mathematics

MDPI

Article

RMPT: Reinforced Memory-Driven Pure Transformer for Automatic Chest X-Ray Report Generation

Caijie Qin [1], Yize Xiong [1], Weibin Chen [2] and Yong Li [1,*]

[1] Institute of Information Engineering, Sanming University, Sanming 365004, China
[2] Qingdao Nuocheng Chemicals Safty Technology Co., Ltd., Qingdao 266071, China
* Correspondence: 20070477@fjsmu.edu.cn

Abstract: Automatic generation of chest X-ray reports, designed to produce clinically precise descriptions from chest X-ray images, is gaining significant research attention because of its vast potential in clinical applications. Recently, despite considerable progress, current models typically adhere to a CNN–Transformer-based framework, which still fails to enhance the perceptual field during image feature extraction. To solve this problem, we propose the Reinforced Memory-driven Pure Transformer (RMPT), which is a novel Transformer–Transformer-based model. In implementation, our RMPT employs the Swin Transformer to extract visual features from given X-ray images, which has a larger perceptual field to better model the relationships between different regions. Furthermore, we adopt a memory-driven Transformer (MemTrans) to effectively model similar patterns in different reports, which is able to facilitate the model to generate long reports. Finally, we present an innovative training approach leveraging Reinforcement Learning (RL) that efficiently steers the model to focus on challenging samples, consequently improving its comprehensive performance across both straightforward and complex situations. Experimental results on the IU X-ray dataset show that our proposed RMPT achieves superior performance on various Natural Language Generation (NLG) evaluation metrics. Further ablation study results demonstrate that our RMPT model achieves 10.5% overall performance compared to the base mode.

Keywords: chest X-ray report generation; transformer; image-to-text; reinforcement learning

JEL Classification: 68T50

check for updates

Academic Editor: Ioannis Tsoulos

Received: 21 March 2025
Revised: 27 April 2025
Accepted: 29 April 2025
Published: 30 April 2025

Citation: Qin, C.; Xiong, Y.; Chen, W.; Li, Y. RMPT: Reinforced Memory-Driven Pure Transformer for Automatic Chest X-Ray Report Generation. *Mathematics* **2025**, *13*, 1492. https://doi.org/10.3390/math13091492

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Writing accurate chest X-ray reports requires radiologists to meticulously scrutinize image details, which demands a substantial investment of time and technical expertise. Recently, radiologists have been confronted with the formidable task of reviewing a substantial volume of medical images and composing diagnostic reports that possess precise content, standardized structure, and coherent semantics due to the burgeoning patient population, which presents significant challenges for them. In this situation, there is a growing demand for automated generation of chest X-ray reports, which has garnered increasing attention from the medical community [1–3].

Chest X-ray report generation refers to discovering and locating lesions through computer vision and natural language processing technology and describing them through a text report to lighten the workload of radiologists. Recently, drawing inspiration from image captioning techniques [4,5], the prevailing models for chest X-ray report generation [6,7]

have predominantly embraced the conventional encoder–decoder architecture. Within this framework, a CNN-based visual feature extractor, such as ResNet [8] or DenseNet [9], is employed to distill visual characteristics from the provided images. Subsequently, a report generation module, often based on the Transformer model [10], is utilized to translate these extracted visual features into a coherent paragraph that accurately depicts the input images. While these CNN–Transformer-based models have demonstrated impressive performance, they are not without their shortcomings, which include the following issues: (1) Lack of effective visual feature extraction methods. The traditional CNN-based methods fail to effectively model the relationships between different regions due to their limited perceptual field. (2) Difficulty in generating long sentences. The direct application of traditional image caption models to chest X-ray report generation is inadequate due to the lengthy nature of these reports. (3) Disregarding of the inter-word connections. Most models [11,12] employ cross entropy for optimization, which solely focuses on word-level errors.

To tackle the aforementioned issues, we propose the Reinforced Memory-driven Pure Transformer (RMPT) model. Specifically, to address the first problem, we construct a Transformer–Transformer framework, which eliminates the requirement for CNN. Compared with most existing models [13,14], our Transformer–Transformer framework considers the relationship among image regions. To solve the second problem, we introduce the memory-driven Transformer (MemTrans). In implementation, our MemTrans leverages a relational memory to capture vital information throughout the generation process. Furthermore, it incorporates Memory-driven Conditional Layer Normalization (MCLN), seamlessly integrating this information into the Transformer decoder. Compared with the vanilla Transformer, our MemTrans exhibits the capability of generating long reports. To solve the third problem, we propose a policy-based RL to exploit the inter-word connections. Compared with Self-Critical Sequence Learning (SCST) [15], our RL is adept at steering the model to focus more intently on challenging examples. This targeted emphasis not only elevates the model's overall performance but also ensures robustness across both general and complex scenarios. The effectiveness of our RMPT is evaluated through experiments conducted on the IU X-ray dataset, demonstrating superior performance across various NLG metrics. In summary, our contributions are:

1. We propose a Pure Transformer framework for automatic chest X-ray report generation, which eliminates the requirement for CNN.
2. We introduce the MemTrans to augment the model's capacity to generate long reports.
3. We introduce a novel RL-based training strategy to enhance the model's performance by exploiting the inter-word connections.
4. Evaluations on the IU X-ray dataset reveal that our proposed RMPT outperforms other models across various NLG metrics. Moreover, generalizability tests on the MIMIC-CXR dataset further validate its effectiveness.

The subsequent sections are structured as follows: Section 2 offers an extensive overview of pertinent literature. Section 3 initially outlines our RMPT, subsequently providing an in-depth examination of the Pure Transformer framework, MCLN, and the RL. For detailed information on the experiments, please refer to the Section 4, while the Section 5 emphasizes the key points.

2. Related Work

2.1. Chest X-Ray Report Generation

Most chest X-ray report generation models [16,17] employ the encoder–decoder framework. The encoder typically employs a CNN-based neural network to extract high-level visual features from chest X-ray images. The decoder typically employs a language model

to generate accurate reports. In this section, we divide the current chest X-ray report generation models into the following two categories based on the language model they employ.

(1) CNN-LSTM-based models utilize the Hierarchical LSTM as the report generator. For example, Jing et al. [18] propose the Co-Attention model, which performs an extra multi-label classification step to locate abnormal regions. On this basis, Park et al. [19] present the mDiTag model to take full advantage of the differences between normal and abnormal features. In addition to this, Zhang et al. [20] utilize a graph convolutional network to introduce prior knowledge. In this way, the model can easily discover various diseases and the relationship between diseases. The CNN-LSTM framework plays an important role in chest X-ray report generation tasks. However, limited by training efficiency and expressiveness, researchers have recently started to explore CNN–Transformer-based models.

(2) CNN–Transformer-based models utilize the Transformer as the report generator. For example, Liu et al. [21] employ contrast attention to obtain the differences between normal and lesion features and use them to guide the prediction of the model. The Progressive Transformer proposed by Nooralahzadeh et al. [22] divides the generation of the report into two stages. That is, image-to-text-to-text, which first generates a global concept based on radiological images and then transforms it into a more fine-grained coherent text. The Align Transformer proposed by You et al. [23] adopts the Align Hierarchical Attention and Multi-Grained Transformer to solve the data bias and very long sequence problems of the chest X-ray report generation task, respectively. The AIMNET model proposed by Shi et al. [24] uses an adaptive approach to focus on image features or disease tag features.

Both of the aforementioned models utilize conventional CNN-based architectures for visual feature extraction. In recent times, the Swin Transformer [25] and its diverse variants have achieved outstanding performance in tasks including object detection and image segmentation. Nevertheless, their application in the automatic generation of chest X-ray reports remains unexplored, thereby prompting us to develop the RMPT model.

2.2. Reinforcement Learning for Image-to-Text

J. Rennie et al. [15] propose a policy gradient-based SCST to train on a single NLG metric, thereby enhancing the precision of word prediction. The outstanding performance of SCST has led to its widespread adoption in various image-to-text conversion tasks [26–28]. In chest X-ray report generation, Qin et al. [29] adopt the SCST to facilitate the alignment between visual and textual features. Yi et al. propose the TSGET model to capture critical global features and utilize the SCST to further improve model performance. To relieve the error accumulation, Jin et al. [30] optimize the model by SCST after a regular training progress. Unlike SCST, which relies solely on a single NLG evaluation metric (BLEU-4 [31]) for reward calculation, our RL framework assigns varying degrees of significance to three pivotal NLG evaluation metrics [31–33]. Furthermore, instead of adopting a one-size-fits-all strategy that treats all samples equally regardless of their learning complexity, our RL adeptly guides the model to focus more intensively on challenging samples. This strategic approach markedly elevates the model's overall performance, ensuring robust handling of both simple and complex scenarios.

3. Proposed Method

The pipeline of our RMPT is shown in Figure 1.

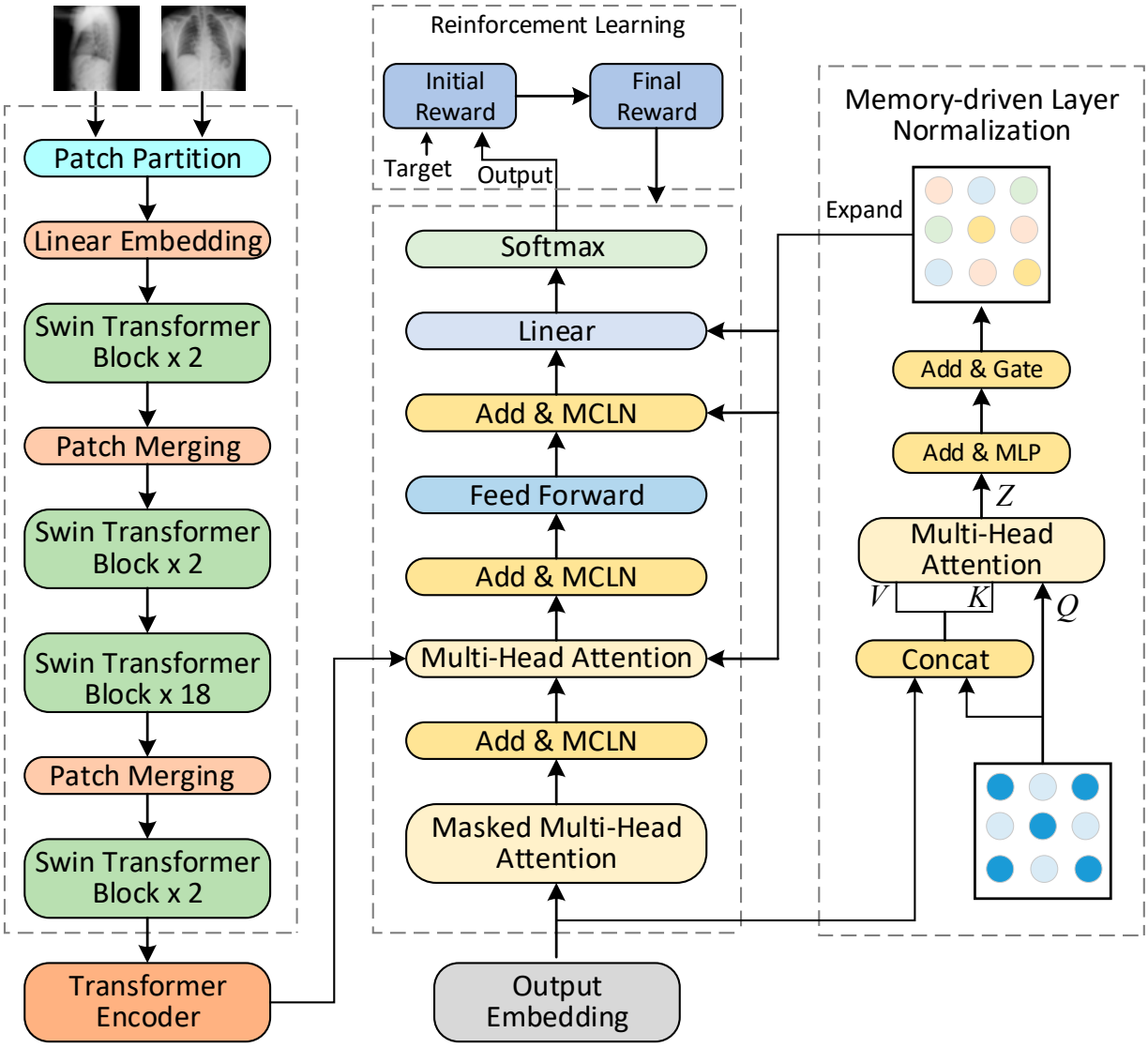


Figure 1. Overview of the RMPT model.

As shown in Figure 1, our RMPT model consists of four major components. (1) Visual extractor: we employ a pre-trained Swin Transformer [25] to extract visual features from input images. For a fair comparison, we utilize both the lateral and frontal view chest X-ray images as input. (2) Encoder: following most chest X-ray report generation models [1–3], we adopt a three-layer Transformer encoder for modeling the relationship between different targets. (3) Memory-driven Transformer: we integrate the MCLN into the Transformer decoder to generate long reports, which can model similar patterns in different radiology reports. (4) Reinforcement learning: we introduce an RL-based fine-tuning method to further improve the model performance.

3.1. Transformer–Transformer Framework

In this paper, we first construct a Transformer–Transformer-based framework. Instead of employing the traditional CNN-based models, such as ResNet [8] or DenseNet [9], our framework adopts the Swin Transformer [25] as the visual extractor, which has a larger perceptual field to better model the relationships between different regions. The Swin Transformer draws on the strengths of both CNN and Transformer, which combines a sliding window and a multi-headed attention mechanism. Figure 2 shows two consecutive Swin Transformer blocks.

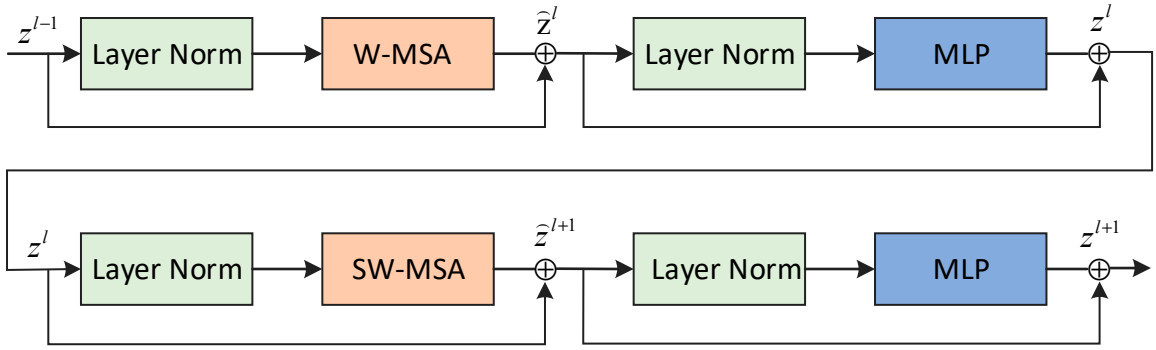


Figure 2. Swin Transformer block.

In the Swin Transformer, the traditional multi-head self-attention (MSA) module within each Transformer block is replaced with a novel shifted-window module, with all other layers preserved. As Figure 2 shows, each block in the Swin Transformer comprises layer normalization, a multi-head attention mechanism, a residual connection, and a two-layer MLP featuring GELU non-linearity. The block's structure can be represented as:

$$\widetilde{X}^{l} = WMSA(LN(X^{l-1})) + X^{l-1} \tag{1}$$

$$X^{l} = MLP(LN(\overset{\frown}{X}^{l})) + \overset{\frown}{X}^{l} \tag{2}$$

$$\overset{\frown}{X}^{l+1} = SWMSA(LN(X^{l})) + X^{l} \tag{3}$$

$$X^{l+1} = MLP(LN(\overset{\frown}{X}^{l+1})) + \overset{\frown}{X}^{l+1} \tag{4}$$

where $\overset{\frown}{X}^{l}$ and X^{l} are the outputs of the (S)WMSA module and the MLP module on the l^{th} block. Similar to the previous works, self-attention is calculated using the following equation:

$$Att(Q, K, V) = (\frac{QK^{T}}{\sqrt{d_k}} + B)V \tag{5}$$

where B is derived from the bias matrix.

The strategy of partitioning with shifted windows effectively enhances interactions between adjacent, non-overlapping windows from the preceding layer, showcasing its broad utility across a variety of computer vision tasks. This proven efficacy has inspired us to incorporate it as a visual extractor. To our knowledge, our work represents the pioneering effort in applying the Swin Transformer to the field of radiology report generation.

3.2. MemTrans

Building upon the Transformer architecture [14], our MemTrans incorporates two key modules. We first introduce relational memory to harness similar patterns across diverse reports. A memory matrix is employed to capture essential pattern information, with its states dynamically propagated throughout the generation phases:

$$M_t = f_g \odot M_{t-1} + i_g \odot M' \tag{6}$$

where f_g refers to the forget gate, i_g refers to the input gate, and M' refers to the temp state.

$$f_g = \sigma(y_{t-1}W_f + \tanh(M_{t-1})U_f) \tag{7}$$

$$i_g = \sigma(y_{t-1}W_i + \tanh(M_{t-1})U_i) \tag{8}$$

where σ refers to the sigmoid function, y_{t-1} refers to the previously generated word, W_f, U_f, W_i, U_i are learnable parameters, and M_{t-1} refers to the previous memory.

$$M' = f(Z + M_{t-1}) + Z + M_{t-1} \tag{9}$$

where $f()$ refers to the MLP.

$$Z = soft\max(\frac{QK^T}{\sqrt{d_k}})V \tag{10}$$

where $Q = W_q \cdot M_{t-1}$, $K = [M_{t-1}, y_{t-1}] \cdot W_k$, $V = [M_{t-1}, y_{t-1}] \cdot W_v$.

Based on the relational memory, we further propose the mcln to incorporate the relational memory to enhance the decoding of the Transformer. The calculation process can be formulated as follows:

$$f_{MCLN}(\psi) = \gamma'_t \odot \frac{\psi - \mu}{\nu} + \beta'_t \tag{11}$$

where ν and μ are the standard and mean deviation of ψ.

$$\gamma'_t = \gamma + f(m_t) \tag{12}$$

$$\beta'_t = \beta + f(m_t) \tag{13}$$

where $f()$ refers to the MLP, and γ and β are two major parameters, which are used to scale and translate features.

3.3. Reinforcement Learning

Our RL, grounded in policy gradients, is a reinforcement learning approach designed to enhance our model's overall performance. Within our RL framework, the memory-driven Transformer we propose acts as an agent engaging with the external environment. Consequently, the parameters θ of our proposed memory-driven Transformer delineate a policy that prompts an action. Upon producing the end token, the agent receives a reward, r. The training process aims to maximize the expected reward.

$$L(\theta) = -E_{y \sim P_\theta}[r(Y)] \tag{14}$$

where Y refers to the generated words. In implementation, we typically adopt a single sample drawn from P_θ to estimate the $L(\theta)$:

$$L(\theta) \approx -r(Y) \tag{15}$$

Next, we employ the reinforcement algorithm to compute the gradient:

$$\nabla_\theta L(\theta) = -E_{y \sim P_\theta}[r(Y) \nabla_\theta \log P_\theta(Y)] \tag{16}$$

We employ a single Monte Carlo sample to estimate the gradient:

$$\nabla_\theta L(\theta) = -r(Y) \nabla_\theta \log P_\theta(Y) \tag{17}$$

Furthermore, we introduce a reference reward b to reduce the high variance:

$$\nabla_\theta L(\theta) = -E_{y \sim P_\theta}[(r(Y) - b) \nabla_\theta \log P_\theta(Y)] \tag{18}$$

Therefore, the final gradient is estimated by utilizing a single sample:

$$\nabla_\theta L(\theta) = -r((Y) - b) \nabla_\theta \log P_\theta(Y) \tag{19}$$

Different from the SCST, we employ three representative NLG metrics as the initial reward:

$$r_i = x_1 B + x_2 M + x_3 R \tag{20}$$

where x_1, x_2, x_3 refer to the weight of the corresponding metric. B, M, and R refer to the BLEU-4 [31], METEOR [32], and ROUGE-L [33] evaluation metrics, respectively. Furthermore, we introduce a factor to adjust the initial reward, which can be formulated as follows:

$$\eta = (1 - \beta)^{\alpha} + 1 \tag{21}$$

$$\beta = f(r_i) \tag{22}$$

where $f()$ refers to the tanh function. The final reward is formulated as follows:

$$r = \eta r_i \tag{23}$$

4. Experiments

4.1. Datasets and Evaluation Metrics

Our RMPT model is experimentally validated on the IU X-ray dataset [34]. The statistics of the IU X-ray dataset are illustrated in Table 1. In alignment with the prevalent methods [1,13], we have partitioned the dataset into training, validation, and testing subsets in a ratio of 7:1:2. We employ the traditional NLG evaluation metrics, including BLEU [31], METEOR [32], and ROUGE-L [33], to verify the effectiveness of our RMPT.

Table 1. The statistics of the IU X-ray dataset.

	Train	**Val**	**Test**
Image	5226	748	1496
Report	2770	395	790
Patient	2770	395	790
Average length	37.56	36.78	33.62

4.2. Experimental Settings

Datasets pre-processing details. In the image pre-processing stage, we filter out images lacking corresponding reports and standardize their dimensions to 224 × 224 pixels. During the report pre-processing phase, we convert all words to lowercase through tokenization, subsequently eliminating special tokens and words that appear less than three times.

Model details. For feature extracting, we employ the Swin Transformer-B to extract visual features from given images. The proposed MemTrans is based on the Transformer with randomly initialized parameters, where the number of encoder–decoder layers is set to three, and the dimension of the model is set to 512. The memory slot in our MemTrans is set to three.

Training details. The RMPT model is initially trained for a duration of 100 epochs, using CE loss, with a mini-batch of 16. We employ the Adam optimizer [35], and the learning rate is set to 5×10^{-5}. Subsequently, the RMPT model undergoes fine-tuning with our proposed RL for an additional 100 epochs, this time utilizing a mini-batch size of 10. We set the max length of generated reports to 60. We use both the frontal and lateral images as input. The weight combination for BLEU-4/METEOR/ROUGE-L in our proposed RL is set to 5:1:5, and the penalty factor is set to 2.0.

4.3. Comparison with Previous Studies

We compare our RMPT with 15 recent methods to assess the efficacy of our proposal. The compared models include image caption models: Grounded [36], M2Transformer [37];

CNN-LSTM-based chest X-ray report generation models: Co_Att [18], mDiTag(-) [19], SentSAT + KG [20], CMAS-RL [38]; CNN–Transformer-based chest X-ray report generation models: R2Gen [13], CMN [12], METransformer [7], Align Transformer [23], PPKED [39], MMTN [17], RL-CMN [29], GSKET [40], CECL [41], CmEAA [42] and Transformer–Transformer based chest X-ray report generation Models: PureT [43]. Experimental results are presented in Table 2. There are four major observations.

Table 2. The performance comparison results of our proposed RMPT model with other models on the test sets of the IU X-ray dataset. (The experimental results are presented as percentages).

Models	B-1	B-2	B-3	B-4	M	R
Image Captioning Models						
Grounded [36]	46.3	31.8	21.4	15.5	-	33.5
M2Transformer [37]	44.6	30.1	23.7	17.6	-	34.3
CNN-LSTM-based Chest X-ray Report Generation Models						
Co_Att [18]	27.4	15.9	10.5	6.8	-	21.0
mDiTag(-) [19]	32.9	19.8	13.5	9.4	19.4	27.3
SentSAT + KG [20]	44.1	29.1	20.3	14.7	-	36.7
CMAS-RL [38]	46.4	30.1	21.0	15.4	-	36.2
CNN–Transformer-based Chest X-ray Report Generation Models						
R2Gen [13]	47.0	30.4	21.9	16.5	18.7	37.1
CMN [12]	47.5	30.9	22.2	17.0	19.1	37.5
METransformer [7]	48.3	32.2	22.8	17.2	19.2	38.0
Align Transformer [23]	48.4	31.3	22.5	17.3	20.4	37.9
PPKED [39]	48.3	31.5	22.4	16.8	-	37.6
MMTN [17]	48.6	32.1	23.2	17.5	-	36.1
RL-CMN [29]	49.4	32.1	23.5	18.1	20.1	38.4
GSKET [40]	49.6	32.7	23.8	17.8	-	38.1
CECL [41]	48.5	33.1	24.1	18.1	20.8	40.6
CmEAA [42]	48.1	31.9	23.4	18.1	-	39.2
Transformer–Transformer-based Chest X-ray Report Generation Models						
PureT [43]	49.6	31.9	24.1	17.5	-	37.7
RMPT (Ours)	**49.7**	**34.9**	**25.7**	**19.4**	20.2	39.5

Firstly, compared to conventional image captioning models, Grounded [36] and M2Transformer [37], which are designed to generate a short sentence to describe the major content of given images, the RMPT model achieves greater scores on all evaluation metrics. This observation confirms that designing a specific model for generating long sentences is necessary.

Secondly, compared to CNN-LSTM-based chest X-ray report generation models, Co_Att [18], mDiTag(-) [19], SentSAT + KG [20], and CMAS-RL [38], which utilize LSTM as the report generator, our RMPT achieves greater scores on all evaluation metrics. This observation confirms the influence of Transformer architecture.

Thirdly, compared to CNN–Transformer-based chest X-ray report generation models—R2Gen [13], CMN [12], METransformer [7], Align Transformer [23], PPKED [39], MMTN [17], RL-CMN [29], GSKET [40], CECL [41], and CmEAA [42]—our RMPT achieves greater scores on most evaluation metrics. This observation confirms the effectiveness of the Transformer–Transformer framework.

Fourthly, compared to the Transformer–Transformer-based chest X-ray report generation model, PureT [43], which utilizes the Vision Transformer [44] as the visual extractor,

our RMPT achieves greater scores on all evaluation metrics. This observation confirms the effectiveness of the Swin Transformer and our proposed RL.

4.4. Ablation Study

We conducted ablation studies on the IU X-ray dataset to validate the effectiveness of our MemTrans and RL. The experimental results are displayed in Table 3.

Table 3. Ablation study results on IU X-ray dataset. (The experimental results are presented as percentages).

Models	B-1	B-2	B-3	B-4	M	R	Improve
Base	48.0	30.9	22.0	16.3	19.6	36.7	-
Base + MemTrans	47.5	32.4	24.4	18.9	19.7	**39.8**	6.6%
Base + MemTrans + RL	**49.7**	**34.9**	**25.7**	**19.4**	**20.2**	39.5	**10.5%**

In this paper, we design a Transformer–Transformer-based framework, serving as our base model. Based on this framework, we first introduce the MemTrans to help to generate long reports. As Table 3 shows, compared to the base model, the incorporation of our MemTrans can dramatically enhance the quality of generated reports, which results in an average performance improvement of 6.6%. The experimental results demonstrate the effectiveness of MemTrans. Furthermore, we propose an RL training strategy designed to effectively steer the model towards allocating greater attention to challenging examples. As Table 2 shows, compared to the base model, the incorporation of our RL can dramatically enhance the quality of generated reports, which results in an average performance improvement of 10.5%. The experimental results demonstrate the effectiveness of RL.

4.5. Discussion

In this section, we undertake a series of experiments on the IU X-ray dataset to examine the impact of various settings on model performance. These include different configurations of the Swin Transformer, diverse visual extractors, varied reinforcement learning training strategies, and distinct memory slot arrangements.

4.5.1. Effect of Different Configurations of Swin Transformer

To better validate the impact of different configurations of Swin Transformer on model performance, we use four different configurations of Swin Transformer as visual feature extractors, Swin-T, Swin-S, Swin-B, and Swin-L. The experimental results obtained using the above four visual feature extraction models are shown in Table 4. As shown in Table 4, our model performs better and better as the number of model parameters increases, and when Swin-B is used as the visual feature extractor, the best effect is achieved. However, when the number of model parameters continues to increase, the model performance deteriorates dramatically. The reason behind this might be that the IU X-ray dataset is relatively small. Therefore, we select the Swin-B as our visual extractor.

Table 4. Performance of our model (only training with cross-entropy loss) using different configurations of Swin Transformer on the IU X-ray dataset. (The experimental results are presented as percentages).

Backbone	B-1	B-2	B-3	B-4	M	R
Swin-T	47.0	30.1	21.3	15.8	19.5	36.0
Swin-S	**49.5**	31.2	21.5	15.7	18.6	37.3
Swin-B	47.5	**32.3**	**24.3**	**18.8**	**19.7**	**39.5**
Swin-L	43.2	27.0	18.8	14.0	18.1	34.1

4.5.2. Effect of Different Visual Extractors

To further validate the impact of Swin Transformer, we compare it with traditional CNN-based extractors, including Vgg-16 [45], Vgg-19 [45], GoogLeNet [46], ResNet-18 [8], ResNet-50 [8], ResNet-101 [8], ResNet-152 [8], and DenseNet-121 [9]. The experimental results are shown in Table 5. As shown in Table 5, our model achieves better performance when Swin-B is used as the visual feature extractor, which demonstrates the effectiveness of our pure transformer-based framework.

Table 5. Performance of our model (only training with cross-entropy loss) using different CNN-based extractors on the IU X-ray dataset. (The experimental results are presented as percentages).

Backbone	B-1	B-2	B-3	B-4	M	R
Vgg-16 [45]	44.6	28.7	21.1	16.5	17.4	35.0
Vgg-19 [45]	43.3	27.1	19.4	14.9	17.9	34.1
GoogLeNet [46]	45.9	29.3	21.0	15.9	18.1	35.9
ResNet-18 [8]	43.7	28.8	21.0	16.4	18.3	36.7
ResNet-50 [8]	43.1	28.2	21.1	16.8	17.6	35.9
ResNet-101 [8]	**47.6**	30.9	22.4	17.2	19.6	37.1
ResNet-152 [8]	44.0	28.3	20.6	15.9	17.9	35.4
DenseNet-121 [9]	43.2	28.2	20.5	15.9	18.2	36.2
Swin-B [25]	47.5	**32.4**	**24.4**	**18.9**	**19.7**	**39.8**

4.5.3. RL vs. SCST

To rigorously evaluate the effectiveness of our RL, we have conducted a comparative analysis against the established conventional SCST method [15]. In this vein, we have separately fine-tuned our model using our novel RL strategy and the traditional SCST technique [15]. The comparative performance outcomes are detailed in Table 6. As Table 6 shows, compared to the SCST, our RL achieves superior performance on all NLG evaluation metrics, which demonstrates its effectiveness.

Table 6. Performance comparisons of our model fine-tuned with SCST and our RL on the IU-Xray dataset. (The experimental results are presented as percentages).

Backbone	B-1	B-2	B-3	B-4	M	R
Base	47.5	32.4	24.4	18.9	19.7	**39.8**
SCST [15]	45.8	32.8	25.0	19.3	18.2	38.2
RL (Ours)	**49.7**	**34.9**	**25.7**	**19.4**	**20.2**	39.5

4.5.4. Effect of Different Memory Slots

We also conduct experiments to validate the impact of the quantity of memory slots. To this end, we train our model with varying memory slots. The comparative results are detailed in Table 7. As evidenced in Table 7, our RMPT achieves superior performance across most evaluation metrics when the memory slot is set to 3, except BLEU-1 and METEOR scores. Accordingly, the memory slot will be uniformly maintained at 3 for the rest of this study.

Table 7. Performance comparisons of our RMPT (only training with CE loss) using different memory slots. (The experimental results are presented as percentages).

Backbone	B-1	B-2	B-3	B-4	M	R
1	**49.9**	31.3	21.6	15.6	18.7	37.3
2	47.8	30.5	21.6	15.9	**19.8**	36.4
3	47.5	**32.3**	**24.3**	**18.8**	19.7	**39.5**
4	46.7	30.1	22.1	17.3	19.5	36.8
5	43.9	27.7	19.8	15.2	17.8	33.9
6	44.3	28.5	20.9	16.5	18.0	35.7

4.6. Generalizability Analysis

To validate our method's robustness and ensure its practical applicability, we evaluate our method on MIMIC-CXR dataset. As Table 8 shows, compared to the base model, the incorporation of our MemTrans and RL can dramatically enhance the quality of generated reports, which results in an average performance improvement of 6.8% and 14.0%, respectively. The experimental results demonstrate the effectiveness of method.

Table 8. Ablation study results on MIMIC-CXR dataset. (The experimental results are presented as percentages).

Models	B-1	B-2	B-3	B-4	M	R	Improve
Base	29.1	18.3	12.4	8.9	12.0	26.8	-
Base + MemTrans	32.1	19.8	13.3	9.6	12.8	26.8	6.8%
Base + MemTrans + RL	**33.1**	**21.4**	**14.7**	**10.7**	**12.7**	**29.2**	**14.0%**

5. Conclusions

In this paper, we propose a novel Reinforced Memory-driven Pure Transformer model. In implementation, instead of employing CNN-based models, our RMPT employs the Swin Transformer, which has a larger perceptual field to better model the relationships between different regions, to extract visual features from given X-ray images. Moreover, we adopt MemTrans to effectively model similar patterns in different reports, which can facilitate the model to generate long reports. Furthermore, we develop an innovative policy-driven RL training strategy that efficiently steers the model to focus more on some challenging samples, thereby uplifting its comprehensive performance across both typical and complex situations. Experimental results on the IU X-ray dataset show that our proposed RMPT achieves superior performance on various NLG evaluation metrics. Further ablation study results demonstrate that our RMPT model achieves 10.5% overall performance compared to the base mode. In our future research, we will extend our experiments to include a diverse range of additional datasets to thoroughly assess the robustness and generalizability of our proposed method.

Author Contributions: Conceptualization, C.Q.; Software, C.Q. and Y.X.; Data curation, W.C.; Writing—original draft, C.Q.; Writing—review & editing, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been partially supported by the Natural Science Foundation of Fujian Province (2022J011178), the young and middle-aged teachers education research project of Fujian Province (JAT210422) and the Sanming College Scientific Research Development Fund (B202103).

Institutional Review Board Statement: This study did not include human or animal subjects. Therefore, ethical approval is not applicable.

Data Availability Statement: Data are available on request from the authors.

Acknowledgments: The authors would like to express their gratitude to the editor and the anonymous reviewers for their insightful suggestions, which significantly raised the caliber of this work.

Conflicts of Interest: Author Weibin Chen was employed by the Qingdao Nuocheng Chemicals Safty Technology Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Yi, X.; Fu, Y.; Liu, R.; Zhang, H.; Hua, R. TSGET: Two-Stage Global Enhanced Transformer for Automatic Radiology Report Generation. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 2152–2162. [CrossRef]
2. Huang, Z.; Zhang, X.; Zhang, S. Kiut: Knowledge-injected u-transformer for radiology report generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023, Vancouver, BC, Canada, 18–22 June 2023; pp. 19809–19818.
3. Gao, D.; Kong, M.; Zhao, Y.; Huang, J.; Huang, Z.; Kuang, K.; Wu, F.; Zhu, Q. Simulating doctors' thinking logic for chest X-ray report generation via Transformer-based Semantic Query learning. *Med. Image Anal.* **2024**, *91*, 102982. [CrossRef] [PubMed]
4. Tang, Z.; Yi, Y.; Yu, C.; Yin, A. PCATNet: Position-Class Awareness Transformer for Image Captioning. *Comput. Mater. Contin.* **2023**, *75*, 6007–6022. [CrossRef]
5. Farkh, R.; Oudinet, G.; Foued, Y. Image Captioning Using Multimodal Deep Learning Approach. *Comput. Mater. Contin.* **2024**, *81*, 3951–3968. [CrossRef]
6. Zhang, J.; Shen, X.; Wan, S.; Goudos, S.K.; Wu, J.; Cheng, M.; Zhang, W. A Novel Deep Learning Model for Medical Report Generation by Inter-Intra Information Calibration. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 5110–5121. [CrossRef] [PubMed]
7. Wang, Z.; Liu, L.; Wang, L.; Zhou, L. METransformer: Radiology Report Generation by Transformer with Multiple Learnable Expert Tokens. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023, Vancouver, BC, Canada, 18–22 June 2023; pp. 11558–11567.
8. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
9. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269. [CrossRef]
10. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
11. Tanida, T.; Müller, P.; Kaissis, G.; Rueckert, D. Interactive and Explainable Region-guided Radiology Report Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 7433–7442.
12. Chen, Z.; Shen, Y.; Song, Y.; Wan, X. Cross-modal memory networks for radiology report generation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, 1–6 August 2021; Volume 1, pp. 5904–5914. [CrossRef]
13. Chen, Z.; Song, Y.; Chang, T.; Wan, X. Generating radiology reports via memory-driven transformer. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020. [CrossRef]
14. Najdenkoska, I.; Zhen, X.; Worring, M.; Shao, L. Uncertainty-aware report generation for chest X-rays by variational topic inference. *Med. Image Anal.* **2022**, *82*, 102603. [CrossRef] [PubMed]
15. Rennie, S.J.; Marcheret, E.; Mroueh, Y.; Ross, J.; Goel, V. Self-Critical Sequence Training for Image Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1179–1195. [CrossRef]
16. Yi, X.; Fu, Y.; Yu, J.; Liu, R.; Zhang, H.; Hua, R. LHR-RFL: Linear Hybrid-Reward based Reinforced Focal Learning for Automatic Radiology Report Generation. *IEEE Trans. Med. Imaging* **2024**, *44*, 1494–1504. [CrossRef] [PubMed]
17. Cao, Y.; Cui, L.; Zhang, L.; Yu, F.; Li, Z.; Xu, Y. MMTN: Multi-modal memory trans former network for image-report consistent medical report generation. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 277–285. [CrossRef]
18. Jing, B.; Xie, P.; Xing, E. On the Automatic Generation of Medical Imaging Reports. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 2577–2586. [CrossRef]
19. Park, H.; Kim, K.; Yoon, J.; Park, S.; Choi, J. Feature Difference Makes Sense: A Medical Image Captioning Model Exploiting Feature Difference and Tag Information. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Online, 5–10 July 2020; pp. 95–102. [CrossRef]

20. Zhang, Y.; Wang, X.; Xu, Z.; Yu, Q.; Yuille, A.; Xu, D. When Radiology Report Generation Meets Knowledge Graph. In Proceedings of the 35th AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020.
21. Liu, F.; Yin, C.; Wu, X.; Ge, S.; Zhang, P.; Sun, X. Contrastive attention for automatic chest x-ray report generation. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online, 1–6 August 2021; pp. 269–280.
22. Nooralahzadeh, F.; Gonzalez, N.; Frauenfelder, T.; Fujimoto, K.; Krauthammer, M. Progressive transformer-based generation of radiology reports. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 2824–2832.
23. You, D.; Liu, F.; Ge, S.; Xie, X.; Zhang, J.; Wu, X. Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In Proceedings of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention, Strasbourg, France, 27 September–1 October 2021; pp. 72–82.
24. Shi, J.; Wang, S.; Wang, R.; Ma, S. Aimnet: Adaptive image-tag merging network for automatic medical report generation. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 7737–7741. [CrossRef]
25. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
26. Yan, C.; Hao, Y.; Li, L.; Yin, J.; Liu, A.; Mao, Z.; Chen, Z.; Gao, X. Task-Adaptive Attention for Image Captioning. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 43–51. [CrossRef]
27. Cao, S.; An, G.; Zheng, Z.; Wang, Z. Vision-Enhanced and Consensus-Aware Transformer for Image Captioning. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 7005–7018. [CrossRef]
28. Meng, L.; Wang, J.; Yang, Y.; Xiao, L. Prior Knowledge-guided Trans former for Remote Sensing Image Captioning. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4706213. [CrossRef]
29. Qin, H.; Song, Y. Reinforced cross-modal alignment for radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2022*; Association for Computational Linguistics: Dublin, Ireland, 2022; pp. 448–458. [CrossRef]
30. Jin, Y.; Chen, W.; Tian, Y.; Song, Y.; Yan, C. Improving radiology report generation with multi-grained abnormality prediction. *Neurocomputing* **2024**, *600*, 128122. [CrossRef]
31. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Stroudsburg, PA, USA, 7–12 July 2002; pp. 311–318.
32. Banerjee, S.; Lavie, A. 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, Ann Arbor, MI, USA, 29 June 2005; pp. 65–72.
33. Lin, C. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*; Association for Computational Linguistics: Barcelona, Spain, 2004; pp. 74–81.
34. Demner-Fushman, D.; Kohli, M.D.; Rosenman, M.B.; Shooshan, S.E.; Rodriguez, L.; Antani, S.; Thoma, G.R.; McDonald, C.J. Preparing a collection of radiology examinations for distribution and retrieval. *J. Am. Med. Inform. Assoc.* **2016**, *23*, 304–310. [CrossRef] [PubMed]
35. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the 3th International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
36. Zhou, Y.; Wang, M.; Liu, D.; Hu, Z.; Zhang, H. More grounded image captioning by distilling image-text matching model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 4777–4786. [CrossRef]
37. Cornia, M.; Stefanini, M.; Baraldi, L.; Cucchiara, R. Meshed-memory transformer for image captioning. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10578–10587. [CrossRef]
38. Jing, B.; Wang, Z.; Xing, E. Show, describe and conclude: On exploiting the structure information of chest x-ray reports. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July—2 August 2019; pp. 6570–6580.
39. Fenglin, L.; Xian, W.; Shen, G.; Fan, W.; Zou, Y. Exploring and distilling posterior and prior knowledge for radiology report generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 13753–13762.
40. Yang, S.; Wu, X.; Ge, S.; Zhou, S.; Xiao, L. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Med. Image Anal.* **2022**, *80*, 102510. [CrossRef] [PubMed]
41. Liu, X.; Xin, J.; Shen, Q.; Li, C.; Huang, Z.; Wang, Z. End-to-End Clustering Enhanced Contrastive Learning for Radiology Reports Generation. *IEEE Trans. Emerg. Top. Comput. Intell.* **2024**, *9*, 1780–1794. [CrossRef]

42. Huang, X.; Han, Y.; Li, R.; Wu, P.; Zhang, K. CmEAA: Cross-modal enhancement and alignment adapter for radiology report generation. In Proceedings of the 31st International Conference on Computational Linguistics, Abu Dhabi, United Arab Emirates, 19–24 January 2025; pp. 8546–8556.
43. Wang, Z.; Han, H.; Wang, L.; Li, X.; Zhou, L. Automated radiographic report generation purely on transformer: A Multi-Criteria supervised approach. *IEEE Trans. Med. Imaging* **2022**, *41*, 2803–2813. [CrossRef] [PubMed]
44. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 3–7 May 2021.
45. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA, 7–9 May 2015; pp. 1–10.
46. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A.; Liu, W.; et al. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.