.1002	0.0493	0.0282	0.0805	0.1758	0.0678	0.6146
	CMR-TARGET	**0.2704**	**0.1377**	0.0762	0.0484	**0.1040**	0.2086	0.1173	0.6346
	CMR-TARGET-Ind	0.2603	0.1371	**0.0841**	**0.0594**	0.1037	**0.2097**	**0.2391**	**0.6363**

entities or phrases in the training data, rather than accurately predicting the less frequent name entities.

Impact of warm-starting with indications CMR-TARGET-Ind, which was warm-started with the indications for the study, consistently outperforms both the baseline models and the base CMR-TARGET. The model outperformed the next best baseline model by 22.9% averaged across all metrics. It also outperformed the base CMR-TARGET by 10.8%. This is reflected in the clinical accuracy where it achieved the best F1 score in 4/5 entities except for mitral regurgitation. In particular, it further improved results in pericardial effusion. Pericardial effusion has specific clinical indications compared to other ventricular or valvular diseases evaluated here. Therefore, commenting on the finding is relatively rare outside of when it is specifically requested. Warm-starting the report generation is important for accurate generation of clinical entities in the impressions.

External validation All models did not translate as well in the external dataset with respect to automated metrics, although CMR-TARGET and CMR-TARGET-Ind both consistently outperformed baseline models in the external dataset. The drop of performance is not as evident in the clinical accuracy, where the F1 scores are relatively similar. The formatting of the external CMR reports are significantly different from the training data. The training data had a semi-structured format which did not support dictation. In contrast, the external validation set did support dictation and therefore, tended to be more unscripted and train-of-thought. However, the similarity in F1 scores of clinical entities would suggest the actual content of impressions was not impacted.

3.4. Human Evaluation

Table 3 shows the results of the CMR physician's manual evaluation of the generated reports and Table 4 show representative examples. The grading criteria is in 5. Both the base CMR-TARGET and CMR-TARGET-Ind achieved higher scores than other report generation models, achieving an average score of 2.4 and 2.7 respectively. Furthermore, both models had more reports graded as 3 or above than any other model (7 and 10 respectively). However, none of the models were able to consistently achieve reports which had minimal errors and no significant omissions, corresponding with the low F1 scores and low automated evaluation scores. The reviewer picked the report generated by our model as the best one among the generated reports in 45% of the cases. In 4 cases, all models generated reports which were completely wrong. Of note, including the indication for the study did force CMR-TARGET-Ind to answer the clinical question more often than if it only had the image embeddings as inputs (CMR-TARGET).

4. Discussion

This work proposes CMR-TARGET and CMR-TARGET-Ind models which automatically generate radiology report impressions from CMR studies. The model takes inspiration from human interpretation. First, it treats the study as a spatially and temporally correlated set of videos, reflecting the underlying anatomy's inherent structure. This could be particularly useful for CMR given a single view of the heart can be imaged multiple times with different image contrasts, visualizing different aspects of morphology and/or physiology. Second, the model leverages

Table 2: Accuracy of important findings contained in generated reports compared to original. CMR-TARGET more consistently generates these findings compared to other techniques.

Location	Finding	# of cases	Retrieval	M^2 Trans	CvT-21	CMR-TARGET	CMR-TARGET-Ind
Internal	Aortic regurgitation	172	0.4941	0.	0.4782	0.4832	**0.5934**
	Mitral regurgitation	549	0.5136	0.4967	0.6425	**0.6584**	0.6535
	Tricuspid regurgitation	138	0.5076	0.	0.5308	0.5178	**0.5362**
	Hypertrophic cardiomyopathy	549	0.5120	0.7336	0.8143	0.8319	**0.8325**
	Pericardial effusion	241	0.4854	0.	0.4940	0.6878	**0.7234**
External	Aortic regurgitation	6	0.5107	0.	0.	0.4938	**0.5438**
	Mitral regurgitation	16	0.4428	0.4869	0.4840	0.4868	**0.5042**
	Tricuspid regurgitation	12	0.5044	0.	0.5456	**0.6024**	0.4882
	Hypertrophic cardiomyopathy	14	0.4817	0.7001	0.4491	**0.7372**	0.5900
	Pericardial effusion	20	0.4663	0	0.4869	0.6934	**0.7109**

Table 3: Human scoring of the generated impressions. Higher numbers indicate more factual correctness and syntactically pleasing. The impressions were scored on a 1-5 scale.

Model	Picked best report	Average Score	Cases > 2
Retrieval	0	1.1	1
M^2	1	1.8	4
CvT21	3	2.1	5
CMR-TARGET	3	2.4	7
CMR-TARGET-Ind	**9**	**2.7**	**10**

warm-starting for the generated impressions using the indication for the study. This indication can be seen as an anchor for which to necessitate the inclusion of certain clinical entities within the impression, similar to how radiologists often read prior clinical history to get insight on potential findings. We show the model achieves better results compared to other report generation frameworks in both automated metrics and human evaluation. The method outperforms the next best method (CvT-2 (Nicolson et al., 2023)) by 22.8%.

One issue limiting the development of AI techniques for CMR is the lack of data. Unlike other radiological tests with publicly available datasets containing 10,000-100,000 instances (echocardiogram and chest X-ray respectively), CMR is performed at much lower volumes clinically. For comparison, in a six year period between 2012 and 2017, Medicare recorded 89.5k CMR exams (Goldfarb and Weber, 2021). In the same period, Medicare recorded over 5 million echocardiograms. This difference in the volume of data makes developing AI tools for CMR difficult. However, we were able to achieve similar automated text generation metrics compared to those in public chest X-ray data, despite multiple orders of magnitude less data. This speaks towards the importance of pre-training the vision encoder.

Many radiology report generation frameworks treat the problem similar to image captioning in that the only input necessary for accurate interpretation is the image itself. However, this is not true in practice as clinical history often changes the interpretation (Gunderman et al., 2001). Many studies have shown that access to an accurate clinical history or indication for exam improves the interpretation accuracy and reporting confidence (Castillo et al., 2021). Furthermore, radiology reports (among other clinical tests) do not happen in a vacuum. Rather, they must answer a specific clinical question as dictated by the patient's health status. Therefore, the clinical history is important to writing a relevant report (Obara et al., 2015), which often includes not only the diagnosis but also ideally includes a possible plan of care or plan for further diagnostics. This work uniquely shows that even a short snippet of clinical information can be used to improve report generation results.

There have been many recent innovations towards general purpose multimodal generative models such as Video-LLaMA (Zhang et al., 2023), Macaw-LLM (Lyu et al., 2023), and Video-ChatGPT (Maaz et al., 2023). These models use a similar setting where the input of the model is a combination of video and text, with the output being some description of the video clip in question. There are three major differences between such general video-text models and our proposed model. First and foremost, general image embeddings typically do not do well in healthcare tasks (Huix et al., 2024). Medical images comprise of very little of the training data used for general purpose foundation models. Of the medical data that is in-

Table 4: Examples of generated reports from different models.

Technique	Impression
Ground Truth	Severe septal hypertrophic cardiomyopathy resulting near obliteration lv cavity. Mild sam mild lvot flow acceleration rest patchy minimal midmyocardial enhancement consistent fibrosis. Normal qualitative rv structure systolic function. Normal structure function aortic mitral tricuspid valves.
Retrieval	Hypertrophic cardiomyopathy asymmetric septal hypertrophy. No evidence myocardial fibrosis delayed enhancement imaging. Papillary muscles appear thickened apically displaced no evidence abnormal attachments. Mild mitral regurgitation.
M^2 Trans	Hypertrophic cardiomyopathy moderate asymmetric septal hypertrophy normal left ventricular size systolic function no regional wall motion abnormalities no evidence myocardial fibrosis delayed enhancement imaging. papillary muscles appear normal no evidence abnormal thickening attachments. Systolic anterior motion mitral valve mitralseptal contact resulting significant flow acceleration within left ventricular outflow tract moderate mitral regurgitation. Normal right ventricular size systolic function qualitative assessment.
CvT-21	Left ventricle normal size shape normal function no discrete myocardial fibrosis delayed enhancement imaging suggest infiltrative process prior myocardial ischemic injury. Right ventricle appears normal size systolic function qualitative assessment . No significant valvular abnormalities.
CMR-TARGET	Hypertrophic cardiomyopathy asymmetric septal hypertrophy normal left ventricle size systolic function mild delayed gadolinium enhancement midmyocardial distribution affecting basal inferoseptum consistent interstitial fibrosis pattern typical hcm. Papillary muscles appear multiheaded no evidence abnormal thickening attachments systolic anterior motion mitral valve mitralseptal contact resulting significant flow acceleration within left ventricular outflow tract mild mitral regurgitation. Normal right ventricle size systolic function. Normal aortic tricuspid valve function.
CMR-TARGET-Ind	Hypertrophic cardiomyopathy moderate asymmetric septal hypertrophy described mild chordal systolic anterior motion mitral valve mild associated flow acceleration within left ventricular outflow tract. Moderate mitral regurgitation mild moderate left atrial enlargement. Mild moderate delayed gadolinium enhancement midmyocardial distribution affecting basal anteroseptum consistent interstitial fibrosis pattern typical hcm. Papillary muscles appear multiheaded without abnormal thickening attachment.

cluded, it is often scraped from public sources such as PubMed or Wikipedia which do not conform with clinical practice standards. CMR is also a relatively rarely used modality, representing less than 1% of cardiac imaging volume. Second, the generated outputs are semantically and syntactically oriented towards layperson speak. This may not be acceptable in clinical practice which involves highly specific jargon. Finally, the outputs answer a single specific question: "What part of the heart is being shown?" Whereas reports need to be much more comprehensive, commenting on several features in the exam.

4.1. Limitations

Although our method represents a significant improvement in report generation accuracy, there is still significant work to get results to a clinically useful state. As shown in Table 3, most of the generated reports still contain significant errors which may affect diagnosis. Furthermore, our results do not include certain CMR imaging types such as flow, parametric mapping, perfusion, etc. Although these imaging types are important, their clinical use is not pervasive and would result in a significant amount of missing

Table 5: Grading criteria for radiologist review of generated reports.

Score	Criteria
1	Superb accuracy, quality, report language similar or better than the clinical report.
2	Good accuracy, with minimal errors or language imprecision that does not impact the imaging findings clinical significance. No significant omissions.
3	Moderate accuracy, with some errors or language imprecision that is distracting but still communicates the clinically pertinent findings. Some missing components, but not critical to diagnosis or management decisions.
4	Limited accuracy with significant errors or missing imaging findings, but with some correct features.
5	Completely wrong findings, or completely illogical language.

data. Their inclusion also requires a larger vision encoder which is not possible with our limits in computational power. Finally, the indications for the test was taken from the ground truth reports themselves. This potentially requires the radiologist to manually summarize history, making CMR-TARGET-Ind's potential impact in practice limited. However, we will explore using the indications from the radiology order which is often recorded in the electronic health record.

Overall, this work provides evidence that CMR reports can be automatically generated using a combination of space-time encoder and an auto-regressive text generator which was warm-started using the indication for the study. Such tools are necessary to reduce radiologist documentation burden given the increasing workloads, particularly with respect to CMR which is a highly complex imaging modality. Automated report generation may also have the potential to increase access given that CMRs require highly specialized training to interpret, thereby largely limiting its use to academic healthcare centers.

5. Bibliography

Acknowledgments

We would also like to thank Dr. Richard Grimm and Charles and Loraine Moore Endowed Chair in Cardiovascular Imaging for funding the powerful infrastructure necessary for this work and the Software Development/Data Science team in Imaging Informatics for facilitating transfer of imaging studies to our computational platform.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. URL https://aclanthology.org/W05-0909.pdf.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. Issue: 3.

William Boag, Tzu-Ming Harry Hsu, Matthew McDermott, Gabriela Berner, Emily Alesentzer, and Peter Szolovits. Baselines for chest x-ray report generation. In *Machine learning for health workshop*, pages 126–140. PMLR, 2020. ISBN 2640-3498.

Chelsea Castillo, Tom Steffens, Lawrence Sim, and Liam Caffery. The effect of clinical information on radiology reporting: a systematic review. *Journal of Medical Radiation Sciences*, 68(1):60–74, 2021.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Trevor Darrell. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):677–691, April 2017. ISSN 1939-3539. doi: 10.1109/TPAMI.2016.2599174.

Kevin Donnelly et al. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279, 2006.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, and Sylvain Gelly. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Jesse M. Ehrenfeld and Jonathan P. Wanderer. Technology as friend or foe? Do electronic health records increase burnout? *Current Opinion in Anesthesiology*, 31(3), 2018. ISSN 0952-7907.

James W Goldfarb and Jonathan Weber. Trends in cardiovascular mri and ct in the us medicare population from 2012 to 2017. *Radiology: Cardiothoracic Imaging*, 3(1):e200112, 2021.

Richard B Gunderman, Micheal D Phillips, and Mervyn D Cohen. Improving clinical histories on radiology requisitions. *Academic radiology*, 8(4): 299–303, 2001.

Joana Palés Huix, Adithya Raju Ganeshan, Johan Fredin Haslum, Magnus Söderberg, Christos Matsoukas, and Kevin Smith. Are natural domain foundation models useful for medical image classification? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7634–7643, 2024.

Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. *arXiv preprint arXiv: 08195*, 2017.

Ming Kong, Zhengxing Huang, Kun Kuang, Qiang Zhu, and Fei Wu. TranSQ: Transformer-Based Semantic Query for Medical Report Generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 610–620. Springer, 2022.

Christopher M. Kramer, Jörg Barkhausen, Chiara Bucciarelli-Ducci, Scott D. Flamm, Raymond J. Kim, and Eike Nagel. Standardized cardiovascular magnetic resonance imaging (CMR) protocols: 2020 update. *Journal of Cardiovascular Magnetic Resonance*, 22(1):17, February 2020. ISSN 1532-429X. doi: 10.1186/s12968-020-00607-1.

Guang Li, Shubo Ma, and Yahong Han. Summarization-based video caption via deep neural networks. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1191–1194, 2015.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation. In *Machine Learning for Healthcare Conference*, pages 249–269. PMLR, 2019. ISBN 2640-3498.

Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093*, 2023.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

Makiya Nakashima, Donna Salem, HW Wilson Tang, Christopher Nguyen, Tae Hyun Hwang, Ding Zhao, Byung-Hak Kim, Deborah Kwon, and David Chen. Reducing contextual bias in cardiac magnetic resonance imaging deep learning using contrastive self-supervision. In *Machine Learning for Healthcare Conference*, pages 473–488. PMLR, 2023.

Aaron Nicolson, Jason Dowling, and Bevan Koopman. Improving chest X-ray report generation by leveraging warm starting. *Artificial Intelligence in Medicine*, 144:102633, 2023. ISSN 0933-3657. doi: https://doi.org/10.1016/j.artmed.2023.102633. URL https://www.sciencedirect.com/science/article/pii/S0933365723001471.

Piotr Obara, Merlijn Sevenster, Adam Travis, Yuechen Qian, Charles Westin, and Paul J Chang. Evaluating the referring physician's clinical history and indication as a means for communicating chronic conditions that are pertinent at the point of radiologic interpretation. *Journal of digital imaging*, 28:272–282, 2015.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Ali Pourvaziri, Anand K. Narayan, David Tso, Vinit Baliyan, McKinley Glover, Bernardo C. Bizzo, Bashar Kako, Marc D. Succi, Michael H. Lev, and Efren J. Flores. Imaging Information Overload: Quantifying the Burden of Interpretive and Non-Interpretive Tasks for Computed Tomography Angiography for Aortic Pathologies in Emergency Radiology. *Current Problems in Diagnostic Radiology*, 51(4):546–551, July 2022. ISSN 0363-0188. doi: https://doi.org/10.1067/j.cpradiol.2022.01.008.

Jielin Qiu, Peide Huang, Makiya Nakashima, Jaehyun Lee, Jiacheng Zhu, Wilson Tang, Pohao Chen, Christopher Nguyen, Byung-Hak Kim, Deborah Kwon, Douglas Weber, Ding Zhao, and David Chen. Multimodal Representation Learning of Cardiovascular Magnetic Resonance Imaging. In *Machine Learning for Multimodal Healthcare Data*, Honolulu, 2023. Springer.

Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretraining for multimodal video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17959–17968, 2022.

Lu Tang, Kaiyue Diao, Qiao Deng, Xi Wu, Pengfei Peng, Xun Yue, Tao Wu, Wei Cheng, Yangjie Li, Xiaoyue Zhou, Jens Wetzl, Yucheng Chen, Wenjun Yue, and Jiayu Sun. Comparison between pre- and post-contrast cardiac MRI cine images: the impact on ventricular volume and strain measurement. *The International Journal of Cardiovascular Imaging*, 39(5):1055–1064, May 2023. ISSN 1875-8312. doi: 10.1007/s10554-023-02809-x. URL https://doi.org/10.1007/s10554-023-02809-x.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. URL http://openaccess.thecvf.com/content_cvpr_2015/html/Vedantam_CIDEr_Consensus-Based_Image_2015_CVPR_paper.html.

A. Wallis and P. McCoubrie. The radiology report — Are we getting the message across? *Clinical Radiology*, 66(11):1015–1022, November 2011. ISSN 0009-9260. doi: https://doi.org/10.1016/j.crad.2011.05.013.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022a.

Zhanyu Wang, Mingkang Tang, Lei Wang, Xiu Li, and Luping Zhou. A Medical Semantic-Assisted Transformer for Radiographic Report Generation. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 655–664, Cham, 2022b. Springer Nature Switzerland. ISBN 978-3-031-16437-8.

Yuan Xue, Tao Xu, L. Rodney Long, Zhiyun Xue, Sameer Antani, George R. Thoma, and Xiaolei Huang. Multimodal recurrent model with attention for automated radiology report generation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pages 457–466. Springer, 2018. ISBN 3-030-00927-0.

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, pages 721–729. Springer, 2019. ISBN 3-030-32225-4.

Mihai Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Spatio-temporal attention models for grounded video captioning. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part IV 13*, pages 104–119. Springer, 2017. ISBN 3-319-54189-7.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Yuanen Zhou, Yong Zhang, Zhenzhen Hu, and Meng Wang. Semi-autoregressive transformer for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3139–3143, 2021.