



Full length article

C2M-DoT: Cross-modal consistent multi-view medical report generation with domain transfer network

Ruizhi Wang ^{a,c}, Zhenghua Xu ^a*, Xiangtao Wang ^a, Weipeng Liu ^b, Thomas Lukasiewicz ^c

^a State Key Laboratory of Intelligent Power Distribution Equipment and System, School of Health Sciences and Biomedical Engineering, Hebei University of Technology, Tianjin, China

^b School of Artificial Intelligence, Hebei University of Technology, Tianjin, China

^c Institute of Logic and Computation, Vienna University of Technology, Vienna, Austria

ARTICLE INFO

Keywords:

Multi-view contrastive learning
Domain transfer
Cross-modal consistency
Medical report generation

ABSTRACT

Objectives: In clinical practice, multiple medical images from different views provide valuable complementary information for diagnosis. However, existing medical report generation methods struggle to fully integrate multi-view data, and their reliance on multi-view input during inference limits practical applicability. Moreover, conventional word-level optimization often neglects the semantic alignment between images and reports, leading to inconsistencies and reduced diagnostic reliability. This paper aims to address these limitations and improve the performance and efficiency of medical report generation.

Methods: We propose C2M-DoT, a cross-modal consistent multi-view medical report generation method with domain transfer network. C2M-DoT (i) uses semantic-based contrastive learning to fuse multi-view information to enhance lesion representation, (ii) uses domain transfer network to bridge the gap in inference performance across views, (iii) uses cross-modal consistency loss to promote personalized alignment of multi-modalities and achieve end-to-end joint optimization.

Novelty and Findings: C2M-DoT pioneered the use of multi-view contrastive learning for the high semantic level of report decoding, and used a domain transfer network to overcome the data dependency of multi-view models, while enhancing the semantic matching of images and reports through cross-modal consistency optimization. Extensive experiments show that C2M-DoT outperforms state-of-the-art baselines and achieves a BLEU-4 of 0.159 and a ROUGE-L of 0.380 on the IU X-ray dataset, and a BLEU-4 of 0.193 and a ROUGE-L of 0.385 on the MIMIC-CXR dataset.

1. Introduction

With the advancement of medical imaging technologies, automatic medical report generation has emerged as a promising tool to enhance diagnostic efficiency and reduce the workload of radiologists. Deep learning-based methods, which combine visual encoders and language decoders, have significantly improved the quality of generated reports by capturing semantic features from images and translating them into coherent medical narratives.

However, the complexity of anatomical structures, variability in pathological manifestations, and the diversity of clinical language still pose major challenges to generating accurate and clinically useful reports. Recently, large language models (LLMs), such as GPT [1], have shown remarkable potential in this domain [2]. Yet, their high computational costs, limited interpretability, and dependency on large-scale annotated data constrain their practical deployment in clinical settings [3].

In contrast, conventional deep learning approaches offer advantages such as faster training, higher efficiency, and better task-specific interpretability. Enhancing these methods remains essential to achieving a more practical balance between performance, interpretability, and resource efficiency. In this work, we focus on improving traditional deep learning-based medical report generation models by addressing three key limitations:

First, most existing methods rely on single-view medical images to generate reports [4–6], overlooking the complementary information available in multi-view data [7]. While different views provide varied spatial and diagnostic perspectives, current models often treat them independently or naively concatenate them, without explicitly modeling inter-view relationships [8–10]. This limits the ability of models to extract comprehensive clinical insights, resulting in reports that may be incomplete or redundant.

* Corresponding author.

E-mail address: zhenghua.xu@hebut.edu.cn (Z. Xu).

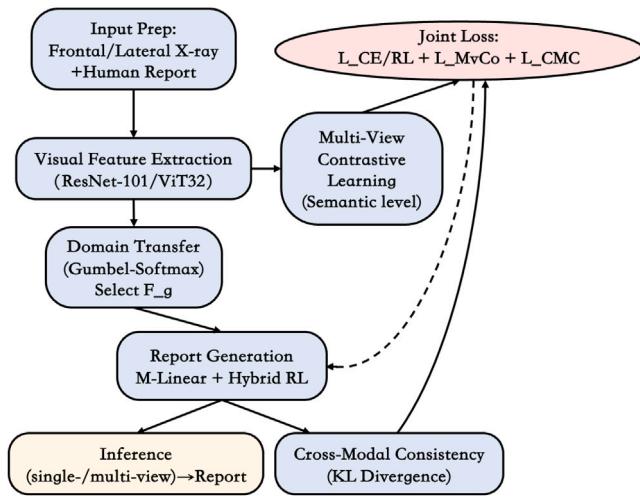


Fig. 1. Flow diagram summarizing the key steps of our proposed C2M-DoT.

Second, our experiments show that incorporating multi-view images improves performance, but also introduces a domain shift issue [11, 12]. Specifically, models trained on multi-view inputs often perform poorly when only single-view data is available during inference. This mismatch in input distributions leads to performance degradation and limits generalization across real-world clinical scenarios.

Third, current optimization strategies are typically unimodal and text-centric. Most models are trained using cross-entropy loss on text, which focuses solely on word-level accuracy while ignoring semantic fidelity and coherence at the report level. Although reinforcement learning has been applied to enhance sentence-level structure [9,13, 14], these methods still rely on text-only reward signals. Such approaches fail to ensure that generated descriptions are semantically aligned with the image content. To improve clinical reliability, optimization objectives should incorporate multimodal alignment that directly assesses image-text consistency.

In this paper, we propose a **Cross-modal Consistent Multi-view medical report generation with Domain Transfer network** (abbreviated by C2M-DoT), to address the three key limitations identified above. The key components are illustrated in Fig. 1.

First, we incorporate a **Multi-view Contrastive Learning (MvCo)** strategy into our prior reinforcement learning-based generation framework [9], to better leverage the complementary information from multi-view medical images. Although images from different views vary visually, they originate from the same patient and reflect the same underlying clinical findings [7]. Consequently, their corresponding textual descriptions should exhibit strong semantic consistency. MvCo exploits this property by enforcing alignment between the high-level semantic representations of the generated reports from each view. Unlike previous contrastive learning approaches [15–17], which operate at the encoder level and typically focus on improving image-text representations within a single view, MvCo performs contrastive learning directly on decoder outputs across views. This design promotes inter-view consistency in generated reports and improves the overall clinical completeness and coherence of the output.

Furthermore, to address the domain shift caused by inconsistent input distributions between training and inference, we propose a **Domain Transfer Network (DoT)** that incorporates a probability driven mechanism for selecting input views. During training, DoT dynamically chooses between single-view inputs and fused multi-view features based on a learned probability distribution. This approach allows the model to be exposed to both types of inputs rather than relying on a fixed or deterministic configuration. The selection process is guided by the semantic richness of the visual content and is designed to mimic

the clinical reasoning employed by radiologists, who assess different image perspectives and determine which views are most relevant for diagnosis. The mechanism is implemented through a probabilistic sampling strategy, which retains a certain degree of randomness. As a result, even less informative views have a non-zero probability of being selected during training. This design offers three main advantages: (i) it exposes the model to a more diverse and representative input distribution, thereby narrowing the performance gap between multi-view training and single-view inference; (ii) it allows the model to select the most informative view according to the content of each sample, which preserves the model's capacity for effective feature learning; and (iii) the occasional inclusion of single-view inputs in the generation branch helps to reduce the discrepancy in information between the report generation component and the contrastive learning component, improving consistency and generalizability.

Finally, to enhance the semantic alignment between image and report, we propose a **Cross-modal Consistency (CMC)** loss that extends the conventional unimodal text-based loss to a multimodal optimization objective. Specifically, CMC compares the semantic similarity between medical images and both predicted and ground-truth reports, and minimizes the divergence between the two similarity distributions. This encourages the generated report to remain semantically faithful to the image content, thereby improving both accuracy and interpretability.

Overall, the contributions of this paper are as follows:

- We identify three key limitations in existing medical report generation methods: (i) the inability to effectively utilize mutual information among multi-view images, (ii) performance degradation caused by domain shift between training and inference inputs, and (iii) reliance on unimodal, text-only optimization objectives. To address these issues, we propose a new framework named C2M-DoT (Cross-modal Consistent Multi-view medical report generation with Domain Transfer).
- The improvements in the proposed C2M-DoT are threefold: (i) we introduce a Multi-view Contrastive Learning (MvCo) strategy, which leverages the complementary information in multi-view chest X-ray images to improve report generation; (ii) we incorporate a Domain Transfer Network (DoT) to ensure the model maintains high performance even when only single-view inputs are available during inference; and (iii) we propose a Cross-modal Consistency (CMC) optimization objective, which uses image-text semantic alignment to guide the model in generating more semantically faithful reports.
- Extensive experiments are conducted on two publicly available medical report generation benchmarks. Results show that our proposed C2M-DoT significantly outperforms state-of-the-art baselines in six commonly used standard natural language generation metrics, and the generated reports exhibit better semantic alignment with the input images. Ablation studies further validate the effectiveness of MvCo, DoT, and CMC. Additionally, we demonstrate that C2M-DoT can achieve nearly the same performance using only single-view inputs, which enables its applicability to incomplete or unpaired clinical images and helps reduce unnecessary X-ray exposure, benefiting both diagnostic efficiency and patient safety.

2. Related works

Medical report generation usually uses convolutional neural networks as visual encoders and recurrent neural networks for sentence generation. Due to the emergence of transformers, many works have also used this to improve the quality of long text generation [5,6]. The ViT image encoder [24] and Transformer text decoder combined with the multi-task strategy resulted in an overall improvement in reported performance [21,23]. To understand and describe complex lesions

Table 1
Summary of advantages and limitations of some popular report generation methods.

Method	Advantages	Limitation
Top-down [18]	Classic encoder-decoder architecture uses a top-down attention mechanism to focus on the regional features of the image.	General image description methods are difficult to generate multi-sentence long reports.
MRMA [10]	Combined with the multimodal recurrent attention mechanism, reports are generated sentence by sentence to enhance the ability to generate long texts.	Based on RNN sentence-by-sentence generation, model training is complex and prone to accumulated errors.
RTMIC [14]	Reinforcement learning strategy optimization directly improves the matching degree and clinical relevance of reports with reference texts.	Reinforcement learning training is complex and strategy convergence is difficult.
X-LAN [19]	X-linear bilinear attention mechanism can capture high-order interactions between image features and text features.	Without optimization for specific medical fields, the generated text has limited clinical accuracy.
R2Gen [5]	The Transformer decoder is enhanced by memory units, which enhances the ability to generate long texts.	The memory mechanism increases model complexity and storage overhead, and slows down the inference speed.
HReMRG-MR [9]	The report generation framework using comprehensive weighted hybrid reinforcement learning is used to comprehensively improve the report quality.	A large amount of labeled data is required to train the hybrid reward weights, and the cost of model migration is high.
MMTrans [20]	Incorporating general common sense knowledge into a multimodal generative model to improve report completeness.	Large pre-trained models (ViT and GPT-2) are used, which are large in size and expensive to train.
RRGTrans [21]	Using a multi-task strategy on a pure Transformer architecture makes the model better at distinguishing fine-grained abnormal differences in images.	The training strategy is complex and sensitive to hyperparameters.
CDGPT2 [22]	Use image labels to guide report generation and ensure that important abnormalities can be noticed by the model.	If the image label is missed or misdetected, the report may omit important information or introduce errors.
TranSQ [23]	Medical report generation is considered as a semantic query set prediction task to improve the interpretability of the generation process.	The sentence selection strategy may be inaccurate in complex cases, resulting in missing details or generating general statements.

more accurately, a series of spatial and linguistic channel attention mechanisms have been carefully designed [6,9,10,18,19,25], especially cross-modal attention, which significantly helps to describe important visual abnormalities [20]. At the same time, more additional data [22] and knowledge [26] are often explored to help generate more accurate and comprehensive reports [4]. **Table 1** summarizes the advantages and limitations of existing medical report generation methods.

However, they all ignore the clinically generated multi-view medical images. In fact, the large associations naturally present in patient metadata benefit the learning of visual representations [7]. Encouraged by the benefits of multi-view images, [8–10] tried to use multi-view medical images to generate reports, but only concatenated them and directly fed them into the model. This rudimentary method does not deeply explore and utilize the mutual information between different views, which cannot bring great benefits but may cause information redundancy. [27] further attempts to fuse different views to obtain more information, but due to the generality of the operation, it is not good enough in understanding the semantics of lesions and learning personalized sample features.

Contrastive learning has an excellent performance in learning personalized features of samples. Semantic feature expressions are obtained by comparing anchor samples with positive and negative samples [28–31]. Usually, augmented forms of the original data are selected as positive samples because they have strong semantic consistency with the original data, such as rotated or cropped images [28], text sentences replaced by synonyms [28]. Proper selection of positive and negative examples is crucial to the superior performance of contrastive learning. [30] found that using images from different perspectives of the same natural scene as positive examples can obtain more meaningful visual representations than using augmented images. Different views have greater representational differences but strong consistency at the semantic level.

Therefore, we propose a medical report generation method based on multi-view contrastive learning. Compared with existing methods, the

C2M-DoT method proposed in this paper has three advantages: (i) We innovatively use contrastive learning on multi-view medical images, and use contrastive learning for decoded semantic vectors to achieve comprehensive improvement of report quality. (ii) Proposes a domain transfer network that enables multi-view medical report generation models to achieve accurate inference using a single view. (iii) Also adopts cross-modal consistency optimization to enhance the semantic association between reports and images.

High matching between inference reports and original images is an important requirement and ultimate goal of medical report generation. This cross-modal matching mode is commonly found in visual language pre-training models [32]. There is currently a lot of work dedicated to using paired image-text pre-training to improve visual understanding of medical images. [16] uses multimodal contrastive learning to change the distance between visual and textual representations in latent space, and then [15] picks more difficult negative samples on this basis to optimize intra-class difference feature learning. [33–35] aligns visual regions and disease labels to learn multi-grained feature representations. [36,37] use medical knowledge-based semantic matching to learn relations between entities. Most of these methods follow Contrastive Language-Image Pre-training (CLIP) [32], which maximizes the similarity of paired text images while minimizing the similarity of unpaired elements to learn cross-modal matching relations. Furthermore, [38] fine-tunes it on medical data to better adapt CLIP to downstream tasks in the medical field.

Thanks to this efficient multimodal mechanism, we optimize the semantic similarity of the inference report to the original image. It is worth noting that: (i) Different from the visual model pre-training work, which separates the upstream visual model training from the downstream medical image analysis tasks, we set the end-to-end learning goal to make the semantic similarity of the image and text of the inference report and The image-text semantic similarity of the real report is consistent, and it is directly optimized for the generation of medical reports. (ii) In cross-modal consistency optimization, we

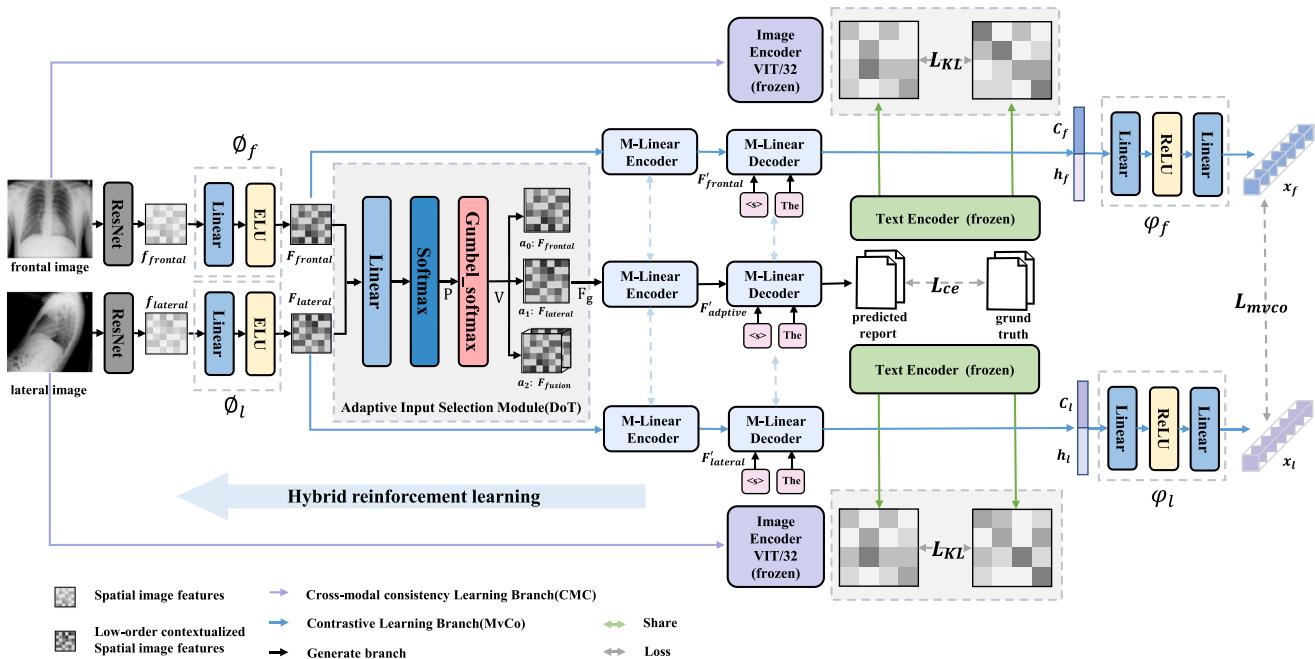


Fig. 2. The architecture of our proposed C2M-DoT network.

introduce a new supervisory signal: image-text semantic similarity. Compared with the single-modal matching of text based on n-grams, the consistency of semantic similarity between images and texts can capture the subtle semantic differences between texts and promote the semantic consistency between the generated report and the input image. This allows the generated reports to more accurately describe the content and features in the image. (iii) We further explore the combination of this cross-modal consistency with multi-view medical report generation.

3. Methodology

We propose a Cross-modal Consistent Multi-view medical report generation with Domain Transfer network (C2M-DoT). Intuitively, we believe that multi-view contrastive learning module will enhance the model's ability to explore lesion information, resulting in more accurate reports. Besides, we believe that domain transfer network will further bridge the performance gap between different views for report generation, and further improve the comprehensive performance of report generation. Moreover, we believe that cross-modal consistency optimizations will help generate semantically consistent descriptions and also help improve the correctness of reported results.

Specifically, as shown in Fig. 2, we adopt the architecture of a two-stream network for multi-view comparative learning and report generation, respectively. The multi-view contrastive learning branch uses frontal and lateral views of medical images independently. In order to focus on the important medical findings of chest X-rays and abstract accurate semantic representations, the original image is first sent to the pre-trained ResNet-101 and Transformer-like encoding network M-Linear encoder, and then the obtained high-order multi-dimensional visual embedding F' is sent to the multi-modal reasoning network M-Linear decoder. The context information c and hidden state h of the last time step of the decoding process participate in semantic contrastive learning. In the generation branch, medical images from different views are fed into the domain transfer network together, where the adaptive sampling module decides the visual embedding F_g that is finally used to generate the report according to the comprehensive information of the current research case. In addition, the decoded predicted report R_{pre} is further optimized for cross-modal consistency with the real report

Table 2
Frequently used symbols.

Symbol	Explanation
f_*	spatial visual features, $* \in \{\text{frontal}, \text{lateral}\}$
F_\diamond	global visual features, $\diamond \in \{\text{frontal}, \text{lateral}, \text{fusion}\}$
F_g	final input feature in the generation branch
F'_\diamond	high-order multi-dimensional visual embeddings, $\diamond \in \{\text{frontal}, \text{lateral}, \text{adaptive}\}$
A	input sampling action space
a_i	input sampling action
P	action sampling probability distribution
V	action value sampled from <i>Gumbel-softmax</i>
c_f, c_l	context vectors decoded from multiple views
h_f, h_l	hidden states decoded from multiple views
x_f, x_l	semantic embeddings for multiple views
u_f, u_l	visual semantic features for multiple views
R_{pred}, R_{true}	predicted report and real report
t_{pred}, t_{true}	text semantic features for predicted and real report
SFP_i^{2v}, SLP_i^{2v}	cross-modal similarity scores from multi-view images to predicted report text
SFP_i^{2v}, SLP_i^{2v}	cross-modal similarity scores from predicted report text to multi-view images
SFT_i^{2v}, SLT_i^{2v}	cross-modal similarity scores from multi-view images to real report text
SFT_i^{2v}, SLT_i^{2v}	cross-modal similarity scores from real report text to multi-view images

R_{real} . We present our network design and implementation details in the following subsections. The frequently used symbols are included and explained in Table 2.

3.1. Multi-view contrastive learning

Although existing medical report generation work has attempted the use of multi-view medical images, no research has explored the positive impact of their extensive associations on medical report generation. Multi-view contrastive learning was first used in natural scenes, and the resulting visual representations achieved excellent performance in downstream tasks such as segmentation and detection. We use multi-view contrastive learning in the task of medical report generation to explore the semantic consistency between different views and help generate reports.

Existing contrastive learning methods are often used in vision model pre-training to optimize representations. However, due to the target differences of upstream and downstream tasks, upstream visual features often cannot generalize well on downstream tasks. To make up for this deficiency, we introduce contrastive learning into an end-to-end medical report generation model to directly compare the decoded semantic embeddings and explore the impact of mutual information between different views of medical images on report generation.

Specifically, we propose a semantic-based multi-view contrastive learning (MvCo) method based on the backbone network of hybrid reinforcement learning medical report generation with M-Linear attention mechanism [9]. First, the pre-trained ResNet-101 [39] is used to initially extract the spatial visual features $f_{frontal}$ and $f_{lateral}$ of the frontal and lateral images of one case. In order to enable multi-view contrastive learning to better utilize the differences of different views for advanced semantically consistent representation learning, we further project the spatial visual features of each view to obtain more distinguishing visual information embeddings.

$$F_{frontal} = \phi_f(f_{frontal}), F_{lateral} = \phi_l(f_{lateral}) \quad (1)$$

where $\phi_f(\cdot)$ and $\phi_l(\cdot)$ are modeled as fully connected layers with ELU activations. $F_{frontal}$ and $F_{lateral}$ are the frontal and lateral view visual embeddings focusing on the difference of view information, which are fed into two weight-shared report generation networks with m-linear encoder-decoder.

To directly affect the quality of generated reports, multi-view contrastive learning is applied to the decoded semantic embeddings. We concatenate contextual semantic representations c_f, c_l and hidden layer information h_f, h_l decoded from different views, and then project onto the same implicit space for comparison.

$$x_f = \psi(Concat(c_f, h_f)), x_l = \psi(Concat(c_l, h_l)) \quad (2)$$

where $\psi(\cdot)$ is modeled as two fully connected layers with ReLU activations, according to [40]. We define the similarity between different elements in terms of cosine distance:

$$sim(m, n) = \frac{m \cdot n^\top}{\|m\| \|n\|}, sim(n, m) = \frac{n \cdot m^\top}{\|n\| \|m\|} \quad (3)$$

Since the lesion semantics presented by medical images from different views should be highly consistent, we maximize the similarity between the semantic embeddings of the frontal and lateral views of the same patient, while minimizing the similarity between the semantic embeddings of different patients. The multi-view contrastive loss L_{MvCo} is defined as:

$$\mathcal{L}_{MvCo} = -\log \frac{\exp(sim(x_l, x_f)/\tau_c)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq l]} \exp(sim(x_l, x_k)/\tau_c)} \quad (4)$$

where τ_c is temperature parameter. In addition, the effects of multi-view contrastive learning using different feature embeddings are detailed in Section 4.7.1.

3.2. Domain transfer network

The excellent properties of multi-view data can improve the semantic expression of the characteristics of abnormal lesions, and help to generate accurate high-quality medical reports. However, an important shortcoming of existing multi-view medical report generation schemes is that since multi-view data complement each other, multi-view data is required not only in the training phase but also in the inference phase, which limits its application in clinical practice. When a multi-view medical report generation model uses single-view data for inference, due to the gap in the input distribution between the training and inference stages, it will inevitably cause performance degradation in report inference, known as the domain shift problem. In order to overcome the above problems, we propose a domain transfer network. During the

training process, the model will receive a comprehensive input distribution of various single-view or multi-view images, and adaptively select the current most beneficial input to feed the generative model.

Specifically, we first define the input distribution as an action space $A \in \mathbb{R}^{1 \times 3}$:

$$a_i = \begin{cases} F_{frontal}, & i = 0 \\ F_{lateral}, & i = 1 \\ F_{fusion}, & i = 2 \end{cases} \quad (5)$$

where, $F_{frontal}$ and $F_{lateral}$ represent the front and lateral single-view visual feature input respectively, while F_{fusion} represents multi-view view input, which is obtained by adding the features of multi-views instead of concatenating them. The same input width can keep the process consistent between the multi-view generation branch and the contrast learning branch using the front single view respectively to enhance the overall performance of the network.

Then, in order for the model to obtain the most useful information input and better balance the use of frontal and lateral view information, we adaptively decide to input a single feature or a mixture of features through action sampling. This form of non-determinism enables the model to adaptively select the best input to obtain the maximum amount of visual information for each image.

To circumvent the technical problem that binary sampling actions can not be differentiated to participate in backpropagation, we utilize random sampling based on *Gumbel-Softmax* distribution. This reparameterization trick has been used in reinforcement learning for making discrete decision [41]. Non-differentiable action values will be replaced by differentiable samples from the *Gumbel-Softmax* distribution. Specifically, we concatenate the global visual features $F_{frontal}$ and $F_{lateral}$, which are multi-scale fusions of frontal and lateral views in the latent space and taken as a comprehensive information basis for the current action selection. It is sent to a linear layer through the fully connected layer to obtain the action confidence warehouse $P \in \mathbb{R}^{1 \times 3}$.

$$P = softmax(W_c(Concat(F_{frontal}, F_{lateral}))) \quad (6)$$

where W_c represents the fully connected layer parameter matrix. Subsequently, the sampling module will generate action values $V \in \mathbb{R}^{1 \times 3}$, which defined as

$$V(a) = \frac{\exp((\log(P_i(a)) + g_i(a))/\tau_s)}{\sum_{j=1}^3 \exp((\log(P_j(a)) + g_j(a))/\tau_s)}, \text{ for } i = 1, 2, 3 \quad (7)$$

where g represents the noise sampled from the standard *Gumbel-Softmax* distribution, and τ_s is the temperature parameter. The final input strategy is gained after V through *argmax* layer. During the inference stage, V is generated according to the input directly. Sample the action whose sample value in A is calculated to be 1, and reconstruct only the features corresponding to the action into the final input feature F_g .

$$F_g = A(a_i), V(a_i) = 1 \quad (8)$$

This view selection mechanism, guided by the amount of semantic information in the image, effectively simulates the clinical decision-making process of radiologists who assess image quality from different perspectives and decide whether to incorporate reference views. Moreover, since the *Gumbel-Softmax* retains a certain degree of strategic randomness, the model maintains a non-zero probability of selecting only the lateral view during training—even when the frontal or fused views carry more information. This design ensures the model remains practical and effective in clinical scenarios with incomplete inputs.

3.3. Cross-modal consistency

Traditional medical report generation methods have always used cross-entropy as an optimization method. As shown in Fig. 3(a), assuming the real report text sequence $R = \{r_1, r_2, \dots, r_Q\}$, and the

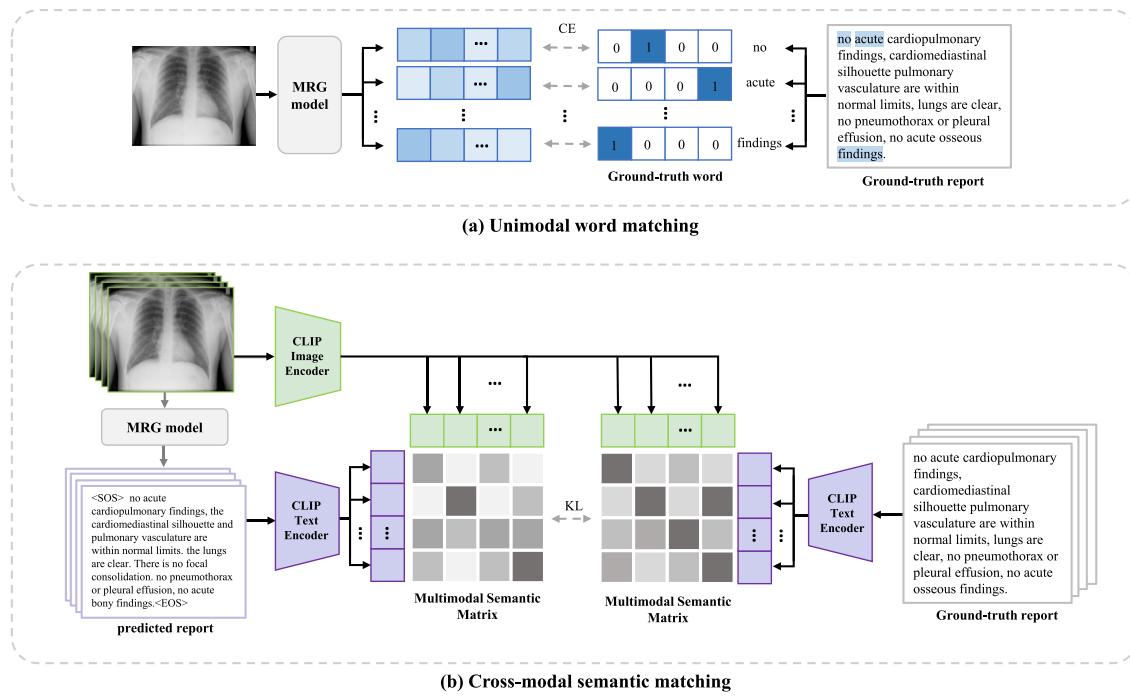


Fig. 3. Schematic diagram of (a) traditional single-modal optimization based on word matching and (b) our cross-modal optimization based on consistency of image-to-text semantic similarity matrix.

probability distribution sequence of the text sequence generated by the model $B = \{b_1, b_2, \dots, b_Q\}$. The cross-entropy loss is calculated as:

$$\mathcal{L}_{CE}(R, B) = - \sum_{q=1}^Q \sum_{d=1}^D r_{q,d} \cdot \log(b_{q,d}) \quad (9)$$

where D is the size of vocabulary, and Q is the sequence length. $b_{q,d}$ represents the probability that the word at the q th position in the sequence is the d th word in the vocabulary, while, $r_{q,d}$ is a one-hot vector, indicating whether the word at the q th position in the real sequence is the d th word in the vocabulary. Only one element of this vector is 1, which represents the real word position, and the other are 0.

The cross-entropy loss improves the quality of generated text by minimizing the difference between the probability distribution generated by the model and the real text distribution. However, since the calculation of each time step of the text sequence is independent, it can only focus on a single word, but cannot effectively capture the contextual relationship between words; moreover, all words in the text are mapped to numerical values representing probabilities, completely ignoring the semantic information of the text.

Although there are many reinforcement learning methods that use natural language indicators as rewards to further optimize the model and better capture the relationship between text contexts to improve coherence, only relying on limited language pattern matching still cannot truly understand the semantics of text. In addition, previous optimization methods are usually limited to the processing of single-modal text data, which means that these methods mainly focus on text information and ignore other important medical data modalities (such as medical images), and cannot effectively adapt to the multi-modal reasoning task of medical report generation. Therefore, we introduce cross-modal consistency loss to enhance the semantic consistency relationship between reports and images to ensure that the generated reports correctly describe and interpret image-related information.

Specifically, we define a visual encoder V_{CLIP} and a language encoder T_{CLIP} to extract features of medical images and report texts respectively. Referring to the setting of PubMedCLIP [38], we use the VIT-B/32 Vision Transformer [24] fine-tuned by medical dataset

Radiology Objects in COntext (ROCO) [42] to encode the frontal and lateral original images into visual features v_f and v_l ; and encode the generated reports and real reports into textual features t_{pred} and t_{true} :

$$v_f = V_{CLIP}(I_f), v_l = V_{CLIP}(I_l) \quad (10)$$

$$t_{pred} = T_{CLIP}(R_{pred}), t_{true} = T_{CLIP}(R_{true}) \quad (11)$$

where, I_f is the frontal image, I_l is the lateral image, R_{pred} is the prediction report, and R_{true} is the real report. Then, we define the similarity between two modalities via the cosine phase similarity distance shown in Eq. (3) and apply softmax normalization to it.

$$SFP_i^{v2t} = \frac{\exp(\text{sim}(v_f, t_{pred})) / \tau_m}{\sum_{j=1}^N \exp(\text{sim}(v_f, t_{pred_j})) / \tau_m} \quad (12)$$

$$SFP_i^{t2v} = \frac{\exp(\text{sim}(t_{pred}, v_{f_i})) / \tau_m}{\sum_{j=1}^N \exp(\text{sim}(t_{pred}, v_{f_j})) / \tau_m} \quad (13)$$

$$SFT_i^{v2t} = \frac{\exp(\text{sim}(v_f, t_{true})) / \tau_m}{\sum_{j=1}^N \exp(\text{sim}(v_f, t_{true_j})) / \tau_m} \quad (14)$$

$$SFT_i^{t2v} = \frac{\exp(\text{sim}(t_{true}, v_{f_i})) / \tau_m}{\sum_{j=1}^N \exp(\text{sim}(t_{true}, v_{f_j})) / \tau_m} \quad (15)$$

where τ_m is a learnable temperature parameter, and N is the number of training pairs. SFP_i^{v2t} , SFP_i^{t2v} , SFT_i^{v2t} and SFT_i^{t2v} are softmax normalized similarity scores from frontal image to predicted text, predicted text to frontal image, frontal image to real text, and real text to frontal image. Similarly, the normalized similarity scores from lateral image to predicted text, predicted text to lateral image, lateral image to real text and real text to lateral image are calculated as SLP_i^{v2t} , SLP_i^{t2v} , SLT_i^{v2t} and SLT_i^{t2v} , respectively.

Since the major human organs and obvious abnormalities in the X-rays are described in the report, there are inevitably more or less semantic similarities. Therefore, it is unreasonable to simply maximize the diagonal similarity of the similarity matrix to 1 and minimize the off-diagonal similarity to 0. Finally, we use Kullback-Leibler (KL) to optimize the similarity matrix. Our goal is to make the similarity matrix

between the predicted text and the image close to the similarity matrix between the real text and the image:

$$\mathcal{L}_{CMC}^F = \frac{1}{2} \epsilon_{(v,t) \sim \Theta} \left(KL(SFP_i^{v2t}, SFT_i^{v2t}) + KL(SFP_i^{t2v}, SFT_i^{t2v}) \right) \quad (16)$$

where \mathcal{L}_{CMC}^F is the frontal image text similarity loss, and the lateral image text similarity loss \mathcal{L}_{CMC}^L is calculated in a similar way. Finally, similarity losses for image and text modalities are combined with multi-views for cross-modal consistency optimization.

$$\mathcal{L}_{CMC} = \mathcal{L}_{CMC}^F + \mathcal{L}_{CMC}^L \quad (17)$$

4. Experiments

4.1. Datasets

To evaluate the performance of our proposed C2M-DoT, extensive experiments are conducted on two publicly available datasets. As shown in Table 3, both datasets contain chest X-ray images and paired free-text reports.

(i) IU X-ray [43] is one of the most commonly used medical image description datasets, collected by Indiana University. (ii) MIMIC-CXR [44] is currently the largest publicly available medical image description dataset, proposed by the Massachusetts Institute of Technology. Each imaging study may contain one or more images, including posteroanterior (PA) or anteroposterior (AP) views and lateral (LL) views. The medical report corresponding to the imaging results consists of multiple sentences, in which the two parts *impression* and *findings* summarize the main diagnostic results.

We preprocess the above two datasets as follows: First, for multi-view contrastive learning, we filter the data to only retain cases with frontal and lateral medical images and complete reports. The number of cases in each dataset is: (i) IU X-ray: 6222 images, 3111 corresponding reports, (ii) MIMIC-CXR: 153,448 images, 76,724 corresponding reports. Then, we resize the image to 224x224. For reports, *impression* and *findings* will be generated simultaneously. We convert all words to lowercase and remove special characters. Thereafter, we tokenize the reports to build word lists. In order to filter out many uncommon words, simplify the model structure and prevent overfitting, we only keep the words that appear more than 5 times, and replace the discarded words with the UNK token. The number of word tokens per dataset is: (i) IU X-ray: 776 tokens, (ii) MIMIC-CXR: 2991 tokens. During preprocessing, we further check and ensure that no patient-identifying information is retained. Since these data are all from public and compliant data sources, there is no privacy risk to specific patients. All experimental links of this study are based on the above anonymized data.

Finally, for all datasets, randomly select 70% of the datasets for training, 10% for validation, and 20% for testing, and make sure there is no overlap between datasets.

4.2. Evaluation metrics

We employed six of the most commonly used metrics for medical-report generation to evaluate model performance, including BLEU-n [45], METEOR [46], and ROUGE-L [47], where BLEU-n refers to the four n-gram-based measures (BLEU-1 through BLEU-4). Specifically, BLEU quantifies exact n-gram matches, METEOR takes word-order information into account, and ROUGE-L is based on the longest common subsequence. These natural-language metrics can, to some extent, reflect the quality of the generated reports [48]. To further assess clinical correctness, we also conducted a detailed case analysis in Section 4.5.

Table 3
Datasets information.

Datasets	Images	Views	Reports
IU X-ray	7470	Multi	3955
MIMIC-CXR	377,110	Multi	227,827

4.3. Baselines

We compare our method with six state-of-the-art image captioning and medical report generation models: (i) our re-implementation of the top-down model [18], which is a classic encoder-decoder-based model for image captioning employing a conventional attention mechanism that calculates the contribution of regional features to the texts to be generated, and (ii) MRMA [10], an encoder-decoder-based model specially designed for medical report generation, in which reports are generated sentence by sentence with a recurrent way to generate long paragraphs. (iii) RTMIC [14], which is a state-of-the-art medical report generation method based on reinforcement learning, enhancing the capacity of the generation model with reinforcement learning, and increasing the clinical accuracy with a transformer. (iv) X-LAN [19], which is an image captioning model employing x-linear attention and improving it with reinforcement learning. As image captioning is similar to our task to some extent, we also take this model as our baseline. (v) HReMRG-MR [9], which is a medical report generation model that utilizes a hybrid reinforcement learning method and uses a high-order attention mechanism to repeat the penalty mechanism to improve reports. (vi) R2Gen [5], a memory unit-based medical report generation method. Models and memorizes similar patterns between reports, thereby facilitating Transformer to generate more informative long-text explanation reports. For our implemented methods, we use the same visual features and train/val/test split on both datasets.

4.4. Implementation details

We utilize ResNet-101 pre-trained on ImageNet [49] to extract 2048 dimensional region-level image features from the last convolutional layer. After being converted to visual embeddings of size 1024, the encoder exploration with four stacks of M-linear attention blocks yields high-order synthetic features. During the decoding process, we set the size of hidden layer, word embedding dimension, and the latent dimension of the projection layer to 1024. During training, we first pre-train the model with a batch size of 6 for 60 epochs using NVIDIA RTX 2080Ti GPUs. We set the base learning rate to $1e-4$, paired with a Norm decay strategy with 10,000 warm-up steps, and used the ADAM [50] optimizer. We set τ_c to 0.1 and τ_s to 0.3. Finally, we train the model with the batch size of 2 for 60 epochs of reinforcement learning [51] using beam search [52] with a beam size of 2 to further improve the model performance. We set the indicator-weighted mixed reward as our training reward [9], where the weights of BLEU-1, BLEU-2, BLEU-3, BLEU-4, METERO, and ROUGE-L, are 2, 2, 1, 1, 2, and 2, respectively; and the base learning rate is reduced to $1e-5$ and decayed by cosine annealing with a period of 15 epochs.

4.5. Main results

Table 4 shows the experimental results of our proposed C2M-DoT and six baselines on six natural language generation metrics, where all baselines are re-implemented by us. Furthermore, Fig. 4 presents some examples of reports generated by these models.

In general, C2M-DoT outperforms all state-of-the-art baselines among all natural language evaluation metrics in Table 4, and Fig. 4 shows that C2M-DoT also generates more comprehensive and accurate reports (with more matches). Specifically, in Table 4, Top-down and MRMA perform poorly in long text generation without reinforcement learning; as shown in Fig. 4, RTMIC and X-LAN cannot use high-order

Table 4

Experimental results of C2M-DoT and the state-of-the-art baselines on IU X-ray (upper part) and MIMIC-CXR (lower part). The best results are bold and the second best ones are underlined.

Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
IU X-ray	Top-down	0.2822	0.1866	0.1241	0.0830	0.1455	0.3330
	MRMA	0.3820	0.2520	0.1730	0.1200	0.1630	0.3090
	RTMIC	0.3448	0.2188	0.1484	0.1063	0.1509	0.2890
	X-LAN	0.3826	0.2724	0.1949	0.1405	0.1750	0.3441
	HReMRG-MR	0.4265	0.3025	0.2119	0.1502	0.1871	0.3608
	R2Gen	<u>0.4349</u>	0.2802	0.1868	<u>0.1510</u>	0.1773	0.3509
MIMIC-CXR	C2M-DoT (ours)	0.4579	0.3214	0.2302	0.1593	0.2037	0.3803
	Top-down	0.2371	0.1548	0.1201	0.0989	0.1352	0.3211
	MRMA	0.3610	0.2440	0.1820	0.1410	0.1570	0.3300
	RTMIC	0.3701	0.2490	0.1812	0.1299	0.1506	0.3276
	X-LAN	0.3656	0.2670	0.1881	0.1315	0.1703	0.3421
	HReMRG-MR	0.4696	<u>0.3251</u>	<u>0.2412</u>	0.1877	<u>0.1993</u>	0.3742
	R2Gen	<u>0.4700</u>	0.3098	0.2390	<u>0.1911</u>	0.1905	0.3609
	C2M-DoT (ours)	0.4842	0.3450	0.2579	0.1925	0.2098	0.3850

Image	Ground-Truth	MRMA	X-LAN	HReMRG-MR	C2M-DoT(Ours)
	no acute cardiopulmonary abnormality. the lungs are clear and without focal airspace opacity . the cardiomedastinal silhouette is normal in size and contour and stable.there is no pneumothorax or large pleural effusion.	no acute cardiopulmonary abnormality . the lungs are clear bilaterally . there is no pneumothorax or pleural effusion . the heart and mediastinum are within normal limits . bony structures are intact.	no acute cardiopulmonary findings . lungs are clear bilaterally . cardiac and mediastinal silhouettes are normal . pulmonary vasculature is normal . no pneumothorax or pleural effusion . no acute bony abnormalities.	no acute cardiopulmonary abnormality . the cardiomedastinal silhouette and pulmonary vasculature are normal in size . the lungs are clear . there is no pneumothorax or pleural effusion . no focal airspace consolidation. no acute bony findings.	no acute cardiopulmonary abnormality . the lungs are clear . there is no focal airspace opacity . the cardiomedastinal silhouette is within normal limits and contours are stable. no pleural effusion or pneumothorax . there are no acute bony abnormality.
	vascular congestion with probable mild interstitial pulmonary edema . no focal consolidations to suggest pneumonia. small bilateral pleural effusions, improved compared to prior. the patient is status post cabg. sternotomy wires are intact and appropriately aligned. there is moderate enlargement of the cardiomedastinal silhouette . no pneumothorax . there are no acute osseous abnormalities.	no acute cardiopulmonary process . the lungs are clear . the cardiomedastinal silhouette is within normal limits . atherosclerotic calcifications are noted at the aortic arch , no acute osseous abnormalities .	no acute cardiopulmonary process . the lungs are clear without focal consolidation . there is mild pulmonary edema and pleural effusion . the cardiomedastinal silhouette is within normal limits . median sternotomy wires and mediastinal clips are again noted . no acute osseous abnormalities .	mild cardiomegaly . no acute cardiopulmonary process . ap and lateral views of the chest . the patient is status post median sternotomy and cabg . the heart size is mildly enlarged . the mediastinal and hilar contours are unchanged . there is mild pulmonary edema . the lungs are clear . there is no focal consolidation , pleural effusion or pneumothorax is seen . there are no acute osseous abnormalities . degenerative changes are noted in the thoracic spine .	cardiomegaly with mild pulmonary vascular congestion and small bilateral pleural effusions . no evidence of pneumonia . ap upright and lateral views of the chest , the patient is status post median sternotomy wires are again noted . there is mild pulmonary edema and pleural effusion . the lungs are clear . there is no focal consolidation or pneumothorax . the cardiac silhouette and the mediastinal are enlarged . the hilar contours are stable . no acute osseous abnormality is seen in the right.

Fig. 4. Example of reports generated by our C2M-DoT model and baselines. The upper case is from the IU X-ray dataset, and the lower case is from the MIMIC-CXR dataset. The matching medical keywords are shown in bold.

attention modules to capture visual features for multimodal reasoning, and the report accuracy is low; R2Gen achieves the highest BLEU score among all baselines due to the use of memory units to generate coherent reports; HReMRG-MR uses hybrid reinforcement learning and generally improves on most metrics, and is the best performer on METEOR and ROUGE-L among all baselines. On this basis, C2M-DoT achieves the best results on all metrics because (i) our multi-view contrastive learning adequately performs multi-view mutual information learning to obtain superior performance, (ii) the same input distribution for multi-view training and single-view testing is maintained, and the task gap between contrastive learning and generation branches is narrowed, avoiding the domain shift problem. (iii) Cross-modal consistency optimization makes inference report semantics and image semantics consistent.

In addition, the visualization results in Fig. 4 corroborate these findings. Specifically, the reports generated by MRMA for the two cases are noticeably shorter and contain fewer keyword matches than those of the other methods, indicating that the RNN-based architecture struggles with complex long-text generation. X-LAN benefits from the Transformer framework and achieves longer reports with more keywords, yet it introduces false positives. For instance, in the MIMIC-CXR case the ground-truth report describes the cardiomedastinal silhouette as “moderate enlargement”, whereas X-LAN incorrectly states it

is “within normal limits”. HReMRG-MR exhibits similar errors: the ground truth notes “small bilateral pleural effusions”, but HReMRG-MR claims “no pleural effusion”. In contrast, C2M-DoT correctly captures “small bilateral pleural effusions”, with no missed findings or misinterpretations in either case. By leveraging multi-view information and cross-modal optimization, C2M-DoT is better at detecting subtle lesions and expressing them in clinically appropriate language. For example, in the IU X-ray case it produces the precise phrase “the cardiomedastinal silhouette is within normal limits and contours are stable”, which matches radiology terminology”.

4.6. Ablation study

In this section, we report on a series of ablation experiments, using C2M-DoT and five incrementally implemented intermediate models to show the effectiveness of using the proposed multi-view contrastive learning, domain transfer module and cross-modal consistency in our work. Specifically, we implement five incrementally implemented intermediate models: (i) We take the reinforcement learning-based report generation model as the base model, and use the concatenated features of different views as the input, called Base-Cat; (ii) introduce a multi-view contrastive learning branch on the base model, called

Table 5

Automatic natural language evaluation on IU X-ray (upper part) and MIMIC-CXR (lower part). Ablation studies, where MvCo indicates multi-view contrastive learning, DoT indicates domain transfer network, and CMC indicates cross-modal consistency. The best results are bold and the second best ones are underlined.

Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
IU X-ray	Base-Cat	0.4175	0.2813	0.1915	0.1400	0.1820	0.3604
	MvCo-Cat	0.4373	0.3062	0.2139	0.1482	0.1933	0.3609
	MvCo-Fus	0.4440	0.3130	0.2196	0.1571	0.1953	0.3698
	MvCo-DoT	0.4533	0.3180	0.2228	0.1568	0.1958	0.3743
	MvCo-CMC	0.4581	<u>0.3193</u>	<u>0.2245</u>	<u>0.1583</u>	0.1934	<u>0.3772</u>
	C2M-DoT (Ours)	<u>0.4579</u>	0.3214	0.2302	0.1593	0.2037	0.3803
MIMIC-CXR	Base-Cat	0.4380	0.2995	0.2132	0.1626	0.1817	0.3647
	MvCo-Cat	0.4521	0.3153	0.2389	0.1704	0.1896	0.3670
	MvCo-Fus	0.4671	0.3211	0.2362	0.1668	0.1935	0.3697
	MvCo-DoT	0.4698	<u>0.3286</u>	0.2416	0.1792	<u>0.1984</u>	0.3823
	MvCo-CMC	<u>0.4772</u>	0.3268	0.2423	<u>0.1856</u>	0.1908	0.3859
	C2M-DoT (Ours)	0.4842	0.3450	<u>0.2579</u>	<u>0.1925</u>	0.2098	<u>0.3850</u>

MvCo-Cat; (iii) A multi-view contrastive learning model using fused features from different views as input, called MvCo-Fus; (iv) Using a domain transfer network capable of adaptively selecting inputs based on multi-view contrastive learning, called MvCo-DoT; (v) introducing a cross-modal consistency loss based on multi-view contrastive learning, called MvCo-CMC. In Table 5 we compare the results of the above six models.

4.6.1. Effectiveness of multi-view contrastive learning

By comparing the results of Baes-Cat and MvCo-Cat, we find that using contrastive learning on multi-view medical images in the multi-view base model makes MvCo-Cat significantly outperform Baes-Cat on all natural language metrics. This finding demonstrates that the mutual information mined by the proposed multi-view contrastive learning helps focus on salient lesions, explore deep semantic features and enable multimodal reasoning. In addition, MvCo-Fus has further improved the results compared to MvCo-Cat, which also shows that the fusion of visual features of different views as the input of the generative model is more suitable for multi-view generation tasks than direct Concatenating. Concatenating different views will double the width of multi-view input features, while fusion features can keep the same width as single-view features. MvCo-Fus narrows the input distribution difference between the contrastive learning branch for single-view input and the generation branch for multi-view input, thus achieving better performance.

Additionally, the effectiveness of multi-view contrastive learning can be visualized in Fig. 5, which shows the semantic embeddings for paired multi-view medical image (frontal, lateral) decoding. We randomly select 50 pairs of images from the IU-X-ray dataset. Through the multi-view comparison learning model, the report semantics generated by the medical image is decoded, and mapped to the same latent space to obtain the feature embedding.

Then, t-SNE is used to reduce the dimensionality of the features in these high-dimensional spaces to represent them in 2D images. The closer the frontal semantic (purple) and corresponding lateral semantic (blue) embeddings are, the better the learned multi-view consistency features are. We observe that frontal and side-reported semantic features for the same patient are closer in the latent space of MvCo-Fus compared to Base-Cat. Therefore, the use of multi-view contrastive learning effectively enhances the learning of mutual information between multiple views to decode consistent semantic feature embeddings.

4.6.2. Effectiveness of domain transfers

Next, we use MvCo-Fus and MvCo-CMC as baselines to compare with C2M-DoT and MvCo-DoT, respectively, to demonstrate the effectiveness of domain transfer networks. In Table 5, the results show that whether it is the single-modal optimization model MvCo-DoT or the multi-modal optimization model C2M-DoT, the results after using the domain transfer network are better than the baselines. The adaptive

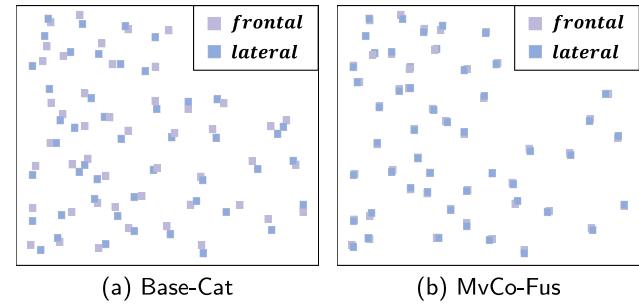


Fig. 5. Comparison of semantic feature embeddings between Base-Cat and MvCo-Fus in latent space.

input selection module selects the most suitable input for the model according to the actual situation to maximize the learning benefits; at the same time, since the generative model is provided with the option of single-view feature input during the training process, the multi-view generation process will turn to single-view reasoning under a certain probability. The more identical input distribution narrows the task gap between the generative branch and the multi-view contrastive learning branch (single-view inference), and the model obtains the optimal representation of both, enabling the two processes to promote each other.

Furthermore, the effectiveness of the domain transfer module can be visualized in Fig. 6, which shows the performance comparison of MvCo-Fus and MvCo-CMC when using single-view and multi-view inputs before (e.g.(a)(c)) and after (e.g.(b)(d)) introducing the domain transfer network. Our further observation: Compared with MvCo-Fus and MvCo-CMC, MvCo-DoT and C2M-DoT introduce a domain transfer network for adaptive input selection during training, which solves the problem of domain shift caused by different input distributions during multi-view training and single-view testing. The models can improve cross-domain transferability and can generate high-score reports given any view as input.

4.6.3. Effectiveness of cross-modal consistency

We then use MvCo-Fus and MvCo-DoT as baselines to compare with MvCo-CMC and C2M-DoT, respectively, to demonstrate the effectiveness of the cross-modal consistency loss. As shown in Table 5, after introducing cross-modal consistency loss, the performance of most natural language metrics of MvCo-CMC and C2M-DoT are further improved, which means that a consistent semantic representation allows the reported text output to be accurate and reliable (closer to the semantics of images).

As shown in Fig. 7, we visualized the multimodal semantic similarity matrices of frontal image reports and profile image reports of different eras on the IU X-ray dataset. The higher the semantic

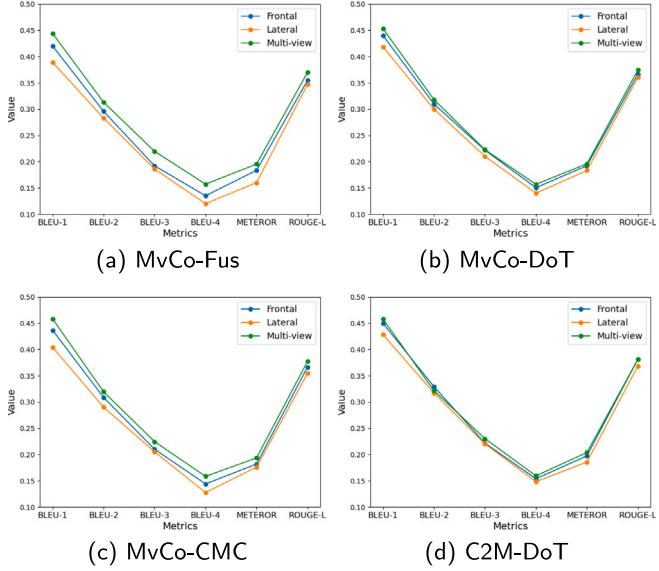


Fig. 6. Performance comparison of models using various view inputs.

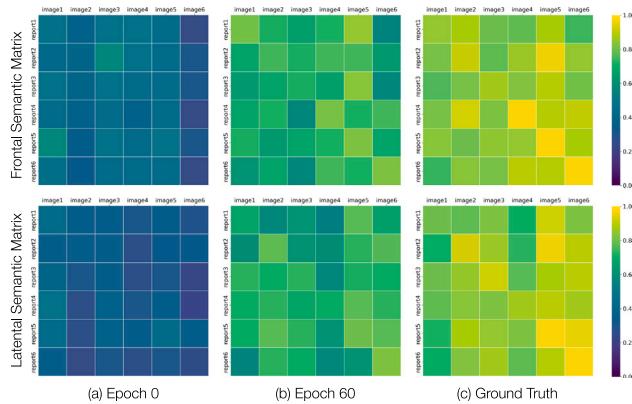


Fig. 7. Variation of Multimodal Semantic Similarity Matrix During Training Phase.

similarity between the report and the image, the lighter the color of the corresponding position in the matrix, otherwise the darker the color. It can be seen that with the increase of training epochs, the graphic-text similarity matrix of the frontal reasoning report and the graphic-text similarity matrix of the lateral reasoning report are more and more similar to the corresponding real report graphic-text similarity matrix. This fully demonstrates the role of our proposed cross-modal consistency loss in image-text matching learning. At the same time, the color distinction in the matrix is gradually obvious, indicating that the multimodal consistency loss helps to generate more sample-individualized report results.

4.7. Additional results

4.7.1. Effects of using multi-view contrastive learning in different positions

In order to further verify the rationality of multi-view contrastive learning based on semantic features, we compared the impact of using contrastive learning at different locations on the model, as shown in Fig. 8. Base-FS represents a fully supervised medical report generation model that does not use multi-view contrastive learning; MvCo-Encoder represents contrastive learning using the feature vectors of the frontal and lateral images output by the encoder; MvCo-Decoder represents using the semantic feature vectors of frontal and lateral reports output by the decoder for comparative learning.

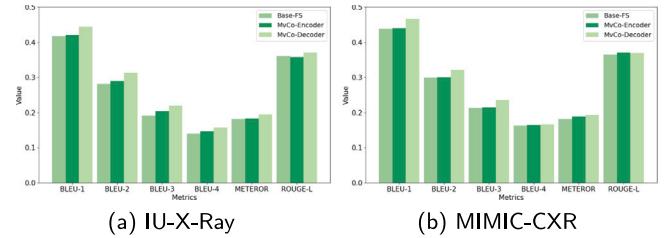


Fig. 8. Comparison of the effect of using multi-view contrastive learning in different positions.

We found that MvCo-Encoder performed better than Base-FS on all metrics, which shows that the mutual information between different views does help focus and understand the lesion characteristics. However, multi-view contrastive learning based on visual features does not significantly improve the performance of report generation. This may be because the spatial features of different views are too different and the visual consistency is limited; in addition, simply updating the encoder to optimize visual feature representation does not help much in end-to-end generation tasks.

When multi-view contrastive learning is used on the decoded semantic vectors, the performance of MvCo-Decoder is further improved. Due to the high degree of agreement between the semantic features decoded from frontal and lateral views of the same patient, more accurate reports were able to be generated. All findings demonstrate that the closer the position using multi-view contrastive learning is to the output, the better the effect. Finally, we use semantic-based multi-view contrastive learning in our research.

4.7.2. Effect of different input sampling methods

In order to further study the rationality of adaptive input selection methods in domain transfer networks, we compared models using three sampling methods: *random*, *argmax*, and *gumbel*, called DoT-Random, DoT-Argmax, and DoT-Gumbel, respectively. Table 6 shows the results of their inference using frontal- or lateral-view, or multi-view inputs.

Specifically, DoT-Argmax achieved the highest scores on BLEU metrics when using multi-view inference, but the performance dropped significantly when using frontal or lateral view inference alone. *Argmax* only samples the input option with the highest probability to participate in training, and the probability value is directly determined by the input information. Since multi-view input tend to contain more feature information (i.e. obtain greater probabilities) than individual frontal- and lateral views, single-view data rarely has a chance to participate in training. The variety of inputs is limited, and thus the improvement on the domain shift problem is limited.

In contrast, the performance of DoT-Random has been further improved when using frontal-view for reasoning. The *random* sampling method ensures the diversity and comprehensiveness of the input, and can effectively alleviate the problem of domain shift. However, due to the inability to select the appropriate input according to the input feature information, the overall reasoning ability obtained is not good.

Finally, we found that DoT-Gumbel achieved the highest scores on almost all metrics when using frontal- and lateral-view reasoning alone, and also achieved the highest METEOR and ROUGE-L results for multi-view reasoning. This is because *gumbel* uses the current input information and adaptively samples based on probability, which can not only select appropriate input features for the model, but also expand the input distribution to a certain extent. Therefore, we adopt *gumbel*-based input sampling in our domain transfer network research.

Table 6
Comparison of different input sampling methods for domain transfer networks.

Dataset	Model	Input		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
		Frontal	Lateral						
IU X-ray	DoT-Argmax	✓	✓	0.4647	0.3296	0.2317	0.1603	0.1926	0.3722
		✓		0.4404	0.2992	0.2176	0.1449	0.1862	0.3636
			✓	0.4283	0.3171	0.2202	0.1482	0.1857	0.3680
	DoT-Random	✓	✓	0.4385	0.2845	0.1751	0.1239	0.1811	0.3124
		✓		0.4382	0.2859	0.2095	0.1392	0.1688	0.3259
			✓	0.4108	0.2762	0.1743	0.1138	0.1631	0.3225
MIMIC-CXR	DoT-Gumbel	✓	✓	0.4579	0.3214	0.2302	0.1593	0.2037	0.3803
		✓		0.4498	0.3291	0.2216	0.1538	0.1973	0.3813
			✓	0.4679	0.3291	0.2302	0.1593	0.2037	0.3813
	DoT-Argmax	✓	✓	0.4874	0.3490	0.2611	0.1984	0.1915	0.3640
		✓		0.4748	0.3189	0.2372	0.1726	0.1906	0.3576
			✓	0.3934	0.2892	0.1777	0.1244	0.1607	0.3005
	DoT-Random	✓	✓	0.4629	0.3205	0.2055	0.1639	0.1801	0.3324
		✓		0.4637	0.3153	0.2268	0.1769	0.1702	0.3222
			✓	0.4491	0.2974	0.2050	0.1437	0.1681	0.3109
	DoT-Gumbel	✓	✓	0.4842	0.3450	0.2579	0.1925	0.2098	0.3850
		✓		0.4712	0.3446	0.2518	0.1916	0.2048	0.3871
			✓	0.4520	0.3305	0.2401	0.1907	0.1996	0.3624

Table 7
Results of varying different consistency losses.

Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
IU X-ray	Base- \mathcal{L}_{CE}	0.4175	0.2813	0.1915	0.1400	0.1820	0.3604
	CMC- \mathcal{L}_{CL}	0.4518	0.3091	0.2180	0.1483	0.1872	0.3665
	CMC- \mathcal{L}_{MSE}	0.4190	0.2983	0.2036	0.1361	0.1912	0.3700
	CMC- \mathcal{L}_{JS}	0.4452	0.3131	0.2135	0.1466	0.1944	0.3740
	CMC- \mathcal{L}_{KL}	0.4579	0.3214	0.2302	0.1593	0.2037	0.3803
MIMIC-CXR	Base- \mathcal{L}_{CE}	0.4380	0.2995	0.2132	0.1626	0.1817	0.3647
	CMC- \mathcal{L}_{CL}	0.4633	0.3153	0.2369	0.1901	0.2008	0.3710
	CMC- \mathcal{L}_{MSE}	0.4672	0.3205	0.2254	0.1894	0.1942	0.3640
	CMC- \mathcal{L}_{JS}	0.4790	0.3300	0.2347	0.1904	0.1922	0.3779
	CMC- \mathcal{L}_{KL}	0.4842	0.3450	0.2579	0.1925	0.2098	0.3850

4.7.3. Effect of using different consistency losses

In order to further study the rationality of cross-modal consistency optimization, we compare the effect of using different loss functions. As shown in Table 7, we use the cross-entropy loss-optimized medical report generation method Base- \mathcal{L}_{CE} as the baseline, and additionally implement four models of cross-modal loss: (i) CMC- \mathcal{L}_{CL} : First, the cross-modal semantic similarity matrix is obtained by using the visual features of the image and the semantic features of the prediction report, and then follow the method in CLIP to make it consistent with the diagonal matrix for contrastive learning (CL) (ii) CMC- \mathcal{L}_{MSE} : Additionally compute a cross-modal semantic similarity matrix between the visual features of an image and the ground-truth reported semantic features, using Mean Squared Error (MSE) for both matrices (iii) CMC- \mathcal{L}_{JS} : For two cross-modal semantic similarity The matrix uses Jensen-Shannon Divergence (JS divergence) (iv) CMC- \mathcal{L}_{KL} : Use Kullback-Leibler Divergence (KL divergence) for two cross-modal semantic similarity matrices.

Intuitively, Base- \mathcal{L}_{CE} yields the worst results among all losses, which strongly supports our previous theoretical analysis that optimization of unimodality using cross-entropy losses ignores reporting semantics and cannot fully optimized. Therefore, we are motivated to introduce a semantic consistency loss across modalities. It can be seen that CMC- \mathcal{L}_{CL} outperforms Base- \mathcal{L}_{CE} by a large margin on all six evaluation metrics by using contrastive learning across pairs of images and texts for cross-modal semantics. However, this improvement is limited. Since the images and reports of different patients may have the same semantics, the one-to-one matching mechanism of contrastive learning often lacks the ability to handle one-to-many samples. In contrast, the three models CMC- \mathcal{L}_{MSE} , CMC- \mathcal{L}_{JS} and CMC- \mathcal{L}_{KL} that perform consistent calculations on two cross-modal semantic similarity

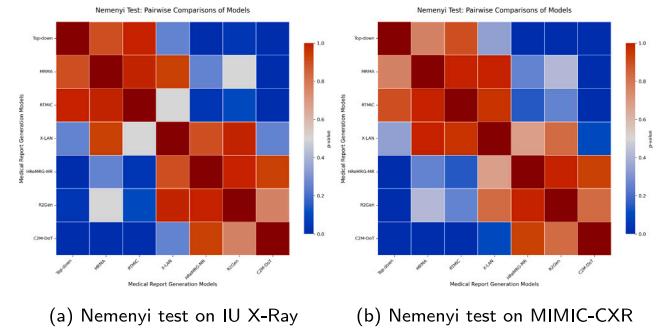


Fig. 9. Nemenyi test results on IU X-ray and MIMIC-CXR.

matrices obtain better scores, and since they achieve accurate and flexible cross-modal optimization by realizing the consistent distribution of the semantic similarity matrix of predicted report and vision and the semantic similarity matrix of real report and vision. Specifically, CMC- \mathcal{L}_{KL} is more suitable for the medical report generation task and achieves the best performance due to its ability to focus on the fine-grained differences in the distribution rather than the overall similarity of the matrix. Ultimately, we choose to use KL divergence to optimize for cross-modal semantic consistency.

4.8. Statistical analysis of MRG methods

To ensure the objectivity and reliability of the performance comparison among the seven medical report generation methods in our experiments, we perform statistical significance tests (Friedman test, Nemenyi test, and Wilcoxon test) to determine whether the differences in performance among the methods are statistically significant.

Based on the results in Table 4, we first conducted the Friedman test. The resulting Friedman statistic value is 33.8571 for IU X-ray and 33.7857 for MIMIC-CXR dataset. With 6 metrics and 7 methods, the statistic follows a Chi-squared (χ^2) distribution with $k - 1 = 6$ degrees of freedom. The critical value of $\chi^2(6)$ for $\alpha = 0.05$ is 12.592. Since our calculated statistics both exceed this critical value, we reject the null hypothesis for both datasets at the 0.05 significance level, indicating the performance of multi methods significantly differs.

Friedman Test indicates that there are overall differences among groups, it does not specify which groups are different. To further investigate those differences we use Nemenyi test and Wilcoxon test.

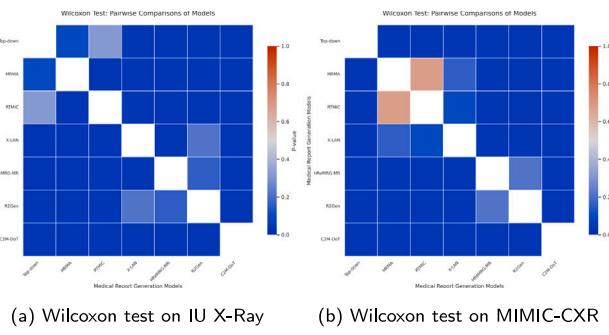


Fig. 10. Wilcoxon test results on IU X-ray and MIMIC-CXR.

Table 8
The computational complexity and the inference time results.

	Params (M)	GFLOPs	Latency (ms)
HReMRG-MR	59.61	4.63	63.49
C2M-DoT	72.20	5.31	73.45

Table 9

Memory and training time results on different datasets with and without cross-modal consistency optimization model.

Dataset	Model	Memory (G)	Time
IU X-ray	w/o cross-modal consistency	5.36	4 m 25 s
	w/ cross-modal consistency (Ours)	6.19	5 m 58 s
MIMIC-CXR	w/o cross-modal consistency	7.73	1 h 45 m 26 s
	w/ cross-modal consistency (Ours)	8.94	2 h 36 m 34 s

As shown in Fig. 9, the p-values represent the pairwise comparisons between different methods, and the diagonal value 1.0 indicate that each method is being compared with itself. To determine the significance between methods, we compare the p-values to an α value of 0.5. If the p -value is less than α , we reject the null hypothesis for that comparison. A higher p -value indicates that the difference between two methods is not statistically significant. For example, the p -value for C2M-DoT and X-LAN is 0.25, indicating a significant difference in their performance.

For further analysis, we performed a pairwise Wilcoxon test. The significance level is set to 0.5. If the observed p -value is less than 0.5, the null hypothesis of the Wilcoxon test is rejected. As shown in Fig. 10, The overall results from the Wilcoxon test suggest that there are significant differences in performance between the models.

4.9. Calculation overhead

Considering the impact of computational resources on clinical deployment, we further analyzed the computational overhead of C2M-DoT. Specifically, we evaluated the model's efficiency and resource requirements during inference using three metrics: Params, GFLOPs, and Latency, which respectively quantify model size, computational complexity, and inference speed. To ensure fairness, we performed inference on 100 randomly selected X-rays from IU X-ray one by one and reported the average results. All experiments were conducted under the same hardware and training configurations. We selected the HReMRG-MR model, which shares the same backbone as ours, as the comparison baseline to better highlight the relationship between the computational overhead and performance gains introduced by our proposed MvCo-DoT, and CMC. The results are summarized in Table 8.

As Table 8 shows, C2M-DoT incurs a moderate increase in overhead compared to HReMRG-MR. Specifically, C2M-DoT has 72.20M parameters versus 59.61M ($\uparrow 21.1\%$). At the same time, its GFLOPs rises from 4.63 to 5.31 ($\uparrow 14.7\%$). Average per-image latency increases from 63.49 ms to 73.45 ms ($\uparrow 15.7\%$). Despite these increases, C2M-DoT

yields substantial quality improvements: as Table 4 shows, on IU X-ray, BLEU-4 climbs from 0.1502 to 0.1593 ($\uparrow 6.1\%$), METEOR from 0.1871 to 0.2037 ($\uparrow 8.9\%$), and ROUGE-L from 0.3608 to 0.3803 ($\uparrow 5.4\%$).

Second, for the training phase, we primarily focused on the overhead introduced by incorporating large-scale pretrained models (e.g., ViT). Specifically, we compared peak GPU memory usage (Memory) and per-epoch training time (Time) during the first training stage on two datasets, with and without the CMC module. The results are shown in Table 9.

As Table 9 shows, the introduction of CMC does lead to an increase in computational resource consumption during training, but it remains feasible to train the model on a single NVIDIA 2080Ti GPU. On IU X-ray, peak memory rises from 5.36 GB to 6.19 GB ($\uparrow 15.5\%$), and training time from 4 m 25 s to 5 m 58 s ($\uparrow 35\%$). On MIMIC-CXR, memory increases from 7.73 GB to 8.94 GB ($\uparrow 15.6\%$), and training time from 1 h 45 m 26 s to 2 h 36 m 34 s ($\uparrow 47\%$). Likewise, this also brings performance gains to the model: as shown in Fig. 7, CMC leads to higher cross-modal semantic similarity between the generated reports and the ground-truth reports. In addition, the model also achieves better performance on natural language metrics. As shown in Table 5, compared to MvCo-DoT, C2M-DoT with CMC achieves a 0.3% improvement in BLEU-4, 0.8% in METEOR, and 0.6% in ROUGE-L on IU X-ray. On MIMIC-CXR, BLEU-4 improves by 1.3%, METEOR by 1.1%, and ROUGE-L by 0.3%.

From the perspective of clinical deployment, the increased computational overhead remains within a controllable range. First, C2M-DoT has a moderate model size (72M parameters), and its average per-image inference latency is around 0.07 s, fully meeting clinical timeliness requirements. In addition, the CMC module, which brings a relatively large training overhead, is only used during training and does not affect inference speed or memory usage during deployment. In practical scenarios, model compression, quantization, or distillation techniques [4] can be further applied to ensure that the model delivers superior report generation performance while meeting the efficiency and scalability requirements of real-world deployment.

5. discussion and future work

At present, publicly available medical report generation datasets are overwhelmingly focused on chest X-rays, so our method has been designed and evaluated primarily for this modality. However, in routine clinical practice a wide range of imaging techniques, including computed tomography (CT), magnetic resonance imaging (MRI) and ultrasound, are equally indispensable. Because these examinations differ in clinical intent and hardware configuration, they often yield only partial views. As a result, the domain-shift problem produced by mismatched data distributions between training and inference is not limited to chest X-rays, and addressing it is therefore of broad clinical significance [11,12]. To this end, in future work, we plan to extend C2M-DoT to additional imaging modalities and multimodal datasets [53]. CT and MRI consist of multiple slices or sequences, and the multi-view contrastive learning scheme together with the domain-adaptation module in C2M-DoT can in principle be transferred to these settings. Nevertheless, because the current framework is tailored to two-dimensional X-ray imagery, suitable feature-extraction networks must be chosen to accommodate the three-dimensional nature of CT and MRI. Furthermore, slice-to-slice correspondence is less straightforward than in X-rays, which may require redefining positive and negative pairs for contrastive learning.

Secondly, our approach leverages large-scale pretrained models such as ViT to speed up convergence and boost the initial performance. Although the ViT encoder in C2M-DoT is kept frozen and only extracts visual features, this design choice can still limit deployment in resource-constrained settings. Simply substituting a lighter encoder risks a drop in accuracy, so future work will explore techniques like knowledge distillation [4] and pruning [54] to maintain strong performance while

improving suitability for low-resource environments. Furthermore, the semantic representations learned during pretraining may not transfer smoothly to other data domains. When computational resources permit, selectively fine-tuning certain layers could better align the model with the target distribution while preserving the advantages of prior knowledge.

Moreover, although C2M-DoT was trained on several public datasets, the risks of data imbalance and sampling bias remain. In future work we will gather additional cross-center, cross-ethnic and cross-lingual data. By learning population characteristics more comprehensively and objectively, the model should become more robust when transferred to new medical institutions, diverse imaging devices and heterogeneous patient cohorts. We also plan to incorporate medical knowledge bases to enrich clinical context and to improve interoperability in multilingual environments [2,55]. Integrating these resources with large language models is expected to deepen the system's grasp of complex medical terminology and discourse.

Finally, ethical compliance and risk management are critical in any clinical application. We emphasize that the present model is intended to assist clinicians rather than supplant their professional judgment. Because most existing metrics cannot fully capture the clinical quality of radiology reports, and because false positives and false negatives may arise during inference, we advocate the establishment of a rigorous human review pipeline [56]. Going forward we will enhance model interpretability by providing attention-map visualizations or causal graphs that reveal decision pathways and salient regions. These tools will enable clinicians to audit and correct model outputs swiftly, thereby minimizing diagnostic delays and reducing medical risk.

6. Conclusions

In this paper, to overcome the above problems, we propose a cross-modal consistent multi-view medical report generation with domain transfer network (C2M-DoT). A semantic-based multi-view contrastive learning is proposed to mine the mutual information between different views of medical images to generate more accurate reports; moreover, we also propose to use a domain transfer network based on adaptive input selection to overcome the input distribution gap between the training and inference stages to make the multi-view medical report generation model adaptable to various single-view reasoning. Multimodal optimization based on cross-modal consistency is also used to help text-image matching. We conduct extensive experiments on publicly available datasets IU X-ray and MIMIC-CXR, demonstrating the superiority and effectiveness of our proposed method.

CRediT authorship contribution statement

RuiZhi Wang: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Zhenghua Xu:** Writing – review & editing, Writing – original draft, Supervision, Resources, Methodology, Funding acquisition, Conceptualization. **Xiangtao Wang:** Writing – review & editing, Validation. **Weipeng Liu:** Writing – review & editing, Supervision, Funding acquisition. **Thomas Lukasiewicz:** Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under the grants 62276089 and 62027813, by the Natural Science Foundation of Tianjin, China, under the grants 24JCJQJC00200, by the Natural Science Foundation of Hebei Province, China, under the grant F2024202064, by the Oversea Returning High-Level Talents Fund of Ministry of Human Resources and Social Security, China, under the grant RSTH-2023-135-1, and by the S&T Program of Hebei, China, under the grant 24464401D. This work was also partially supported by the AXA Research Fund, France.

Data availability

The links of the data used in this experiments have been included in the paper.

References

- [1] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training, 2018.
- [2] K. Singhal, S. Azizi, T. Tu, S.S. Mahdavi, J. Wei, H.W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pföhrl, et al., Large language models encode clinical knowledge, *Nature* 620 (7972) (2023) 172–180, <http://dx.doi.org/10.1038/s41586-023-06291-2>.
- [3] M.U. Hadi, R. Qureshi, A. Shah, M. Irfan, A. Zafar, M.B. Shaikh, N. Akhtar, J. Wu, S. Mirjalili, et al., A survey on large language models: Applications, challenges, limitations, and practical usage, *Authorea Prepr.* (2023) <http://dx.doi.org/10.36227/techrxiv.23589741.v1>.
- [4] F. Liu, X. Wu, S. Ge, W. Fan, Y. Zou, Exploring and distilling posterior and prior knowledge for radiology report generation, in: *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13753–13762, <http://dx.doi.org/10.1109/CVPR46437.2021.01354>.
- [5] Z. Chen, Y. Song, T.-H. Chang, X. Wan, Generating radiology reports via memory-driven transformer, 2020, <http://dx.doi.org/10.18653/v1/2020.emnlp-main.112>, arXiv preprint [arXiv:2010.16056](https://arxiv.org/abs/2010.16056).
- [6] B. Hou, G. Kaassis, R.M. Summers, B. Kainz, Ratchet: Medical transformer for chest X-ray diagnosis and reporting, in: *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021, pp. 293–303, http://dx.doi.org/10.1007/978-3-030-87234-2_28.
- [7] Y.N.T. Vu, R. Wang, N. Balachandar, C. Liu, A.Y. Ng, P. Rajpurkar, Medaug: Contrastive learning leveraging patient metadata improves representations for chest x-ray interpretation, in: *Proceedings of Machine Learning for Healthcare*, 2021, pp. 755–769, <http://dx.doi.org/10.48550/arXiv.2102.10663>.
- [8] A.B. Amjoud, M. Amrouch, Automatic generation of chest X-ray reports using a transformer-based deep learning model, in: *Proceedings of the International Conference on Intelligent Computing in Data Sciences*, 2021, pp. 1–5, <http://dx.doi.org/10.1109/icds53782.2021.9626725>.
- [9] W. Xu, Z. Xu, J. Chen, C. Qi, T. Lukasiewicz, Hybrid reinforced medical report generation with M-linear attention and repetition penalty, 2022, <http://dx.doi.org/10.1109/TNNLS.2023.3343391>, arXiv preprint [arXiv:2210.13729](https://arxiv.org/abs/2210.13729).
- [10] Y. Xue, T. Xu, L. Rodney Long, Z. Xue, S. Antani, G.R. Thoma, X. Huang, Multi-modal recurrent model with attention for automated radiology report generation, in: *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2018, pp. 457–466, http://dx.doi.org/10.1007/978-3-030-00928-1_52.
- [11] J. Zhang, S. Zhang, X. Shen, T. Lukasiewicz, Z. Xu, Multi-ConDoS: Multimodal contrastive domain sharing generative adversarial networks for self-supervised medical image segmentation, *IEEE Trans. Med. Imaging* (2023) <http://dx.doi.org/10.1109/tmi.2023.3290356>.
- [12] S. Zhang, J. Zhang, B. Tian, T. Lukasiewicz, Z. Xu, Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation, *Med. Image Anal.* 83 (2023) 102656, <http://dx.doi.org/10.1016/j.media.2022.102656>.
- [13] G. Liu, T.M.H. Hsu, M. McDermott, W. Boag, W.H. Weng, P. Szolovits, M. Ghassemi, Clinically accurate chest X-Ray report generation, 2019, <http://dx.doi.org/10.48550/arXiv.1904.02633>, arXiv preprint [arXiv:1904.02633](https://arxiv.org/abs/1904.02633).
- [14] Y. Xiong, B. Du, P. Yan, Reinforced transformer for medical image captioning, in: *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019, pp. 673–680, http://dx.doi.org/10.1007/978-3-03-32692-0_77.
- [15] A. Yan, Z. He, X. Lu, J. Du, E. Chang, A. Gentili, J. McAuley, C.-N. Hsu, Weakly supervised contrastive learning for chest x-ray report generation, 2021, <http://dx.doi.org/10.48550/arXiv.2109.12242>, arXiv preprint [arXiv:2109.12242](https://arxiv.org/abs/2109.12242).

- [16] Y. Zhang, H. Jiang, Y. Miura, C.D. Manning, C.P. Langlotz, Contrastive learning of medical visual representations from paired images and text, in: Proceedings of the Machine Learning for Healthcare, 2022, pp. 2–25, <http://dx.doi.org/10.48550/arXiv.2010.00747>.
- [17] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen, et al., Big self-supervised models advance medical image classification, in: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, 2021, pp. 3478–3488, <http://dx.doi.org/10.1109/ICCV48922.2021.00346>.
- [18] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, in: Proceedings of Conference on Computer Vision and Pattern Recognition, 2018, pp. 6077–6086, <http://dx.doi.org/10.1109/cvpr.2018.00636>.
- [19] Y. Pan, T. Yao, Y. Li, T. Mei, X-linear attention networks for image captioning, in: Proceedings of Conference on Computer Vision and Pattern Recognition, 2020, pp. 10971–10980, <http://dx.doi.org/10.1109/CVPR42600.2020.01098>.
- [20] S. Ramineni, S. Shridevi, D. Won, Multi-modal transformer architecture for medical image analysis and automated report generation, *Sci. Rep.* 14 (1) (2024) 19281, <http://dx.doi.org/10.1038/s41598-024-69981-5>.
- [21] Z. Wang, H. Han, L. Wang, X. Li, L. Zhou, Automated radiographic report generation purely on transformer: A multicriteria supervised approach, *IEEE Trans. Med. Imaging* 41 (10) (2022) 2803–2813, <http://dx.doi.org/10.1109/tmi.2022.3171661>.
- [22] O. Alfarghaly, R. Khaled, A. Elkorany, M. Helal, A. Fahmy, Automated radiology report generation using conditioned transformers, *Informatics Med. Unlocked* 24 (2021) 100557, <http://dx.doi.org/10.1016/j.imu.2021.100557>.
- [23] M. Kong, Z. Huang, K. Kuang, Q. Zhu, F. Wu, Transq: Transformer-based semantic query for medical report generation, in: International Conference on Medical Image Computing and Computer Assisted Intervention, 2022, pp. 610–620, http://dx.doi.org/10.1007/978-3-031-16452-1_58.
- [24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, <http://dx.doi.org/10.48550/arXiv.2010.11929>, arXiv preprint <arXiv:2010.11929>.
- [25] B. Jing, P. Xie, E. Xing, On the automatic generation of medical imaging reports, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018, pp. 2577–2586, <http://dx.doi.org/10.18653/v1/p18-1240>.
- [26] M.U. Akhtar, J. Liu, Z. Xie, X. Cui, X. Liu, B. Huang, Multilingual entity alignment by abductive knowledge reasoning on multiple knowledge graphs, *Eng. Appl. Artif. Intell.* 139 (2025) 109660, <http://dx.doi.org/10.1016/j.engappai.2024.109660>.
- [27] J. Yuan, H. Liao, R. Luo, J. Luo, Automatic radiology report generation based on multi-view image fusion and medical concept enrichment, in: Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention, 2019, pp. 721–729, http://dx.doi.org/10.1007/978-3-030-32226-7_80.
- [28] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: Proceedings of International Conference on Machine Learning, 2020, pp. 1597–1607, <http://dx.doi.org/10.48550/arXiv.2002.05709>.
- [29] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738, <http://dx.doi.org/10.1109/CVPR42600.2020.00975>.
- [30] Y. Tian, D. Krishnan, P. Isola, Contrastive multiview coding, in: Proceedings of the European Conference on Computer Vision, 2020, pp. 776–794, http://dx.doi.org/10.1007/978-3-030-58621-8_45.
- [31] T. Gao, X. Yao, D. Chen, Simcse: Simple contrastive learning of sentence embeddings, 2021, <http://dx.doi.org/10.48550/arXiv.2104.08821>, arXiv preprint <arXiv:2104.08821>.
- [32] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: Proceedings of International Conference on Machine Learning, 2021, pp. 8748–8763, <http://dx.doi.org/10.48550/arXiv.2103.00020>.
- [33] D. You, F. Liu, S. Ge, X. Xie, J. Zhang, X. Wu, Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation, in: Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention, 2021, pp. 72–82, http://dx.doi.org/10.1007/978-3-030-87199-4_7.
- [34] S.C. Huang, L. Shen, M.P. Lungren, S. Yeung, Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition, in: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, 2021, pp. 3942–3951, <http://dx.doi.org/10.1109/iccv48922.2021.00391>.
- [35] C. Seibold, S. Reiß, M.S. Sarfraz, R. Stiefelhagen, J. Kleesiek, Breaking with fixed set pathology recognition through report-guided contrastive training, in: Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention, 2022, pp. 690–700, http://dx.doi.org/10.1007/978-3-031-16443-9_66.
- [36] Z. Wang, Z. Wu, D. Agarwal, J. Sun, Medclip: Contrastive learning from unpaired medical images and text, 2022, <http://dx.doi.org/10.18653/v1/2022.emnlp-main.256>, arXiv preprint <arXiv:2210.10163>.
- [37] C. Wu, X. Zhang, Y. Zhang, Y. Wang, W. Xie, Medclip: Medical knowledge enhanced language-image pre-training, *MedRxiv* (2023) 2001–2023, <http://dx.doi.org/10.1101/2023.01.10.23284412>.
- [38] S. Eslami, C. Meinel, G. de Melo, PubMedCLIP: How much does CLIP benefit visual question answering in the medical domain? in: Proceedings of the Association for Computational Linguistics, 2023, pp. 1181–1193, <http://dx.doi.org/10.18653/v1/2023.findings-eacl.88>.
- [39] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778, <http://dx.doi.org/10.48550/arXiv.1512.03385>.
- [40] V. Nair, G.E. Hinton, Rectified linear units improve restricted boltzmann machines, in: Proceedings of International Conference on Machine Learning, 2010, pp. 807–814.
- [41] E. Jang, S. Gu, B. Poole, Categorical reparameterization with gumbel-softmax, 2016, <http://dx.doi.org/10.48550/arXiv.1611.01144>, arXiv preprint <arXiv:1611.01144>.
- [42] O. Pelka, S. Koitka, J. Rückert, F. Nensa, C.M. Friedrich, Radiology objects in context (ROCO): a multimodal image dataset, in: Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis, 2018, pp. 180–189, http://dx.doi.org/10.1007/978-3-030-01364-6_20.
- [43] D. Demner-Fushman, M.D. Kohli, M.B. Rosenman, S.E. Shooshan, L. Rodriguez, S. Antani, G.R. Thoma, C.J. McDonald, Preparing a collection of radiology examinations for distribution and retrieval, *J. Am. Med. Informatics Assoc.* 23 (2) (2016) 304–310, <http://dx.doi.org/10.1093/jamia/ocv080>.
- [44] A.E. Johnson, T.J. Pollard, S.J. Berkowitz, N.R. Greenbaum, M.P. Lungren, C.y. Deng, R.G. Mark, S. Horng, MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports, *Sci. Data* 6 (2019) 317, <http://dx.doi.org/10.1038/s41597-019-0322-0>.
- [45] K. Papineni, S. Roukos, T. Ward, W.J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002, pp. 311–318, <http://dx.doi.org/10.3115/1073083.1073135>.
- [46] S. Banerjee, A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in: Proceedings of the Second Workshop on Statistical Machine Translation, 2005, pp. 65–72, <http://dx.doi.org/10.3115/1626355.1626389>.
- [47] C.Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Proceedings of the Association for Computational Linguistics, 2004, pp. 74–81.
- [48] J.-B. Delbrouck, P. Chambon, C. Bluethegn, E. Tsai, O. Almusa, C. Langlotz, Improving the factual correctness of radiology report generation with semantic rewards, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the Association for Computational Linguistics: Empirical Methods in Natural Language Processing, 2022, pp. 4348–4360, <http://dx.doi.org/10.18653/v1/2022.findings-emnlp.319>.
- [49] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Proceedings of Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255, <http://dx.doi.org/10.1109/cvpr.2009.5206848>.
- [50] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, <http://dx.doi.org/10.48550/arXiv.1412.6980>, arXiv preprint <arXiv:1412.6980>.
- [51] S.J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, V. Goel, Self-critical sequence training for image captioning, in: Proceedings of Conference on Computer Vision and Pattern Recognition, 2017, pp. 7008–7024, <http://dx.doi.org/10.1109/cvpr.2017.131>.
- [52] A.K. Vijayakumar, M. Cogswell, R.R. Selvaraju, Q. Sun, S. Lee, D. Crandall, D. Batra, Diverse beam search: Decoding diverse solutions from neural sequence models, 2016, <http://dx.doi.org/10.48550/arXiv.1610.02424>, arXiv preprint <arXiv:1610.02424>.
- [53] X. Wang, R. Wang, J. Zhou, T. Lukasiewicz, Z. Xu, AMLP: Adaptive masking lesion patches for self-supervised medical image segmentation, 2023, <http://dx.doi.org/10.48550/arXiv.2309.04312>, arXiv preprint <arXiv:2309.04312>.
- [54] X. Lin, L. Yu, K.T. Cheng, Z. Yan, The lighter the better: rethinking transformers in medical image segmentation through adaptive pruning, *IEEE Trans. Med. Imaging* 42 (8) (2023) 2325–2337, <http://dx.doi.org/10.1109/TMI.2023.3247814>.
- [55] M.U. Akhtar, J. Liu, Z. Xie, X. Liu, S. Ahmed, B. Huang, Entity alignment based on relational semantics augmentation for multilingual knowledge graphs, *Knowl.-Based Syst.* 252 (2022) 109494, <http://dx.doi.org/10.1016/j.knosys.2022.109494>.
- [56] R. Tanno, D.G. Barrett, A. Sellergren, S. Ghaisas, S. Dathathri, A. See, J. Welbl, C. Lau, T. Tu, S. Azizi, et al., Collaboration between clinicians and vision-language models in radiology report generation, *Nature Med.* 31 (2) (2025) 599–608, <http://dx.doi.org/10.1038/s41591-024-03302-1>.