Dual-View Vision Transformer with Cross-Modal Memory for Radiology Report Generation

Noureddine LAMRED[1], Norhene GARGOURI[2], and Nesrine CHARFI[3]

***Abstract*— Automatically generating clinical narratives from chest radiographs is an active area of study designed to alleviate radiologist workload and standardize diagnostic reporting.Existing pipelines predominantly encode visual content through Convolutional Neural Networks (CNN), whose locality constraints inherently limit the modelling of anatomically dispersed pathological patterns.To address these limitations, we propose a dual-view framework that integrates a pre-trained Vision Transformer (ViT-B/16) with a Transformer encoder-decoder architecture augmented by a Cross-Modal Memory Network (CMN). Unlike prior memory-based approaches that rely on single-view CNN encoders, our framework simultaneously processes frontal and lateral chest X-ray projections through a shared ViT backbone, enabling richer cross-view global representations. The CMN incorporates a shared learnable memory matrix consulted by both the visual encoder and the textual decoder, effectively bridging the semantic gap between visual patch embeddings and specialized medical terminology. Extensive experiments on the MIMIC-CXR benchmark demonstrate that our framework achieves a B-4 of 0.135, a R-L of 0.302, and a CIDEr of 0.306, surpassing all comparable methods that share the same resource constraints, without the use of external knowledge repositories or foundation model decoders.**

***Index Terms*— Radiology report generation, Vision Transformer, Cross-modal memory, Chest X-ray, MIMIC-CXR, Medical image analysis.**

I. INTRODUCTION

Developing computational methods to automatically draft clinical narratives from radiographic images is an active area of intersection between computer vision and medical informatics. Such automation holds the potential to improve reporting consistency and relieve the diagnostic throughput pressure faced by radiological departments [1]–[3]. Most contemporary solutions rely on standard encoder-decoder frameworks, frequently utilizing Convolutional Neural Networks to extract visual features before decoding them into text via Transformer modules [4]–[7].Despite their established efficacy in extracting localized texture patterns, convolutional encoders remain fundamentally constrained in aggregating inter-regional context, a property critical for interpreting findings distributed across the thoracic cavity.Moreover, standard radiological protocols systematically acquire both frontal and lateral projections, as the combined reading of complementary views is essential for complete lesion characterization yet many existing automated systems process only a single view, inevitably missing crucial diagnostic context. Vision Transformers address these spatial limitations by computing global attention across distinct image patches. However, mapping these complex, non-local visual representations to highly specific medical terminology remains a significant challenge. Relying solely on standard cross-attention mechanisms often leads to generic or repetitive phrasing, as models struggle to consistently anchor rare clinical concepts to their corresponding visual markers. To bridge this semantic gap, we propose a multimodal architecture that combines a dual-view Vision Transformer with a Cross-Modal Memory Network and a sequence-to-sequence Transformer. While the CMN mechanism has been previously explored in CNN-based, single-view contexts, our primary contribution lies in adapting and integrating this memory structure to process global patch embeddings derived from multiple simultaneous radiographic projections. This targeted integration provides the textual decoder with a shared semantic space, enabling the retrieval of specific clinical prototypes rather than reverting to statistically common templates. Empirical evaluation on the MIMIC-CXR dataset [2] demonstrates that this architecture yields reliable structural clarity and clinical accuracy. The primary contributions of this research are summarized as follows:

- We propose a dual-view framework that replaces traditional local CNN backbones with a global Vision Transformer, simultaneously processing frontal and lateral X-rays to maximize spatial context.
- We integrate a Cross-Modal Memory Network into this dual-view pipeline to bridge the semantic divide between global ViT embeddings and specialized medical vocabulary.
- We demonstrate that the targeted combination of multi-view global attention and memory-driven decoding provides robust clinical alignment on the MIMIC-CXR benchmark, offering an effective alternative without requiring external knowledge bases or large-scale foundation models.

II. RELATED WORK

The earliest automated reporting architectures adapted image-to-text captioning pipelines, coupling CNN visual backbones [8], [9] with sequential language decoders based

[1]Noureddine LAMRED is with the National School of Electronics and Telecommunications of Sfax (ENET'Com), Signals systeMs aRtificial intelligence and neTworkS Laboratory (SM@RTS), Digital Research Center of Sfax, Sfax, Tunisia noureddine.lamred@gmail.com

[2]Norhene GARGOURI is an Associate professor with Signals systeMs aRtificial intelligence and neTworkS Laboratory (SM@RTS), Digital Research Center of Sfax, Sfax, Tunisia norhene.gargouri@gmail.com

[3]Nesrine CHARFI is an Assistant professor with the Higher Institute of Technological Studies of Kef (ISET Kef), Kef, Tunisia. charfi.nesrine@gmail.com

on gated recurrent units. [10], [11].Although these methods established proof of concept for vision-to-language medical report synthesis, the convolutional inductive bias prevented adequate modelling of global anatomical context essential for holistic thoracic evaluation. The field subsequently transitioned toward Transformer-based architectures [4], [5], improving long-range linguistic coherence. More recently, large-scale Vision-Language Foundation Models such as BioViL-T [12], CheXagent [13], and MedVersa [14] have achieved strong generalization through massive pre-training, yet demand extreme computational resources that limit their practical deployment in clinical settings.

Patch-based Vision Transformers [6] address this limitation by decomposing radiographs into fixed-size tokens and computing pairwise dependencies across the entire image through self-attention, capturing distributed pathological patterns more effectively than CNN-based encoders. To further bridge the semantic gap between visual features and clinical vocabulary, Parameterized memory structures have been proposed to accumulate reusable semantic prototypes that mediate correspondence between radiographic representations and diagnostic vocabulary. [15], [16]. However, prior works apply such memory matrices exclusively to single-view, CNN-extracted features. Our framework advances this paradigm by deploying the CMN as an alignment mechanism for a dual-view ViT encoder, grounding generated text in a comprehensive multi-perspective spatial context, achieving strong clinical fidelity without the overhead of billion-parameter foundation models.

III. METHODOLOGY

We present an end-to-end trainable architecture designed for structured clinical narrative synthesis from paired chest radiographic projections, based on a Vision Transformer encoder and a Transformer-based decoding unit integrated with a Cross-Modal Memory Network . As depicted in Figure 1, our system processes a pair of chest X-ray scans , specifically frontal and lateral projections to produce a structured clinical narrative. The architecture is divided into two main phases: (A) a Visual Feature Extraction stage that utilizes a Dual-View Vision Transformer with a bi-view integration strategy to map global spatial relationships and (B) a Cross-Modal Generation stage, employing a Transformer Encoder-Decoder structure enhanced by a Cross-Modal Memory Network to ensure accurate alignment between visual evidence and medical language.

A. Visual Feature Extraction

Given a pair of chest X-ray images $I_f, I_l \in \mathbb{R}^{3\times224\times224}$ representing the frontal and lateral views respectively, we employ a Vision Transformer pre-trained on ImageNet-21k as our visual encoder. For each image, the ViT encoder splits the input into $N = 196$ non-overlapping patches of size 16×16 pixels and processes them through 12 transformer blocks. The output consists of a class token x_{cls} representing global image semantics, and N patch tokens representing local visual regions:

$$\{x_{cls}, x_1, x_2, \ldots, x_N\} = \text{ViT}(I) \in \mathbb{R}^{(N+1)\times768} \quad (1)$$

Each patch token is projected to a higher-dimensional space via a linear projection layer followed by Layer Normalization and GELU activation:

$$P = \text{GELU}(\text{LN}(\text{Linear}(x_{1:N}))) \in \mathbb{R}^{N\times d_{vf}} \quad (2)$$

where $d_{vf} = 2048$. The patch features from both views are concatenated to form the final visual representation:

$$\text{att\_feats} = [P_f; P_l] \in \mathbb{R}^{2N\times d_{vf}} \quad (3)$$

Concatenating both projection streams enables the encoder to capture anatomical evidence that is exclusively visible from a single perspective, enriching the composite visual representation, which is important for detecting subtle pulmonary abnormalities.

B. Cross-Modal Memory Transformer

The concatenated visual features are projected to $d_{model} = 512$ via a linear embedding layer, then processed by a Transformer encoder-decoder architecture augmented with a Cross-Modal Memory Network, as proposed in [15]. The CMN consists of a learnable memory matrix $M \in \mathbb{R}^{C\times d_{model}}$ shared between the visual and textual modalities, where $C = 2048$ is the memory size. For visual features, A top-k attention based retrieval step selects the most semantically relevant memory entries prior to the encoding pass.:

$$\tilde{P}^* = \tilde{P} + \text{CMN}(\tilde{P}, M, M) \quad (4)$$

The CMN operates via a top-k attention mechanism that selects the $k = 32$ most relevant memory slots for each query, allowing efficient retrieval without attending over the full memory:

$$\text{CMN}(Q, K, V) = \text{softmax}\left(\text{top-}k\left(\frac{QK^\top}{\sqrt{d_k}}\right)\right)V \quad (5)$$

The same memory matrix is consulted by the textual decoder embeddings before decoding:

$$e_t^* = e_t + \text{CMN}(e_t, M, M) \quad (6)$$

where e_t is the token embedding at step t. This shared memory bridges the visual and textual modalities, encouraging alignment between visual patches and medical terms. **Encoder:** The enriched visual features $\tilde{P}^*$ are encoded through $L = 3$ layers of multi-head self-attention and feed-forward networks:

$$\mathcal{H} = \text{Encoder}(\tilde{P}^*) \in \mathbb{R}^{2N\times d_{model}} \quad (7)$$

Decoder: The decoder generates the report token by token using masked self-attention over previously generated tokens and cross-attention over the encoded visual memory $\mathcal{H}$:

$$\alpha_t = \text{softmax}\left(\frac{Q_t K^\top}{\sqrt{d_k}}\right) \in \mathbb{R}^{1\times2N} \quad (8)$$

$$o_t = \alpha_t V \quad (9)$$

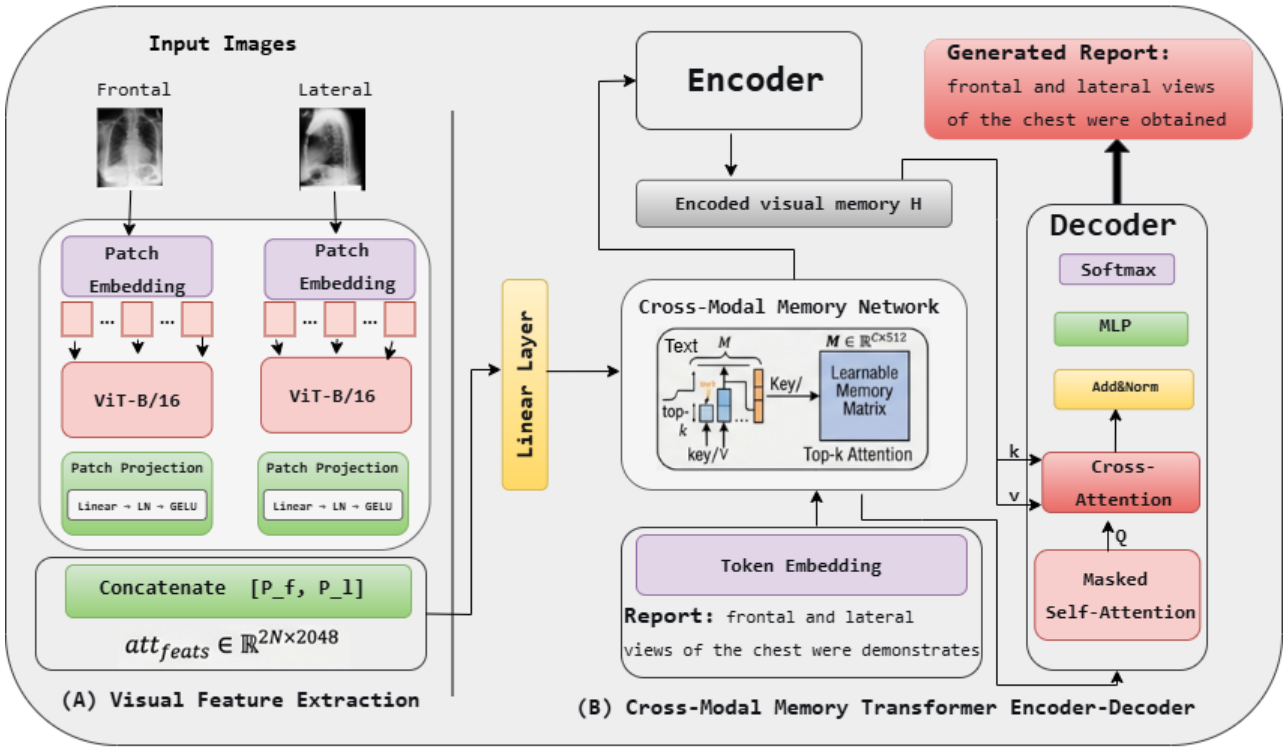

Fig. 1. Overview of the proposed pipeline for radiology report generation.

where Q_t, K, and V are derived from the decoder hidden state and encoder output $\mathcal{H}$ respectively.

Training Objective: The model is trained with the standard cross-entropy loss over the target report tokens:

$$\mathcal{L} = \mathcal{L}_{CE} = -\sum_{t=1}^{T} \log P(w_t \mid w_{<t}, \mathcal{H}) \tag{10}$$

IV. EXPERIMENTAL SETUP

A. Datasets

We evaluate the proposed framework on the widely used MIMIC-CXR benchmark [2], which contains 377,110 chest radiographs and 227,835 associated reports from 64,588 patients. Following established preprocessing protocols [4], [15], only the *findings* section is retained as the target text. To support the proposed dual-view architecture, we further filter the dataset to include only studies containing both frontal and lateral chest X-ray projections. The resulting dataset distribution is presented in Table I.

TABLE I
MIMIC-CXR DATASET SPLITS (AFTER DUAL-VIEW FILTERING)

Split	Patients	Reports	Image Pairs
Training	46252	46252	92504
Validation	13213	13213	26426
Testing	6613	6613	13226
Total	**66078**	**66078**	**132156**

B. Evaluation Metrics

The quality of the generated reports is assessed using both standard Natural Language Generation (NLG) criteria and clinical efficacy metrics. For NLG, we report BL (1-4) [17] for n-gram precision, R-L [18] for sequence-level structural similarity, and CIDEr [19] for TF-IDF weighted consensus, which is particularly effective in penalizing generic phrases and rewarding specific pathological terminology. To ensure diagnostic reliability, we additionally evaluate clinical correctness using the CheXpert labeler to compute Precision, Recall, and F1-score over 14 pathological observations.

C. Implementation Details

The visual feature extraction relies on a ViT-B/16 model pre-trained on ImageNet-21k. During training, the ViT weights were fine-tuned to ensure optimal feature adaptation. Patch embeddings are projected from $d_{vit} = 768$ to $d_{vf} = 2048$. The encoder-decoder network consists of 3 layers with 8 attention heads, a hidden size $d_{model} = 512$, and a feed-forward size $d_{ff} = 512$. The cross-modal memory matrix M is configured with $N_m = 2048$ memory slots. Optimization is performed using the Adam optimizer with disparate learning rates to ensure stable convergence: 5×10^{-5} for the visual backbone and 1×10^{-4} for the Transformer components. A StepLR scheduler with a 0.1 decay factor is applied. The model is trained for a maximum of 100 epochs with an early stopping patience of 10 epochs.Given the memory demands of processing concurrent dual-projection inputs at full resolution on a single GPU, mini-batches of 4 studies were used throughout training. During inference, reports

are decoded using beam search with a beam width of 3 and trigram blocking. Tokens appearing fewer than 3 times are filtered, yielding a vocabulary of 4,756 distinct tokens including special symbols.The complete model comprises 105.3M trainable parameters and processes a dual-view chest X-ray study in approximately 0.14 seconds on a single NVIDIA Tesla T4 GPU 15.7 GB.

V. RESULTS AND DISCUSSION

A. Comparison with State-of-the-Art

Table II compares our framework with recent state-of-the-art methods on the MIMIC-CXR benchmark. The proposed dual-view ViT architecture combined with the Cross-Modal Memory Network achieves a B-4 of 0.135, a R-L of 0.302, and a CIDEr score of 0.306, demonstrating competitive performance among recent CNN-based and ViT-based approaches without relying on external knowledge bases or large language models.

Compared with earlier memory-driven Transformer models such as R2Gen [4] and R2G-Cross [15], our framework consistently improves B-4 and R-L scores, indicating better alignment between visual features and generated clinical descriptions. In comparison with the recent ViT-based DM-RRG framework [23], our method achieves higher B-4 and R-L performance while maintaining a lightweight architecture.

Furthermore, the pronounced CIDEr gain reflects the capacity of the shared memory to steer token prediction toward pathology-specific vocabulary, reducing reliance on generic reporting templates. Although PromptMRG [22] achieves a slightly higher B-1 score, it relies on prompt-driven supervision and larger language modeling components.

Overall, these results indicate that combining dual-view global visual representations with memory-augmented decoding provides a lightweight framework with consistent performance across both linguistic and clinical metrics for radiology report generation.

B. Clinical Efficacy Evaluation

Beyond standard NLG metrics, we evaluate the clinical reliability of the generated reports through Clinical Efficacy (CE) metrics computed over the 14 CheXpert pathological observations (Table II). Our framework achieves a CE Precision of 0.533, a CE Recall of 0.414, and a CE F1-score of 0.466. These results surpass all CNN-based competitors, demonstrating that the CMN module not only improves linguistic fluency but also enhances diagnostic accuracy. Compared to the baseline without CMN (F1=0.275), the addition of the memory network yields a substantial gain of $+0.191$ in CE F1-score, confirming that the shared memory matrix effectively bridges the semantic gap between visual anomalies and appropriate clinical vocabulary. It is noteworthy that while PromptMRG and DM-RRG achieve competitive CE F1 scores (0.476 and 0.486 respectively), these methods rely on external prompting strategies or larger backbone architectures. Our approach achieves comparable clinical precision with a significantly more lightweight and reproducible design. The slightly lower CE Recall of our model (0.414 vs. 0.509 for PromptMRG) reflects a conservative generation behaviour, which is a known limitation of attention-based decoding on complex multi-pathology studies. Improving recall on such cases remains an important direction for future work.

C. Ablation Study

To validate the contribution of each component, we conduct an ablation study analyzing the architecture design, CMN hyperparameters, and decoder settings.

1) Impact of Architecture and Input Strategy: Table III evaluates the contribution of the dual-view strategy and the CMN module using three configurations: (1) Dual-View ViT with a standard Transformer encoder decoder, (2) Single-View ViT with CMN, and (3) the complete Dual-View ViT + CMN framework.

Configuration (1) achieves a B-4 of 0.106 and a CIDEr score of 0.199, showing that global patch attention alone remains insufficient for precise clinical grounding. Introducing the CMN in Configuration (2) improves B-4 to 0.121 and increases the CE F1-score from 0.275 to 0.445, confirming that the memory module plays an important role in improving clinical alignment.

The complete architecture further improves B-4 to 0.135 and CIDEr to 0.306 by combining complementary frontal and lateral information with CMN-based alignment. The CE F1-score reaches 0.466, while recall increases from 0.395 to 0.414, indicating that the lateral view provides additional pathological cues not always visible in frontal projections.

2) Sensitivity to CMN Hyperparameters: Table V reports the impact of the memory size C and the top-k retrieval parameter.

Memory size C. Configurations with $C < 2048$ obtain similar B-4 scores (0.121), suggesting limited memory capacity for representing diverse clinical patterns. Increasing the memory size to $C = 4096$ reduces performance (B-4=0.113, CIDEr=0.268), likely due to noisy or redundant memory slots. The best performance is obtained with $C = 2048$.

Top-k retrieval. Values of $k \in \{8, 16\}$ produce identical B-4 scores of 0.121, whereas $k = 32$ achieves the best results with B-4=0.135 and CIDEr=0.306. Retrieving too few memory slots limits semantic diversity, while excessive retrieval introduces irrelevant clinical prototypes.

3) Impact of Decoder Settings: Table V also evaluates beam size and Transformer depth.

Beam size. Greedy decoding (Beam Size = 1) underperforms with B-4=0.084 and CIDEr=0.150 due to limited exploration during sequence generation. Beam Size = 3 achieves the best performance, whereas Beam Size = 5 slightly degrades results, likely because of repetitive decoding patterns.

Network depth. A shallow encoder-decoder architecture ($L = 1$) already benefits from CMN integration, achieving B-4=0.121. The best overall performance is obtained with

TABLE II

RESULTS ON MIMIC-CXR DATASET

Method	Year	Encoder	B-1	B-2	B-3	B-4	R-L	CIDEr	CE Precision	CE Recall	CE F1-score
R2Gen [4]	2020	ResNet-101	0.353	0.218	0.145	0.103	0.277	–	0.333	0.273	0.276
R2G-Cross [15]	2021	CNN	0.353	0.218	0.148	0.106	0.278	–	0.334	0.275	0.278
DCL [20]	2023	CNN	–	–	–	0.109	0.284	0.281	0.471	0.352	0.373
Modal Alignment [21]	2023	ResNet-101+BERT	0.386	0.237	0.157	0.111	0.274	0.111	0.420	0.339	0.352
PromptMRG [22]	2024	ResNet-101	0.398	–	–	0.112	0.268	–	0.501	0.509	0.476
CAMANet [7]	2024	DenseNet121	0.374	0.230	0.155	0.112	0.279	0.161	0.483	0.323	0.387
DM-RRG [23]	2026	Swin Transformer	0.405	–	–.	0.114	0.275	–	0.508	0.523	0.486
Ours	**2026**	**ViT-B/16**	**0.377**	**0.247**	**0.177**	**0.135**	**0.302**	**0.306**	**0.533**	**0.414**	**0.466**

Note:"–" denotes metrics not reported in the source paper.

TABLE III

ARCHITECTURAL ABLATION: IMPACT OF BACKBONE, VIEWS, AND CMN

Configuration	B-1	B-2	B-3	B-4	R-L	CIDEr	CE Precision	CE Recall	CE F1-score
(1) Dual-View ViT + Transformer	0.321	0.204	0.141	0.106	0.301	0.199	0.494	0.191	0.275
(2) Single-View ViT + Transformer + CMN	0.340	0.226	0.161	0.121	0.301	0.293	0.511	0.395	0.445
(3) Ours (Dual-View ViT + Transformer + CMN)	**0.377**	**0.247**	**0.177**	**0.135**	**0.302**	**0.306**	**0.533**	**0.414**	**0.466**

TABLE IV

QUALITATIVE EVALUATION OF GENERATED REPORTS VS. GROUND TRUTH. BLUE HIGHLIGHTS CORRECTLY IDENTIFIED FINDINGS, WHILE RED INDICATES HALLUCINATED OR INCORRECT FINDINGS.

Study ID	Ground Truth	Generated (Ours)	Clinical Analysis
Case 1 *p10325631 s55979476*	pa and lateral views of the chest provided . there is no focal consolidation effusion or pneumothorax . the cardiomediastinal silhouette is normal . no free air below the right hemidiaphragm is seen .	pa and lateral views of the chest provided . there is no focal consolidation effusion or pneumothorax . the cardiomediastinal silhouette is normal . imaged osseous structures are intact .no free air below the right hemidiaphragm is seen .	**Complete success.** The generated report closely matches the radiologist interpretation and correctly identifies all major clinical findings with consistent terminology.
Case 2 *p10795257 s54036463*	Frontal and lateral radiographs demonstrate normal heart size, mediastinal and hilar contours. Clear lungs. No pneumothorax or pleural effusion. No radiopaque foreign body.	The lungs are clear without focal consolidation. No pleural effusion or pneumothorax. The cardiac and mediastinal silhouettes are stable.	**Partial success.** The major clinical findings are correctly identified, including clear lungs and absence of pleural effusion or pneumothorax. Although the foreign body observation is omitted, the generated report remains clinically coherent.
Case 3 *p10449408 s54732880*	as compared to the previous image there is unchanged evidence of a relatively widespread right lung multifocal pneumonia. the severity and extent of the changes has slightly increased as compared to the outside hospital film . on the left there is a zone of increased lung opacity around the hilus which could however be atelectatic in origin ...	as compared to the previous radiograph there is no relevant change . moderate cardiomegaly with moderate pulmonary edema and moderate bilateral pleural effusions with moderate pulmonary edema and small bilateral pleural effusions . moderate cardiomegaly persists . no pneumothorax . moderate cardiomegaly .	**Acknowledged limitation.** The model fails to identify multifocal pneumonia and generates unrelated findings such as cardiomegaly and pleural effusions. This example highlights the difficulty of handling complex multipathology studies and the susceptibility of generative models to hallucinated findings.

$L = 3$. Increasing the depth to $L = 6$ slightly improves R-L but reduces B-4 and CIDEr, suggesting mild overfitting. These results indicate that three encoder-decoder layers provide the best trade-off between representation capacity and generalization.

VI. QUALITATIVE ANALYSIS AND CLINICAL VALIDATION

To assess the practical clinical utility of the proposed framework, we conducted a qualitative evaluation comparing ground truth reports with outputs generated by our model. The generated reports were reviewed by a senior radiologist from the Radiology Department of Hedi Chaker University Hospital.

TABLE V
HYPERPARAMETER ABLATION

Parameters	B-4	R-L	CIDEr
Varying Memory Size C (with top-k=32)			
$C = 512$	0.121	0.301	0.301
$C = 1024$	0.121	0.301	0.278
$C = 2048$	**0.135**	**0.302**	**0.306**
$C = 4096$	0.113	0.276	0.268
Varying top-k retrieval (with C=2048)			
$top\text{-}k = 8$	0.121	0.301	0.293
$top\text{-}k = 16$	0.121	0.301	0.293
$top\text{-}k = 32$	**0.135**	**0.302**	**0.306**
Varying Beam Size			
Beam Size = 1	0.084	0.254	0.150
Beam Size = 3	**0.135**	**0.302**	**0.306**
Beam Size = 5	0.109	0.288	0.229
Varying Encoder/Decoder Layers			
Layers = 1	0.121	0.301	0.293
Layers = 3	**0.135**	**0.302**	**0.306**
Layers = 6	0.122	0.303	0.296

Table IV presents three representative cases illustrating different performance scenarios of the proposed framework. Case 1 demonstrates accurate report generation with highly consistent clinical findings. Case 2 shows clinically coherent generation despite minor wording differences and omission of a secondary observation. Case 3 highlights a failure case where the model misses multifocal pneumonia and hallucinates unrelated abnormalities, illustrating the limitations of generative report synthesis on complex multi-pathology studies.

VII. CONCLUSION

In this work, we introduced a dual-view Vision Transformer framework augmented with a Cross-Modal Memory Network for automated radiology report generation. By simultaneously processing frontal and lateral radiographs and aligning global visual patches with a shared medical memory, our architecture successfully bridges the semantic gap between complex medical imaging and clinical terminology. Evaluated on the MIMIC-CXR benchmark, our resource-efficient approach consistently outperforms established baselines in both natural language and clinical efficacy metrics, achieving high diagnostic precision without the massive computational overhead of large-scale language models. Future work will focus on expanding the training corpus to the entire MIMIC-CXR dataset, integrating specialized factual metrics like RadGraph F1, and exploring hybrid frameworks that leverage efficient LLM decoding to further enhance diagnostic reliability and narrative naturalness.

ACKNOWLEDGMENT

The authors would like to thank Pr. Wiem Feki and her team at the Radiology Center of CHU Hedi Chaker for their valuable contribution in validating the generated reports obtained in this work.

REFERENCES

[1] X. Wang *et al.*, "A survey of deep-learning-based radiology report generation using multimodal inputs," *Med. Image Anal.*, vol. 103, p. 103627, 2025.

[2] A. E. W. Johnson *et al.*, "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs," *arXiv preprint arXiv:1901.07042*, 2019.

[3] D. Demner-Fushman *et al.*, "Preparing a collection of radiology examinations for distribution and retrieval," *J. Am. Med. Inform. Assoc.*, vol. 23, no. 2, pp. 304–310, 2016.

[4] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, "Generating radiology reports via memory-driven transformer," in *Proc. EMNLP*, 2020, pp. 1439–1449.

[5] A. Vaswani *et al.*, "Attention is all you need," in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[6] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[7] J. Wang, A. Bhalerao, T. Yin, S. See, and Y. He, "Camanet: class activation map guided attention network for radiology report generation," *IEEE J. Biomed. Health Inform.*, vol. 28, no. 4, pp. 2199–2210, 2024.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[10] B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports," in *Proc. 56th Annu. Meet. Assoc. Comput. Linguistics*, 2018, pp. 2577–2586.

[11] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3156–3164.

[12] S. Bannur *et al.*, "Learning to exploit temporal structure for biomedical vision-language processing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 15016–15027.

[13] Z. Chen *et al.*, "A vision-language foundation model to enhance efficiency of chest X-ray interpretation," *arXiv preprint arXiv:2401.12208*, 2024.

[14] Z. Qiu *et al.*, "MedVersa: A Medical Foundation Model for Versatile Multimodal Processing," *arXiv preprint arXiv:2404.14810*, 2024.

[15] Z. Chen, Y. Shen, Y. Song, and X. Wan, "Cross-modal memory networks for radiology report generation," in *Proc. 59th Annu. Meet. Assoc. Comput. Linguistics*, 2021, pp. 5904–5914.

[16] X. Hou *et al.*, "MKCL: Medical Knowledge with Contrastive Learning model for radiology report generation," *J. Biomed. Inform.*, vol. 146, p. 104496, 2023.

[17] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meet. Assoc. Comput. Linguistics*, 2002, pp. 311–318.

[18] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[19] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4566–4575.

[20] M. Li *et al.*, "Dynamic graph enhanced contrastive learning for chest x-ray report generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 3334–3343.

[21] S. Yang *et al.*, "Radiology report generation with a learned knowledge base and multi-modal alignment," *Med. Image Anal.*, vol. 86, p. 102798, 2023.

[22] H. Jin, H. Che, Y. Lin, and H. Chen, "Promptmrg: Diagnosis-driven prompts for medical report generation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 3, pp. 2607–2615, 2024.

[23] S. Ju *et al.*, "DM-RRG: Radiology image report generation with disease-topic guidance and multi-grained image-text alignment," *Biomedical Signal Processing and Control*, vol. 114, p. 109322, 2026.