Hi-MrGn: Hierarchical Medical Report Generation Network

Anonymous EMNLP submission

Abstract

Numerous deep learning (DL)-based approaches have been developed for medical report generation (MRG), aiming to automate the description of medical images. These reports typically comprise two sections: the findings, which describe visual aspects of the images, and the impression, which summarizes the diagnosis or assessment. Given the distinct abstraction levels of these sections, conventional end-to-end DL methods that generate both simultaneously may not be optimal. Addressing this challenge, we introduce a novel Hierarchical Medical Report Generation Network (Hi-MrGn) designed to better reflect the inherent structure of medical reports. The Hi-MrGn operates in two stages: initially, it generates the findings from input multimodal data including medical images and auxiliary diagnostic texts; subsequently, it produces the impression based on both the findings and images. To enhance the semantic coherence between findings and impression, we incorporate a contrastive learning module within the Hi-MrGn. We validate our approach using two public X-ray image datasets, MIMIC-CXR and IU-Xray, demonstrating that our method surpasses current state-of-the-art (SOTA) techniques in this domain.

1 Introduction

Medical reports are essential in routine clinic. However, for radiologists, medical report writing is a time-consuming and labor-intensive task. Medical report generation (MRG), which can produce reports from input medical images automatically, is highly desired and of great clinical significance. MRG is a special field of image captioning, and in the advance of deep learning (DL), many DL-based image captioning methods have been proposed with success (Vinyals et al., 2015; Karpathy and Fei-Fei, 2015; Anderson et al., 2018; Krause et al., 2017; Cornia et al., 2020).

Despite the great achievement of existing DL-based MRG methods, most of them, as illustrated in Fig. 1(a), generate both findings and impression simultaneously through a one-pass approach using the same features learned through the same deep learning paths. It is known that medical reports are structured by findings and impression. The findings in medical reports provide visual descriptions of medical images, detailing aspects such as the anatomical shape, position, and size of lesions. In contrast, the impression section entails the deduction and final decision-making process, embodying a higher abstract level of semantic information compared to the findings. Existing one-pass generation strategy ignores the semantic hierarchy inherent in radiology reports. Therefore, as shown in Fig. 1(b), we claim that the findings and impression in medical reports should be generated hierarchically to avoid the mixture of information at different abstract levels.

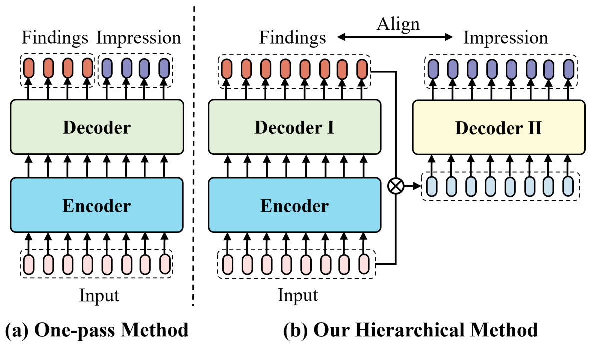

Figure 1: Comparison between one-pass and hierarchical strategies for medical report generation. (a) Existing one-pass methods jointly generate findings and impression, ignoring their semantic hierarchy. (b) Our hierarchical approach adopts a two-stage process with explicit semantic alignment between the two sections.

Some existing works have attempted hierarchical generation strategies (Srinivasan et al., 2020), where findings and impression are generated independently without explicitly modeling their underlying reasoning relationship. As a result, semantic consistency between findings and impression cannot be guaranteed, which may introduce hallucinated content in the impression (Jiang et al., 2025).

Based on the above observation, in this paper, we propose a new DL-based MRG method, the Hierarchical Medical Report Generation Network (Hi-MrGn). The Hi-MrGn separates the generation process of findings and impression, where features learned from input medical images and auxiliary diagnostic texts (e.g., reason for examination) are used for generating the findings, then features from the findings are refined with visual features to produce the impression. A disease classification branch is adopted as an auxiliary task to guide the generation of both findings and impression. Moreover, a contrastive-learning module is integrated in the Hi-MrGn to make the separately generated findings and impression have semantic consistency. In the experiment, two public datasets MIMIC-CXR (Johnson et al., 2019) and IU-Xray (Demner-Fushman et al., 2016) are used. The experimental results show that the proposed method outperforms the state-of-the-art (SOTA) methods. The main contributions of our method are listed below:

- We propose a novel hierarchical MRG framework (Hi-MrGn), explicitly designed to reflect the inherent semantic order of radiology reports by generating the findings and impression in two stages, aligning with real-world clinical reporting practices.
- To bridge the semantic gap between the separately generated findings and impression, we introduce a co-attention module and a contrastive learning module to enforce semantic consistency across the two stages.
- We conduct comprehensive experiments on two datasets, demonstrating that Hi-MrGn outperforms state-of-the-art baselines in both language generation and clinical accuracy.

2 Related Work

2.1 Medical Report Generation

MRG methods adopt encoder–decoder frameworks similar to image captioning. Specifically, the encoder is responsible for extracting visual features from input images, based on which sequence of words describing the input images can be generated by the decoder. For example, Chen *et al.* proposed an attention-based decoder with a relational memory module to record key information during generation. This memory-driven Transformer achieved better language fluency and higher clinical accuracy (Chen et al., 2020b). Jin *et al.* proposed a diagnosis-driven prompting framework that integrates a disease classifier into the generation process. Predicted disease labels are converted into discrete prompt tokens, which guide the Transformer decoder to generate more clinically accurate content (Jin et al., 2024). In addition, various mechanisms in MRG, including reinforcement learning (RL) (Miura et al., 2021; Zhou et al., 2024) and knowledge graph techniques (Li et al., 2023), are employed to enhance the accuracy or fluency of the generated reports.

2.2 Multimodal Learning in MRG

Medical report generation is fundamentally a vision-to-language task, recent studies have shown that incorporating additional textual or semantic information along with images can significantly improve performance. Two representative strategies have emerged. The first one introduces intermediate supervision through semantic tags. For example, the AlignTransformer model first predicts a set of disease tags from the chest X-ray and then uses those tags to guide report generation (You et al., 2021). By integrating an image-derived text representation(the tags), the model was shown to reduce the bias towards describing only normal observations. The other strategy is to enrich input representations through external knowledge or retrieved text. They leverage structured resources such as medical knowledge graphs, existing clinical reports and auxiliary diagnostic texts. For example, knowledge graphs are typically encoded as structured embeddings or relational memory, which are injected into the model to enhance clinical reasoning (Liu et al., 2021; Huang et al., 2023). In (Jin et al., 2024), clinical reports are incorporated through retrieval-augmented generation (RAG) frameworks, where similar cases are retrieved and fused with the current input. In (Nguyen et al., 2021; Liu et al., 2025), auxiliary diagnostic texts, such as the reason for examination, are integrated via multimodal encoders that align textual and visual features, which often follow the architecture of vision-language models (VLMs).

2.3 Hierarchical Generation in MRG

As aforementioned, most MRG models generate the entire report in a one-pass approach. Recently,

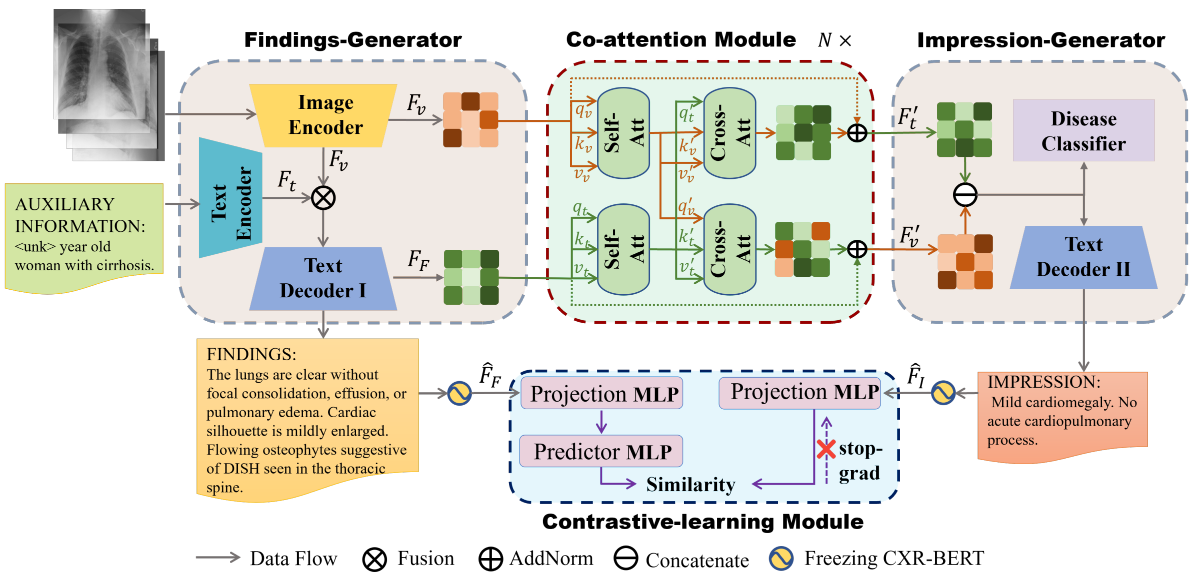

Figure 2: Structure of the Hi-MrGn. It is composed of a findings generator, a co-attention module, an impression generator, and a contrastive-learning module. Visual and textual features derived in the findings generator are refined by the co-attention module for generating the impression in the impression generator. Semantic consistency of the generated findings and impression is ensured by the contrastive-learning module.

a few works have explored hierarchical generation strategies. For example, ORGAN (Hou et al., 2023) adopts a two-stage plan-then-generate approach: it first generates a list of key observations and then elaborating them into a full report. However, ORGAN's observation plan covers only findings and does not explicitly generate an impression section or summary of those findings. In (Srinivasan et al., 2020), a hierarchical Transformer based MRG is proposed. It first generates the findings and then produces the impression based on them, reflecting the hierarchical structure of radiology reports. Our work is closely related to this method but extends this idea in two important aspects: (1) we incorporate original visual features via a co-attention module for richer context, rather than generating the impression solely on the findings and intermediate tag embeddings; and (2) we introduce a contrastive learning module to explicitly align the semantic representations of findings and impression.

3 Methods

The structure of Hi-MrGn is shown in Fig. 2. It is composed of four main components, i.e., the findings generator, the co-attention module, the impression generator, and the contrastive-learning module. The findings generator learns visual features F_v from the input medical images and textual features F_t from auxiliary diagnostic text for generating the findings. Concerning the impression generation, F_v and the findings-related features F_F in the findings generator are further fed to the co-attention module, where self-attention and cross-attention blocks are adopted to explore higher level features, i.e., F_v' and F_F', based on which the impression can be generated by the impression generator. Additionally, F_v' and F_F' are utilized by a disease classification branch to predict the presence of diseases, serving as an auxiliary task to enhance the capacity of feature learning in both generators. The semantic consistency between the separately generated findings and impression is enhanced by the contrastive-learning module. It is worth noting that the Hi-MrGn can be regarded as a hierarchical generation framework, and its image encoder and text decoder can be replaced by any existing ones. Details of each components are discussed below.

3.1 The Findings Generator

The findings generator is a typical encoder-decoder network with multimodal inputs. The encoder is of a dual-branch structure. The first branch is a pre-trained ResNet (He et al., 2016) based image encoder which learns visual features F_v from the input medical images I, and F_v is organized as a set of patch tokens, i.e, $F_v = \{x_1, x_2, ..., x_S\}$, where S denotes the number of patches and $x_i \in \mathbb{R}^d$ represents the visual feature of a patch. To provide rich clinical context and guide report generation with prior knowledge (Nguyen et al., 2021; Liu et al., 2025), a text encoder is added as the second branch, where auxiliary diagnostic texts is encoded through BERT's embedding layer (Devlin et al., 2019) to obtain textual embeddings $F_t \in \mathbb{R}^{N\times d}$.

To enable semantic integration of visual and textual representations, we first concatenated the patch-wise image features and the diagnostic text embeddings to form a joint token sequence $F = [F_v; F_t] \in \mathbb{R}^{(S+N)\times d}$. This fused sequence is then transformed through a stack of L standard Transformer encoder layer:

$$\tilde{F}^{(\ell)} = \mathrm{MSA}(\mathrm{LN}(F^{(\ell-1)})) + F^{(\ell-1)}, \quad (1)$$

$$F^{(\ell)} = \mathrm{FFN}(\mathrm{LN}(\tilde{F}^{(\ell)})) + \tilde{F}^{(\ell)}, \quad \ell = 1, 2, ..., L, \quad (2)$$

where $\mathrm{MSA}(\cdot)$ denotes multi-head self-attention, $\mathrm{FFN}(\cdot)$ is a two-layer feed-forward network and $\mathrm{LN}(\cdot)$ represents layer normalization. We use $F_{\text{fused}} = F^{(L)}$ as the final fused representation.

In the text decoder I, we utilize a Transformer decoder-based architecture to generate the findings. Specifically, the hidden state for each word position $h_i \in \mathbb{R}^d$ in the findings is computed based on the fused features and previous words:

$$h_i = \mathrm{Decoder}(F_{\text{fused}}, w_1, ..., w_{i-1}). \quad (3)$$

where $w_1, w_2, ..., w_{i-1}$ represent previous $i-1$ words. Based on the hidden states $H = \{h_i\}_{i=1}^{N_{\text{FD}}}$ (N_{FD} is the number of words in the findings), the words of findings can be determined, which is defined as:

$$P_{\text{FD}} = \mathrm{softmax}(HW^\top), \quad (4)$$

where $W \in \mathbb{R}^{N_w\times d}$ is the vocabulary matrix and N_w is the vocabulary size. $P_{\text{FD}}(i,j)$ represents the probability in choosing the j-th word from W for the i-th word in the generated findings. The cross entropy is used as the loss function of the findings generator, which is defined as:

$$\mathcal{L}_{\text{FD}} = -\frac{1}{N_{\text{FD}}} \sum_{i=1}^{N_{\text{FD}}} \sum_{j=1}^{N_w} Y_{\text{FD}}(i,j) \log P_{\text{FD}}(i,j), \quad (5)$$

where $Y_{\text{FD}}(i,j)$ is the ground truth. The decoder hidden states $H = \{h_i\}_{i=1}^{N_{\text{FD}}}$ also serve as the token-level semantic representation of findings, which we denote as F_{F} in the following co-attention module.

3.2 The Co-Attention Module

In clinical routine, radiologists can make the deduction and final decision (i.e., the impression) according to the findings and the medical images. Based on this observation and inspired by (Lu et al., 2019), a co-attention module is adopted. It employs attention mechanism that allows cross-modal learning between the findings (represented by the textual feature F_F) and the medical images (represented by the visual features F_v). The resulting representations, denoted as F'_F and F'_v, encode more abstract and complementary semantics for generating the accurate impression.

As shown in Fig. 2, the co-attention module is composed of N attention blocks. Each block contains two symmetrical sub-branches for textual and visual streams. Within each branch, a self-attention layer is first applied to encode intra-modal context, followed by a cross-attention layer that integrates information from the other modality. Formally, given input features F_{F} and F_v, one co-attention block proceeds as:

$$\tilde{F}_{\text{F}} = \mathrm{SelfAtt}_t(\mathrm{LN}(F_{\text{F}})) + F_{\text{F}}, \quad (6)$$

$$\tilde{F}_{\text{v}} = \mathrm{SelfAtt}_v(\mathrm{LN}(F_{\text{v}})) + F_{\text{v}}, \quad (7)$$

$$F'_{\text{F}} = \mathrm{CrossAtt}_t(\mathrm{LN}(\tilde{F}_{\text{F}}), \tilde{F}_{\text{v}}) + \tilde{F}_{\text{F}}, \quad (8)$$

$$F'_{\text{v}} = \mathrm{CrossAtt}_v(\mathrm{LN}(\tilde{F}_{\text{v}}), \tilde{F}_{\text{F}}) + \tilde{F}_{\text{v}}. \quad (9)$$

After passing through all co-attention blocks, the outputs $F'_{\text{F}} \in \mathbb{R}^{N_{\text{FD}}\times d}$ and $F'_{\text{v}} \in \mathbb{R}^{S\times d}$ are used as enriched representations for impression generation.

3.3 The Impression Generator

The output features of the co-attention module (F'_v and F'_{F}) are concatenated as the input of the impression generator (text decoder II) , based on which the impression can be obtained:

$$k_i = \mathrm{Decoder}(F'_v; F'_{\text{F}}; w_1, w_2, ..., w_{i-1}), \quad (10)$$

where k_i is the hidden state of each word in the impression. Following the same way as the findings generator, $K = \{k_i\}_{k=1}^{N_{\text{IP}}}$ (N_{IP} is the number of words in the impression) can be produced, based on which the final impression can be generated. The loss function of the generation of impression is similar as that used for the findings, i.e., the cross entropy as defined in (5).

Additionally, to enhance the learning capacity of related features, disease classification is added as an auxiliary task, where a classification head formed by fully connected layers ($d \rightarrow d/2 \rightarrow 14$) is integrated to predict the presence of 14 distinct thoracic diseases, such as atelectasis and lung opacity, which are widely recognized in the field of chest radiograph report generation (Smit et al., 2020). Binary Cross-Entropy is adopted as the loss function of the disease classifier.

3.4 The Contrastive-Learning Module

Since the findings and impression are generated separately, their semantic consistency is not explicitly enforced. To mitigate this, we incorporate a contrastive learning module into Hi-MrGn to enhance the semantic consistency between the two sections.

Conventional contrastive learning frameworks rely on both positive and negative pairs (Radford et al., 2021; Chen et al., 2020a). While positive pairs (i.e., findings and impression from the same image) are readily available, defining reliable negative pairs is ambiguous, as semantically related findings and impression from different samples may be incorrectly treated as negatives (Wang et al., 2022b). Thus, we adopt SimSiam (Chen and He, 2021) in the contrastive-learning module, which requires positive pairs only.

The module comprises a projection MLP $\mathcal{K}$ and a prediction MLP $\mathcal{H}$. The positive sample pair consists of two features, specifically $\hat{F}_{\mathrm{F}}$ and $\hat{F}_{\mathrm{I}}$, generated by CXR-BERT (Boecking et al., 2022) based on the findings and impression produced by the Hi-MrGn from identical medical images (refer to Fig.2). The corresponding loss is defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{sim}} =& \mathcal{D}(\mathcal{H}(\mathcal{K}(\hat{F}_{\mathrm{F}}))), \text{stopgrad}(\mathcal{K}(\hat{F}_{\mathrm{I}})) \\ &+\mathcal{D}\left(\mathcal{H}(\mathcal{K}(\hat{F}_{\mathrm{I}})), \text{stopgrad}(\mathcal{K}(\hat{F}_{\mathrm{F}}))\right),\end{aligned} \tag{11}$$

where $\mathcal{D}(x, y)$ is the cosine similarity, which is defined as:

$$\mathcal{D}(x, y) = -\frac{x}{\|x\|_2} \cdot \frac{y}{\|y\|_2}. \tag{12}$$

4 Experiments

4.1 Datasets

Two widely applied public datasets, i.e., the MIMIC-CXR (Johnson et al., 2019) and the IU-Xray (Demner-Fushman et al., 2016) are used in our experiment. Specifically, the MIMIC-CXR contains 377,110 chest X-ray images and corresponding medical reports of 65,379 patients. While the IU-Xray contains 7,470 chest X-ray images with medical reports of 3,955 patients. Each dataset is divided into training (70%), validation (10%), and testing (20%) sets, respectively.

4.2 Baselines

Besides the proposed Hi-MrGn, existing SOTA methods, including the TopDown (Anderson et al., 2018), the G-Trans (Lovelace and Mortazavi, 2020), the R2GenCMN (R-CMN) (Chen et al., 2021), the XPRONET (XPRO) (Wang et al., 2022a), the R2GEN (Chen et al., 2020b), the ORGan (Hou et al., 2023), and the PromptMRG(P-MRG) (Jin et al., 2024) are also evaluated.

4.3 Experimental Results

4.3.1 Language Generation Performance

In this section, six widely used natural language generation(NLG) metrics, including BLEU-1 (B-1) to BLEU-4 (B-4) (Papineni et al., 2002), METEOR (MTR) (Banerjee and Lavie, 2005) and ROUGE-L (R-L) (Lin, 2004), are adopted in our experiment. Since the findings and impression are generated separately in the Hi-MrGn, besides the evaluation of the whole generated medical reports, the findings and impression are also evaluated separately. Considering that the medical reports generated by the SOTA methods under evaluation have no clear division of findings and impression, we concatenate the findings and impression using a special delimiter token during traning, which allows us to easily divide the outputs into findings and impression during evaluation.

Evaluation results are shown in Table 1. Clearly, for the findings (F), the impression (I), and the whole medical report (F+I), the Hi-MrGn achieves superior or comparable performance to all SOTA methods.

4.3.2 Clinical Accuracy Performance

While NLG metrics assess the fluency and lexical similarity of generated reports, the ability to accurately identify diseases is crucial in MRG. Following prior works (Liu et al., 2024; Hou et al., 2023), we evaluate the clinical efficacy of our method on the MIMIC-CXR dataset using CheXpert-based metrics (Irvin et al., 2019), including micro-average Precision, Recall, and F1-score. As shown in Table 2, Hi-MrGn achieves the best performance among all baselines, with an F1-score of 0.467. We also observe that models such as ORGAN, P-MRG, and G-Trans achieve relatively high

Table 1: Evaluation results using SOTA methods and the Hi-MrGn on two datasets, where F, I and F+I, represent findings, impression, and the whole reports, respectively.

Methods	**Target**	**MIMIC-CXR**						**IU-Xray**					
		B-1	B-2	B-3	B-4	MTR	R-L	B-1	B-2	B-3	B-4	MTR	R-L
TopDown	F	0.322	0.205	0.142	0.105	0.133	0.281	0.384	0.244	0.161	0.112	0.190	0.322
	I	0.219	0.140	0.094	0.067	0.112	0.358	0.321	0.182	0.121	0.073	0.106	0.393
	F+I	0.321	0.205	0.141	0.103	0.134	0.284	0.404	0.260	0.169	0.116	0.184	0.332
R-CMN	F	0.347	0.221	0.152	0.112	0.144	0.288	0.458	0.296	0.203	0.148	0.188	0.352
	I	0.273	0.173	0.115	0.080	0.128	0.371	0.380	0.218	0.143	0.097	0.126	0.440
	F+I	0.350	0.222	0.152	0.110	0.145	0.291	0.460	0.303	0.205	0.147	0.184	0.363
R2GEN	F	0.352	0.222	0.153	0.113	0.142	0.285	0.467	0.289	0.209	0.160	0.182	0.357
	I	0.271	0.172	0.113	0.079	0.127	0.372	0.373	0.224	0.139	0.105	0.125	0.433
	F+I	0.355	0.224	0.153	0.111	0.144	0.290	0.446	0.297	0.204	0.145	0.173	0.342
G-Trans	F	0.362	0.229	0.158	0.116	0.147	0.288	0.478	0.309	0.210	0.149	0.188	0.346
	I	0.282	0.176	0.115	0.079	0.130	0.367	0.343	0.227	0.160	0.087	0.131	**0.467**
	F+I	0.365	0.230	0.158	0.115	0.148	0.292	0.481	0.319	0.218	0.153	0.189	0.360
P-MRG	F	0.371	0.226	0.149	0.105	0.147	0.268	0.395	0.236	0.160	0.113	0.156	0.310
	I	0.242	0.141	0.093	0.064	0.104	0.255	0.218	0.136	0.091	0.060	0.120	0.301
	F+I	0.369	0.224	0.147	0.103	0.144	0.265	0.430	0.270	0.182	0.127	0.164	0.325
XPRO	F	0.382	0.255	0.182	0.137	0.155	0.296	0.486	0.317	0.225	0.164	0.193	0.343
	I	0.278	0.174	0.114	0.078	0.129	0.367	0.389	0.220	0.128	0.088	0.127	0.446
	F+I	0.376	0.255	0.182	**0.136**	0.156	**0.339**	0.489	0.326	0.225	0.160	0.189	0.361
ORGan	F	0.391	0.257	0.181	0.134	0.157	0.322	0.495	0.325	0.228	0.170	**0.205**	0.367
	I	0.287	0.183	0.123	0.087	0.132	0.373	0.418	0.244	0.162	0.108	0.137	0.451
	F+I	0.387	0.258	0.183	0.135	0.158	0.335	0.496	0.331	0.229	0.168	0.200	0.338
Hi-MrGn	F	**0.415**	**0.292**	**0.221**	**0.176**	**0.182**	**0.340**	**0.506**	**0.335**	**0.240**	**0.178**	0.201	**0.385**
	I	**0.303**	**0.215**	**0.164**	**0.127**	**0.161**	**0.374**	**0.438**	**0.316**	**0.226**	**0.178**	**0.176**	0.451
	F+I	**0.392**	**0.262**	**0.185**	**0.136**	**0.160**	0.334	**0.514**	**0.353**	**0.256**	**0.189**	**0.205**	**0.408**

Table 2: Clinical efficacy comparison on the MIMIC-CXR dataset. We report micro-average Precision, Recall, and F1-score based on CheXpert labels for whole report generation (F+I).

Methods	Precision	Recall	F1-score
TopDown	0.315	0.270	0.291
R-CMN	0.342	0.310	0.325
R2GEN	0.353	0.300	0.324
G-Trans	0.421	0.375	0.397
P-MRG	0.492	0.420	0.453
XPRO	0.385	0.330	0.355
ORGan	0.463	0.405	0.432
Hi-MrGn	**0.518**	**0.455**	**0.467**

clinical accuracy due to their explicit incorporation of disease-specific guidance during training.

4.3.3 Ablation Study

The Hi-MrGn consists of several components, namely, the hierarchical structure for findings and impression generation (H), the multimodal input fusion (M), the co-attention mechanism for enhancing and fusing visual and textual features (Co-A), and the contrastive-learning module (CL) module. To show the contribution of each components, we need to answer the following questions: (1) Does the impression generation benefit from the H? (2) Is multimodal input fusion necessary for accurate findings generation? (3) Is the Co-A module helpful for two-modality feature fusion and enhancing? (4) What does the CL bring to medical report generation? Therefore, the ablation experiments are conducted in our study. Table 3 shows the evaluation results using the MIMIC-CXR and the IU-Xray datasets, and we come to the following conclusions for each component.

Impact of H. The H module provides the ben-

Table 3: Ablation study of the Hi-MrGn. 'H', 'M', 'Co-A', and 'CL' indicate the hierarchical structure, the multimodal input fusion, the co-attention module, and the contrastive-learning module, respectively.

Methods	**Target**	**MIMIC-CXR**						**IU-Xray**					
		B-1	B-2	B-3	B-4	MTR	R-L	B-1	B-2	B-3	B-4	MTR	R-L
Base	F	0.388	0.261	0.188	0.142	0.158	0.312	0.442	0.302	0.222	0.169	0.189	0.384
	I	0.257	0.161	0.111	0.078	0.132	0.369	0.389	0.263	0.187	0.108	0.121	0.422
	F+I	0.378	0.254	0.184	0.137	0.154	0.325	0.449	0.318	0.236	0.179	0.191	0.412
H	F	0.397	0.266	0.189	0.142	0.160	0.327	0.473	0.315	0.229	0.173	0.195	0.379
	I	0.276	0.197	0.145	0.119	0.152	0.373	0.412	0.288	0.203	0.158	0.163	0.435
	F+I	0.384	0.256	0.182	0.134	0.152	0.328	0.482	0.327	0.241	0.183	0.198	0.403
H w/o M	F	0.365	0.246	0.174	0.128	0.150	0.288	0.435	0.287	0.206	0.156	0.181	0.335
	I	0.275	0.173	0.114	0.078	0.130	0.370	0.408	0.280	0.195	0.148	0.157	0.428
	F+I	0.360	0.245	0.160	0.115	0.145	0.330	0.463	0.309	0.225	0.170	0.188	0.385
H+Co-A	F	0.401	0.266	0.188	0.141	0.159	**0.332**	0.481	0.322	0.235	0.176	0.197	0.388
	I	**0.284**	**0.205**	**0.158**	**0.124**	**0.160**	**0.375**	**0.432**	**0.305**	**0.220**	**0.170**	**0.172**	**0.448**
	F+I	0.386	0.255	0.178	0.130	0.156	0.330	0.490	0.335	0.249	0.190	0.203	0.415
H+CL	F	0.**406**	**0.272**	**0.196**	**0.147**	**0.163**	**0.332**	**0.492**	**0.330**	**0.240**	**0.182**	**0.200**	**0.396**
	I	0.282	**0.205**	0.156	0.122	0.158	0.372	0.425	0.298	0.216	0.165	0.168	0.441
	F+I	**0.391**	**0.261**	**0.185**	**0.140**	**0.154**	**0.332**	**0.498**	**0.342**	**0.252**	**0.193**	**0.207**	**0.418**

efit of generating findings and impression from a single model, which is closer to clinic routine than previous SOTA methods that generate the medical report simultaneously through the same network path. To evaluate the effectiveness of hierarchical generation, we compare with a baseline (Base) that generates findings and impression simultaneously using only the findings generator. Careful observation of Table 3 shows that the H module improves the generation quality, especially for the impression.

Impact of M. The findings generator incorporates both visual features from medical images and textual features from auxiliary diagnostic texts. By removing the textual input branch (H w/o M), we observe notable performance degradation, particularly in findings generation. This demonstrates that multimodal fusion provides richer context information for accurate report generation.

Impact of Co-A. The Co-A module enables advanced cross-modal learning by interleaving the textual features of findings with visual features from the corresponding X-ray image. In this way, both textual and visual features can be enhanced for good generation of the impression. Table 3 shows that the generation performance is improved to some extent in terms of all metrics when comparing H with H+Co-A.

Impact of CL. In the Hi-MrGn, the findings and impression are separately generated. Since the semantic information in them should be the same, the CL module is adopted in the Hi-MrGn to validate the consistency constraint. The evaluation results in Table 3 indicate that the consistency constraint can considerably improve the generated reports in terms of all metrics, demonstrating that the CL module indeed contributes to the performance.

4.3.4 Qualitative Analysis and Case Study

To further demonstrate the effectiveness of our proposed Hi-MrGn model in generating clinically accurate and semantically consistent medical reports, we present two representative cases in Fig. 3, comparing the generated reports from our model with those from ORGAN and G-Trans, alongside the ground-truth reports.

In the first case, the ground-truth report describes bilateral peribronchial consolidations. ORGAN, however, presents a clear inconsistency: the findings state *"the lungs are clear bilaterally"*, while the impression reports *"chronic left upper lobe atelectasis"*. This semantic conflict reflects the weakness of one-pass generation, which fails to align the factual content across sections. In contrast, Hi-MrGn maintains consistency throughout: the findings indicate *"peribronchial consolidations present, notably involving the left upper lobe"*, and the impression reinforces this with *"prominent left upper lobe involvement, consistent with a chronic pulmonary condition"*.

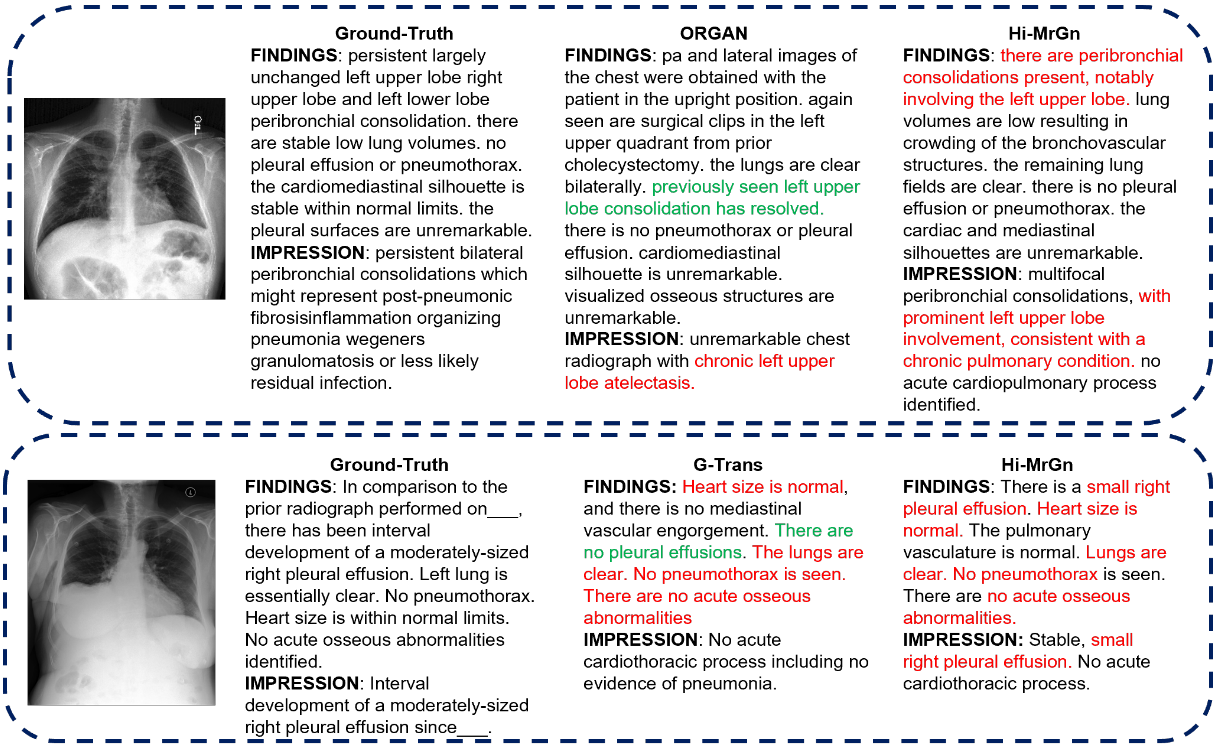

Figure 3: Examples of the reports using SOTA method and the Hi-MrGn. Red and green indicate consistent and inconsistent with the ground truth, respectively.

The second case illustrates Hi-MrGn's strength in accurate pathology recognition. The ground-truth impression highlights an *"interval development of a moderately-sized right pleural effusion"*, capturing the progression of the condition. G-Trans, in contrast, reports that *"there are no pleural effusions"*, and its impression incorrectly concludes with *"no acute cardiothoracic process including no evidence of pneumonia"*, missing the key pathology. Hi-MrGn accurately detects the effusion in the findings, stating *"there is a small right pleural effusion"*, and its impression reinforces this by noting a *"stable, small right pleural effusion"*, aligning well with the reference.

5 Conclusions

In this paper, we proposed a novel hierarchical medical report generation network (Hi-MrGn) to address the issue of generating medical reports, which have two parts different abstraction levels, i.e., the findings and impression. Specifically, the proposed Hi-MrGn generates the findings and impression in a hierarchical way, where the first stage focuses on generating the findings, while the second stage derives information related to the impression based on the refinement of the features learned at the first stage. Additionally, a contrastive learning module is introduced to ensure high semantic consistency between the generated findings and impression. Through the substantial experiments using two public datasets, the experimental results demonstrated that the proposed Hi-MrGn outperformed the latest state-of-the-art methods in both datasets. Further ablation study showed that all the proposed components, including the hierarchical framework, the multimodal input fusion, the co-attention module, and the contrastive-learning module, play effective roles in the medical report generation.

Limitations

Our current framework has some limitations. First, since the impression is generated based on the previously generated findings, its quality inherently depends on the accuracy and completeness of the findings, which may lead to error accumulation throughout the generation process. Second, our current framework is tailored for radiology report generation from chest X-ray images. Future work could explore the generalizability of this approach to other imaging modalities, such as computed tomography (CT) and magnetic resonance imaging (MRI), to extend its clinical applicability.

Ethics Statement

The MIMIC-CXR (Johnson et al., 2019) and IU-Xray (Demner-Fushman et al., 2016) datasets used in this study are publicly available and de-identified, ensuring no protected health information is involved. However, reports generated by Hi-MrGn may contain errors such as misdiagnoses or missed findings, which could impact clinical decisions. Therefore, model outputs should always be reviewed by medical professionals before use.

Similar to other deep learning models, Hi-MrGn may reflect biases in the training data. We encourage careful consideration of fairness and ethical implications when applying the model in real-world settings.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. 2022. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR.

Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5904–5914.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020b. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449, Online. Association for Computational Linguistics.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10578–10587.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Wenjun Hou, Kaishuai Xu, Yi Cheng, Wenjie Li, and Jiang Liu. 2023. ORGAN: Observation-guided radiology report generation via tree reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8108–8122.

Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. 2023. Kiut: Knowledge-injected u-transformer for radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19809–19818.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.

Yue Jiang, Jiawei Chen, Dingkang Yang, Mingcheng Li, Shunli Wang, Tong Wu, Ke Li, and Lihua Zhang. 2025. Comt: Chain-of-medical-thought reduces hallucination in medical report generation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. 2024. Promptmrg: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2607–2615.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–325.

Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. 2023. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3334–3343.

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

Chang Liu, Yuanhe Tian, Weidong Chen, Yan Song, and Yongdong Zhang. 2024. Bootstrapping large language models for radiology report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18635–18643.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13753–13762.

Kang Liu, Zhuoqi Ma, Xiaolu Kang, Yunan Li, Kun Xie, Zhicheng Jiao, and Qiguang Miao. 2025. Enhanced contrastive learning with multi-view longitudinal data for chest x-ray report generation. *arXiv preprint arXiv:2502.20056*.

Justin Lovelace and Bobak Mortazavi. 2020. Learning to generate clinically coherent chest x-ray reports. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1235–1243.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. 2021. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5288–5304, Online. Association for Computational Linguistics.

Hoang TN Nguyen, Dong Nie, Taivanbat Badamdorj, Yujie Liu, Yingying Zhu, Jason Truong, and Li Cheng. 2021. Automated generation of accurate\& fluent medical x-ray reports. *arXiv preprint arXiv:2108.12126*.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. 2020. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*.

Preethi Srinivasan, Daksh Thapar, Arnav Bhavsar, and Aditya Nigam. 2020. Hierarchical x-ray report generation via pathology tags and multi head attention. In *Proceedings of the Asian Conference on Computer Vision*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Jun Wang, Abhir Bhalerao, and Yulan He. 2022a. Cross-modal prototype driven network for radiology report generation. In *European Conference on Computer Vision*, pages 563–579. Springer.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022b. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2022, page 3876.

Di You, Fenglin Liu, Shen Ge, Xiaoxia Xie, Jing Zhang, and Xian Wu. 2021. Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 72–82. Springer.

Zijian Zhou, Miaojing Shi, Meng Wei, Oluwatosin Alabi, Zijie Yue, and Tom Vercauteren. 2024. Large model driven radiology report generation with clinical quality reinforcement learning. *arXiv preprint arXiv:2403.06728*.

A Implementation Details

We train Hi-MrGn using the AdamW optimizer with a learning rate of 5×10^{-5}, a minimum learning rate of 5×10^{-6}, and a warm-up learning rate of 5×10^{-7}. A linear warm-up followed by a cosine decay schedule (LinearWarmupCosineLRScheduler) is applied. The weight decay is set to 0.05, and the dropout rate is 0.1. The model is trained for 30 epochs with a batch size of 64.

The input image resolution is 224×224. The maximum sequence lengths are set to 100 for findings, 50 for impression, and 50 for history. The encoder is a pretrained ResNet-101 provided by PyTorch, and the decoder is a pretrained BERT model from HuggingFace, enhanced with a co-attention module. For contrastive learning, we adopt CXR-BERT as the sentence encoder to extract semantic embeddings of findings and impression.

The model has 544.97M trainable parameters. All experiments are conducted on a single NVIDIA RTX 3090 GPU.

Here are the pretrained models we used:

- BERT(base, uncased):

  https://huggingface.co/google-bert/bert-base-uncased

- CXR-BERT(BiomedVLP-CXR-BERT-general):

  https://huggingface.co/microsoft/BiomedVLP-CXR-BERT-general