VDGen: Visual and Diagnostic Representation Calibration for Radiology Report Generation

Anonymous ACL submission

Abstract

Automated radiology report generation (RRG) offers the potential to reduce clinical workload and enhance diagnostic consistency. However, existing models struggle with degraded visual representations caused by long-tailed lesion distributions, and suffer from limited alignment between image features and diagnostic semantics. We propose **VDGen**, a unified framework for calibrating visual and diagnostic representations to improve disease-aware report generation. VDGen integrates two complementary modules: **Vision Self-Equilibration (VSE)**, a self-supervised contrastive module that mitigates visual feature degradation by promoting structured representation learning; and **Disease Information Distillation (DID)**, a cross-modal distillation mechanism that uses diagnostic reports as teacher signals to guide the extraction of disease-sensitive semantics from visual features. Our end-to-end architecture incorporates a LoRA-adapted large language model (LLM) decoder to generate clinically accurate reports. Experiments on the IU-Xray and MIMIC-CXR datasets show that VDGen achieves state-of-the-art performance on MIMIC-CXR and maintains competitive results on IU-Xray. Code and models will be released upon acceptance.

1 Introduction

Automated radiology report generation (RRG) holds significant promise for alleviating radiologists' workload while improving the interpretability and consistency of diagnostic reports (Wang and Summers, 2012). With rapid advancements in artificial intelligence, automated RRG has emerged as a critical research frontier in multimodal AI. Existing studies have attempted to adapt image caption methods to RRG by training models to generate medical reports from chest X-ray images through supervised learning (Liu et al., 2021a). However, such direct migration faces inherent limitations due to the unique characteristics of radiological data (Gu et al., 2024).

As illustrated in Fig. 1(a), lesion-related pixels in chest X-ray images typically occupy only small regions, resulting in highly imbalanced intra-image distributions. At the dataset level (Fig. 1(b)), normal samples overwhelmingly dominate, limiting the model's exposure to abnormal findings. These dual imbalances—spatial sparsity within images and class imbalance across the dataset—collectively contribute to a long-tailed data distribution that challenges conventional vision-language models (Bu et al., 2024). This distributional bias leads to degraded visual representations, especially for abnormal regions, where traditional visual encoders struggle to extract discriminative features (Liu et al., 2021b). As shown in Fig. 1(c, left), embeddings of lesion areas are often disorganized in the latent space, which in turn confuses the report decoder and impairs the generation of clinically meaningful text. We argue that effective radiology report generation requires addressing both: (i) *intra-modal degradation* of visual features caused by distributional imbalance, and (ii) *inter-modal misalignment* between image features and diagnostic language semantics leads to difficulties in disease information extraction.

In this paper, we propose **VDGen**, a unified framework for Visual and Diagnostic representation calibration in radiology report generation. VDGen integrates two complementary modules: **Vision Self-Equilibration (VSE)** introduces a self-supervised contrastive mechanism that learns to pull together visual embeddings from similar pathological patterns while pushing apart those of dissimilar ones. By performing instance-level regularization within each batch, VSE mitigates the effects of long-tailed and sparse lesion distributions, enhancing intra-modal discrimination without relying on manual labels. **Disease Information Distillation (DID)** introduces a cross-modal distillation mechanism, where diagnostic reports serve as teacher signals to supervise the learning of disease-

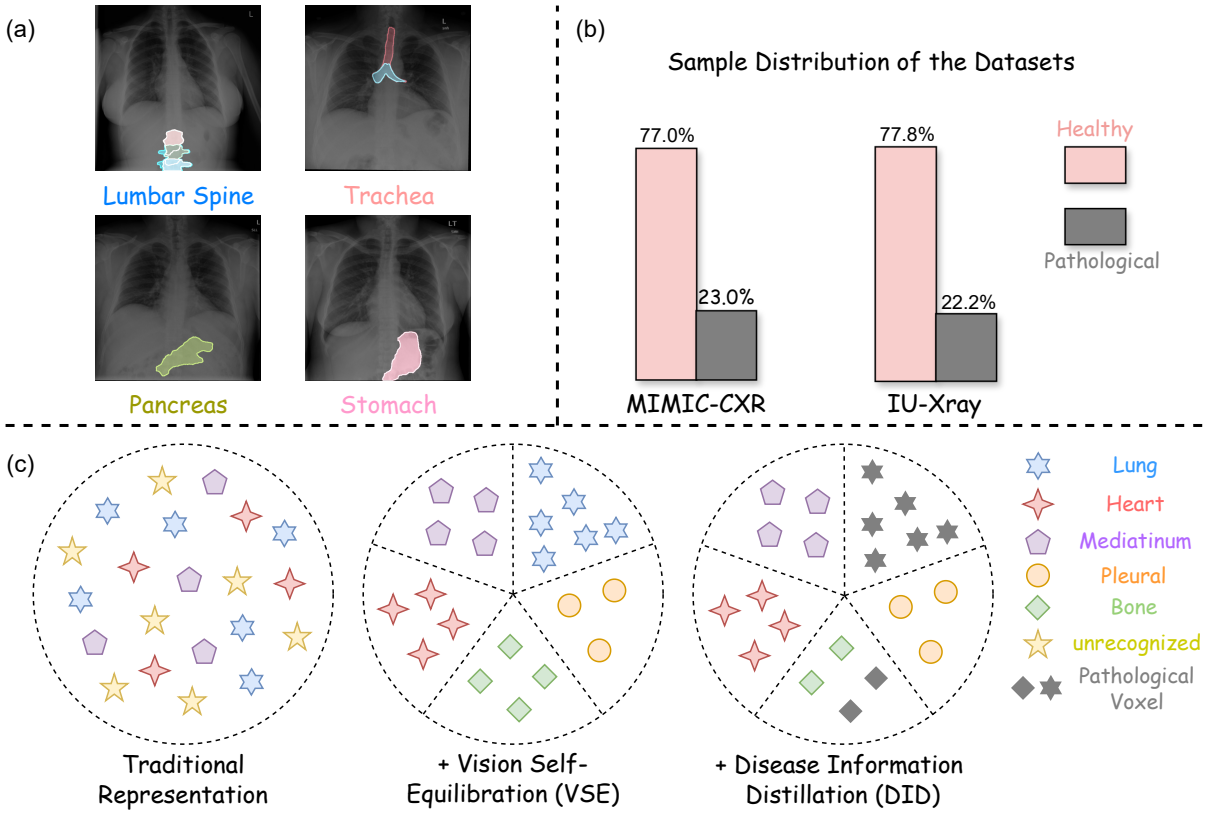

Figure 1: Subgraphs a and b demonstrate the scarcity of pathological samples in radiology report generation, and subgraph c shows the roles of VSE and DID.

sensitive visual semantics. Instead of aligning representations directly, DID performs bidirectional similarity optimization between image features and text-derived embeddings, enabling the visual encoder to internalize fine-grained diagnostic knowledge such as lesion presence and scope. As shown in Fig. 1(c, middle/right), these two mechanisms jointly refine the latent space, enabling more accurate and interpretable report synthesis. We validate VDGen on two widely used datasets, IU-Xray and MIMIC-CXR. Results show that VDGen achieves new state-of-the-art performance on MIMIC-CXR and remains competitive on the low-resource IU-Xray benchmark. Our contributions are summarized as follows:

- We propose **VDGen**, a unified framework for calibrating visual and diagnostic representations to improve the quality and interpretability of radiology report generation.
- We introduce **Vision Self-Equilibration (VSE)**, a self-supervised contrastive learning module that alleviates feature degradation by promoting intra-class compactness and inter-class separability in visual embeddings, without relying on manual labels.
- We develop **Disease Information Distillation (DID)**, a cross-modal distillation mechanism that leverages diagnostic reports as teacher signals to guide the learning of disease-aware visual semantics via bidirectional alignment.
- We conduct extensive experiments on the IU-Xray and MIMIC-CXR datasets. Results show that VDGen achieves state-of-the-art performance on MIMIC-CXR and remains competitive on IU-Xray, demonstrating both scalability and clinical relevance.

2 Related Work

2.1 Image Caption

Image caption aims to automatically generate text descriptions of input natural images. The common approach in this field is to use an end-to-end framework for caption generation. After the Transformer architecture became popular, researchers often used Swin-Transformer for image encoding, and then employed multi-layer Text Transformer to decode the hidden states and finally generate the

corresponding caption text. With the increasing demand for automated report generation, AI scholars have attempted to apply image captioning methods to the task of radiology report generation. However, due to the fundamental differences between natural images and radiological images, general caption models are difficult to be directly transferred to radiology report generation. In this paper, we propose VSE and DID to address the impact of the gap between natural images and radiological images on the performance of the generation model.

2.2 Radiology Report Generation

Radiology report generation is dedicated to helping radiologists reduce the workload of reading images. AI models are used to automatically analyze X-ray images and generate corresponding reports. With the rapid development of the AI field, the accuracy of radiology report generation has been continuously improved. Initially, the R2Gen model proposed by Alice et al. led the way in using memory-based Transformers for report analysis. This technology aimed to solve the consistency problem in report generation through the memory module and achieved good results, laying the foundation for R2Gen to be a baseline in the RRG field. Later, AI researchers discovered the difference in pixel distribution between X-ray image datasets and natural image datasets, and people began to focus on solving this problem. Some works using attention mechanisms to address this difference were proposed one after another, such as Contrastive (Liu et al., 2021b), COMG (Gu et al., 2024), etc. Subsequently, people tried to introduce some prior medical knowledge to improve the interpretability and reliability of model generation, such as PPKED, KiUT, etc. In this paper, we innovatively propose VSE and DID to bridge the gap between radiological images and natural images, thereby enhancing the interpretability and reliability of the model.

3 Methodology

In this section, we introduce the proposed VDGen model in detail. We elaborate on the working principle of VDGen from two aspects: VSE and DID.

3.1 Overview of Radiology Report Generation

The radiology report generation task can be formalized as follows: Given a radiological image $I \in \mathbb{R}^{C\times H\times W}$, where C denotes the number of channels, and H, Wrepresent the height and width of the X-ray image, respectively, the image I is first encoded by the visual encoder M_v of the radiology report generation model M to obtain its feature representation F_e. Typically, M follows an encoder-decoder architecture, where M_v is implemented using ResNet or Swin Transformer. Subsequently, the textual decoder M_t decodes F_e to generate the diagnostic report $\mathcal{R} \in \mathbb{R}^{L\times D_t} = \{T_1, T_2, T_3, ..., T_L\}$ where T_i denotes a token, L is the report length (number of tokens), and D_t represents the embedding dimension of each token. M_t is commonly a large-scale Transformer model. The full report generation process can be recursively formulated as:

$$p(\mathcal{R} \mid I) = \prod_{i=1}^{L} p(T_{i+1} \mid T_1, \ldots, T_i, I) \quad (1)$$

For the training of model M, given a batch of images $X \in \mathbb{R}^{B\times C\times H\times W} = \{I_1, I_2, \ldots, I_B\}$ and corresponding reports $Y \in \mathbb{R}^{B\times L\times D_t} = \{Y_1, Y_2, \ldots, Y_B\}$,, where $b = B$ denotes the batch size, the predominant training objective in medical report generation is to minimize the cross-entropy loss of model M. Thus, the optimization goal for M can be expressed as:

$$\mathcal{L}_{\text{report}} = \mathcal{L}_{\text{CE}}(\theta) = -\sum_{i=1}^{\mathcal{N}} \log(p_\theta(\mathcal{T}_n^* \mid \mathcal{T}_{1:n-1}^\star)) \quad (2)$$

where $i \in (1, \mathcal{N})$, $\mathcal{N}$ is the vocabulary size. When $i = 1$, $n - 1 = 0$, and the input is $\mathcal{T}_{1:n-1}^\star = X$; otherwise, the input is $\mathcal{T}_{1:n-1}^\star = \{X, T_1, T_2, \ldots, T_{n-1}\}$.

3.2 Vision Self-Equilibration (VSE)

To address the representation challenges caused by imbalanced pixel distributions in radiology report generation, we propose the Vision Self-Equilibration (VSE) module. This module automatically calibrates pixel-level representations of the visual encoder while mitigating distribution bias. During training, given a batch of X-ray images X, VDGen first splits X into patches, appends a $[CLS]$ token, and feeds them into a Swin Transformer (Liu et al., 2021c) for visual encoding. The encoded semantic features are defined as $X^e \in \mathbb{R}^{B\times S\times D} = \{I_1^e, I_2^e, I_3^e, \ldots, I_i^e\}$, where S denotes the sequence length, D_v is the embedding dimension, and B represents the batch size. To ensure alignment in a unified embedding space, we apply Layer Normalization to X^e before VSE processing. For each X-ray image, we compute the

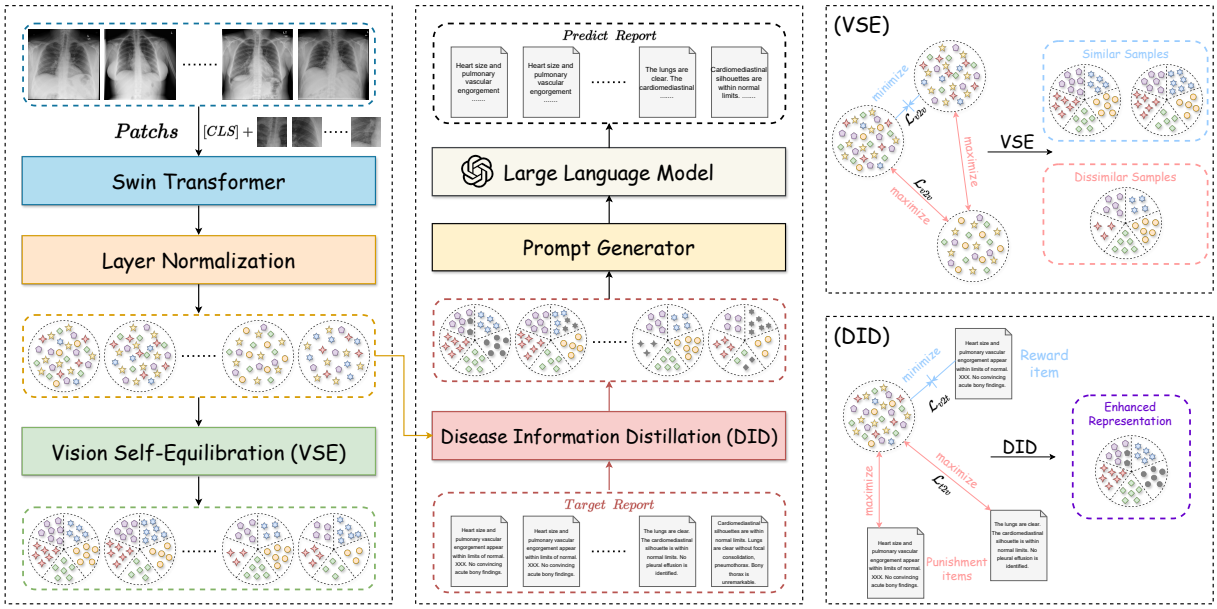

Figure 2: The architecture diagram of the VDGen model. This diagram elaborates on the working principles of VSE and DID, and comprehensively showcases various components of the VDGen model.

mean embedding along the sequence dimension S, yielding $h_i^e = \frac{1}{S}\sum_{s=1}^{S} I_i^{e(s)}$, where $h_i^e \in \mathbb{R}^D$ represents the global visual semantics of the i-th image. To enhance discriminative power, h_i^e is mapped to a higher-dimensional space. The mapped image representation can be expressed as $z_i^v = \mathcal{W}_1^v \cdot h_i^e + \mathcal{B}$, where $\mathcal{W}_1^v$ is learnable matrix, $\mathcal{B}$ is the bias. We expect that X-ray images of similar diseases have similar visual semantic representations, while X-ray images of different diseases have significant differences in semantic representations. In this way, we can alleviate the problem of uneven pixel distribution and achieve a better soft-alignment of visual semantic representations, ultimately achieving a self-equilibrating effect. Our specific approach is as follows: First, we calculate the similarity between different samples within a batch, $S_{ij} = \frac{z_i^{vT} z_j^v}{\tau}$, where τ is the temperature hyperparameter. We set the optimization reward target as itself z_i^v, and other samples z_j^v as penalty terms. Thus, $S_{ii} = -\infty, \forall i \in \{1, 2, 3, \cdots, B\}$. Finally, the objective function that VSE needs to optimize can be expressed as:

$$\mathcal{L}_{vse} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(S_{ii})}{\sum_{j=1}^{B} \exp(S_{ij})} \quad (3)$$

Under the supervision of $\mathcal{L}_{vse}$, VDGen can greatly align the visual semantic representations. Meanwhile, it enhances the ability to represent various visual semantics, including abnormal disease samples and normal samples. This solves the problem of feature degradation in the RRG model caused by the unbalanced pixel distribution in the radiology report generation domain.

3.3 Disease Information Distillation (DID)

While VSE enables discriminative representations of disease symptoms by addressing pixel distribution imbalances in X-ray images, it remains insufficient for characterizing disease presence and extent (e.g., lesion size or scope). Precise identification of abnormalities, such as determining pathological regions in cardiac imaging, critically impacts model performance. To enhance this capability, we propose Disease Information Distillation (DID), which leverages textual embeddings from diagnostic reports (e.g., "minor pleural effusion") as teacher signals to guide VDGen in extracting disease-specific features from visual semantic representations, thereby refining the model's ability to localize and quantify pathologies. For a given target report, after being encoded by a language model, we can obtain $Y = \{Y_1, Y_2, Y_3, \cdots, Y_i\} \in \mathbb{R}^{B \times L \times D_t}$. Before performing the DID operation, we first perform some simple normalization and alignment operations on the visual semantic representations. The operation of visual semantic representations can be expressed as:

$$z_i^{img} = \frac{\frac{1}{S}\sum_{s=1}^{S} I_i^{e(s)}}{\left\| \frac{1}{S}\sum_{s=1}^{S} I_i^{e(s)} \right\|_2} \in \mathbb{R}^D \quad (4)$$

For the features after encoding the target report, we first convert them into probabilities, so $p_i^{text} = softmax(Y_i) \in \mathbb{R}^C$. Then, along the dimension of the report length L, we calculate the average value of the text features to extract the global features. This process can be defined as: $\tilde{z}_i^{text} = \frac{1}{L}\sum_{l=1}^{L} p_i^{text(l)} \in \mathbb{R}^C$. To facilitate the subsequent operations of DID, we need to unify the lengths of the text features and the X-ray visual semantic features. We map $\hat{z}_i^{text}$ into a vector space with the same length as $\hat{z}_i^{img}$ and perform a normalization operation. This process can be defined as follows:

$$z_i^{text} = \frac{W_2\hat{z}_i^{text} + B_2}{\|W_2\hat{z}_i^{text} + B_2\|_2} \in \mathbb{R}^D \tag{5}$$

After obtaining the aligned visual semantics $\hat{z}_i^{img}$ and the text representation $\hat{z}_i^{text}$ of the target report, through the knowledge distillation technique, we can use the text representation $\hat{z}_i^{text}$ of the target report as the output of the teacher model to guide the student model VDGen in the disease information distillation of visual semantics $\hat{z}_i^{img}$. First, we calculate the similarity between the visual representation and the target report representation as $S_{ij} = \frac{z_i^{img^T} z_j^{text}}{\tau_2}$, where τ_2 is the temperature hyperparameter. Next, we calculate the matching probabilities for each X-ray image and the target report features respectively, with the goal of optimizing and aligning the diagonal elements of the similarity matrix. Thus, we can obtain:

$$\begin{aligned} \mathcal{L}_{v2t} &= -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(S_{ii}/\tau)}{\sum_{j=1}^{B}\exp(S_{ij}/\tau)} \\ \mathcal{L}_{t2v} &= -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(S_{ii}/\tau)}{\sum_{j=1}^{B}\exp(S_{ji}/\tau)} \end{aligned} \tag{6}$$

Finally, the objective function that DID needs to optimize can be expressed as:$\mathcal{L}_{DID} = \frac{1}{2}(\mathcal{L}_{v2t} + \mathcal{L}_{t2v})$. Under the constraint of $\mathcal{L}_{DID}$, the representation I_i^e of the X-ray image, guided by the target report representation Y_i, is further enhanced in representing various disease symptoms, thus achieving the extraction of disease information. Specifically, it further enhances the representation of whether a disease occurs and the size of the disease scope. This strengthens the disease perception ability of the downstream LLM during decoding, thereby improving the accuracy of report generation.

3.4 Large Language Model Decoder

After going through VSE and DID, the visual semantic representations are greatly enhanced. We only need to decode them through a Large Language Model (LLM) to generate the final report. Before decoding, we need to construct a prompt based on the visual semantic representations to enhance the generation ability of the large-scale model (Jia et al., 2022). This kind of prompt is more conducive to the reading and understanding of the large-scale model. The prompt constructed in this paper is *"Human: <img>I_i^e</img> Generate a comprehensive and detailed diagnosis report for this chest X-ray image.\nAssistant:"*. Here, I_i^e is the visual representation corrected by VSE and DID. Then this prompt is input into Llama for decoding. The entire decoding process can be represented recursively as:

$$p(\mathcal{R} \mid I_i^e) = \prod_{i=1}^{L} p(T_{i+1} \mid T_1, \ldots, T_i, I_i^e) \tag{7}$$

The report obtained through decoding will calculate the cross-entropy loss ($\mathcal{L}_{CE}$) with the target report (mentioned in Eq. 2), and backpropagation will be carried out to adjust the parameters of the model. Therefore, the loss function that the final model needs to optimize can be expressed as:

$$\mathcal{L} = \mathcal{L}_{report} + \mathcal{L}_{VSE} + \mathcal{L}_{DID} \tag{8}$$

4 Experiment

4.1 Experiment Setting

4.1.1 Dataset

The **MIMIC-CXR** dataset (Johnson et al., 2019), jointly released by the Massachusetts Institute of Technology and Beth Israel Deaconess Medical Center (BIDMC), is a large-scale resource containing approximately 370,000 chest X-ray images from 227,000 patient samples, covering 14 common thoracic diseases with over 200,000 associated reports. Notably, 23% of the samples exhibit abnormalities, reflecting a pronounced class imbalance. In contrast, the **IU-XRay** dataset (Demner-Fushman et al., 2016), curated by Indiana University Hospital, comprises 7,470 chest X-ray images from 3,851 patients paired with 3,955 reports, of which 22.2% (877 samples) are annotated as abnormal. While MIMIC-CXR typically uses single-image-to-single-report pairs for radiology report generation (RRG) training, IU-XRay

Dataset	Model	Year	BLEU1	BLEU2	BLEU3	BLEU4	Meteor	Rouge-L	CIDEr
IU-Xray	SEBTSAT+KG (Zhang et al., 2020)	2020	0.441	0.291	0.203	0.147	–	0.367	0.304
	R2Gen (Chen et al., 2020)	2020	0.470	0.304	0.219	0.165	0.187	0.371	–
	M^2Trans (Nooralahzadeh et al., 2021)	2021	0.402	0.284	0.168	0.143	0.170	0.328	–
	Contrastive (Liu et al., 2021b)	2021	0.492	0.314	0.222	0.169	0.193	0.381	–
	CMCL (Liu et al., 2022)	2022	0.473	0.305	0.217	0.162	0.186	0.378	–
	R2GenCMN† (Chen et al., 2022)	2022	0.475	0.305	0.221	0.171	0.188	0.375	–
	DCL (Li et al., 2023)	2023	–	–	–	0.163	0.193	0.383	0.586
	METransformer (Wang et al., 2023a)	2023	0.483	0.322	0.228	0.172	0.192	0.380	0.435
	KiUT† (Huang et al., 2023)	2023	0.515	0.347	0.241	0.178	0.237	0.406	0.592
	PromptMRG† (Jin et al., 2024)	2024	0.487	0.325	0.234	0.178	0.204	0.401	0.573
	COMG (Gu et al., 2024)	2024	0.482	0.316	0.233	0.184	0.198	0.382	0.529
	R2GenGPT (Wang et al., 2023b)	2024	0.488	0.316	0.228	0.173	0.211	0.377	0.438
	EKAGen (Bu et al., 2024)	2024	**0.517**	**0.351**	**0.258**	**0.191**	0.211	**0.409**	–
	VDGen	2025	0.491	0.325	0.230	0.169	0.195	0.362	0.386
MIMIC-CXR	R2Gen (Chen et al., 2020)	2020	0.353	0.218	0.145	0.103	0.142	0.277	–
	M^2Trans (Nooralahzadeh et al., 2021)	2021	0.332	0.210	0.142	0.101	0.134	0.264	–
	Contrastive (Liu et al., 2021b)	2021	0.350	0.219	0.152	0.109	0.151	0.283	–
	CMCL (Liu et al., 2022)	2022	0.334	0.217	0.140	0.097	0.133	0.281	–
	R2GenCMN† (Chen et al., 2022)	2022	0.348	0.206	0.135	0.094	0.136	0.266	–
	XPRONET (Wang et al., 2022)	2022	0.344	0.215	0.146	0.105	0.138	0.279	–
	DCL (Li et al., 2023)	2023	–	–	–	0.109	0.150	0.284	0.281
	KiUT† (Huang et al., 2023)	2023	0.371	0.233	0.152	0.107	0.146	0.286	0.368
	PromptMRG† (Jin et al., 2024)	2024	0.382	0.227	0.148	0.105	0.157	0.284	**0.375**
	COMG (Gu et al., 2024)	2024	0.346	0.216	0.145	0.104	0.137	0.279	0.352
	R2GenGPT (Wang et al., 2023b)	2024	0.411	0.267	0.186	0.134	0.160	0.297	0.269
	EKAGen (Bu et al., 2024)	2024	0.415	0.254	0.166	0.117	0.154	0.285	–
	VDGen	2025	**0.418**	**0.274**	**0.190**	**0.136**	**0.164**	**0.301**	0.162

Table 1: Main Result on IU-Xray and MIMIC-CXR Dataset. Comparison with SOTA RRG methods on IU X-Ray and MIMIC-CXR benchmark. † indicates the performance evaluated by us. The best results are in **bold**. The second best result is underlined

employs multi-view images (e.g., frontal and lateral) concatenated as input to align with a single report, addressing limited data diversity through multi-perspective fusion. Both datasets highlight the challenge of imbalanced normal-abnormal distributions in RRG model optimization.

4.1.2 Implementation Details

We trained VDGen on a single Nvidia H100 with 80GB of memory. The Swin Transformer used is the base version proposed by Microsoft Corporation. The input image size is set to 224x224. The LLM decoder is based on Llama2-7B (Touvron et al., 2023) that has been aligned through RLHF (Reinforcement Learning from Human Feedback). During the training phase, we used LoRA (Low-Rank Adaptation) (Hu et al., 2022) to fine-tune Llama. We configured the LoRA attention dimension to 16. The alpha hyperparameter for LoRA scaling was also set to 16. Both the training and validation batch sizes were set to 16. The learning rate was set to 1e-4. When VDGen generates reports, the maximum length of the text is controlled to be 100 tokens, and the number of newly added tokens is limited to the range of 80-120. Meanwhile, the repetition penalty coefficient is set to 2.0 and the length penalty coefficient is set to 2.0 to enhance the generation diversity. When generating reports, VDGen adopts a beam search strategy, and the beam size is set to 3.

4.1.3 Evaluation Metrics

In order to compare the performance of VDGen fairly with other methods in the same field, we adopted the mainstream metrics in the RRG field to evaluate the performance of the model. These include commonly used metrics such as BLEU n (Papineni et al., 2002), Meteor (Banerjee and Lavie, 2005), Rouge-L (Chin-Yew, 2004), and CIDEr. BLEU is usually used to measure the similarity of n-grams between the predicted report and the ground truth report. The higher the BLEU score, the closer the generated report is to the ground truth report. Other metrics like Rouge-L are usually used in caption translation tasks. Here, they are used to measure the overall similarity between the report generated by the model and the ground truth report, without calculating from a fine-grained perspective.

By conducting a fair comparison of RRG through these popular metrics, we can observe the performance of VDGen.

4.2 Main Result

To demonstrate the effectiveness of VDGen, we conducted a large number of experiments on the IU-Xray and MIMIC-CXR datasets. The experimental results are shown in Table 1. The best results are presented in bold font, and the second-best results are underlined. The VDGen we proposed did not perform outstandingly on the IU-XRay dataset, differing from the best result by nearly 1.5%. A possible reason is that the IU-Xray dataset is a small-sample dataset with only about 7,000 data points, which is difficult to support the amount of data required for the training of the LLM decoder, resulting in a slight decrease in performance. However, compared with other results on the IU-Xray dataset, VDGen is still comparable. VDGen performed remarkably on the MIMIC-CXR dataset and achieved a new state-of-the-art (SOTA). Specifically, VDGen outperformed the second-best result by 1% on the MIMIC-CXR dataset. This fully demonstrates that VDGen has good performance when there is an adequate amount of training data. The SOTA result on the MIMIC-CXR dataset also establishes the novelty and feasibility of our method. In general, we achieved comparable performance on the IU-XRay dataset and a new SOTA on the MIMIC-CXR dataset.

4.3 Ablation Study

To verify the effectiveness of each module of VDGen, we designed ablation experiments for the VSE, DID, and LoRA modules of VDGen. Table 2 shows that removing VSE led to a 1.26% average performance drop, proving it can address radiological image pixel distribution issues and boost model performance. Removing DID caused a 1.34% average decline, indicating using the ground truth report as a teacher signal for VDGen to extract disease info is effective. VSE is better at capturing fine-grained info, while DID handles both fine - and coarse-grained data. The LoRA ablation experiment showed its fine-tuning benefits the downstream LLM decoder. Overall, VSE and DID are essential for VDGen, and their absence will reduce performance. At the same time, using the LoRA method for training can bring further performance improvements.

4.4 Clinical Efficacy Analysis

To demonstrate the potential of VDGen in clinical applications, we evaluated its performance on the MIMIC-CXR dataset using clinical efficacy metrics (Precision, Recall, and F1-score). The evaluation results are shown in Table 1. Experiments show that VDGen achieved the best Recall and F1-score in the MIMIC-CXR benchmark test, and obtained comparable Precision. Although VDGen's Precision is slightly lower than that of DCL, its Recall and F1-score are 4.5% and 2.9% higher than those of DCL respectively. Therefore, considering comprehensively, VDGen outperforms DCL in clinical applications. Meanwhile, VDGen leads the second-ranked model by 1% in Recall and 1.3% in F1-score, which also highlights the excellent clinical potential of VDGen. Overall, VDGen has shown significant improvements in clinical efficacy metrics, demonstrating good clinical effects and highlighting its potential in clinical applications.

4.5 Case Studies

To further demonstrate the effectiveness and clinical significance of VDGen, we conducted qualitative experiments on the MIMIC - CXR dataset. We compared VDGen with the most popular LLM - based methods on the RRG dataset. The experimental results are shown in Figure 3. We selected two groups of classic cases to analyze the differences between the two methods. The first group mainly consists of normal samples, that is, the images do not contain any diseases. The second group is abnormal samples, that is, the images contain one or more diseases that are difficult to observe with the naked eye. For the first-group samples, the reports generated by VDGen are basically indistinguishable from those described by professional physicians. However, the current mainstream methods fail to diagnose the previous labels, that is, they do not recognize that the patient has had a previous surgery. For the second - group samples, both VDGen and R2GenGPT identified the degenerative changes in the thoracic vertebrae. However, VDGen additionally reported that the patient had a previous surgery, which is crucial information for formulating a treatment plan. R2GenGPT missed this key information. Overall, the reports generated by VDGen are more comprehensive and of greater reference value. This also indirectly reflects the clinical significance of VDGen.

Model	BLEU 1 ↓	BLEU2↓	BLEU3↓	BLEU4↓	Meteor↓	Rouge-L↓	CIDEr↓
Ours	**0.418**	**0.274**	**0.190**	**0.136**	**0.164**	**0.301**	**0.162**
w/o *VSE*	0.395	0.257	0.176	0.128	0.158	0.285	0.158
w/o *DID*	0.402	0.260	0.174	0.124	0.153	0.284	0.154
w/o *LoRA*	0.406	0.263	0.182	0.132	0.161	0.293	0.158

Table 2: Ablation Study on MIMIC-CXR Dataset. We conducted a comprehensive set of ablation studies to analyze individual components of VDGen, where the notation "w/o" (without) indicates the exclusion of specific modules. The experimental results demonstrate that lower metric values observed in configurations without particular modules directly reflect the critical importance of those components, with greater performance degradation quantitatively validating the effectiveness of the removed module.

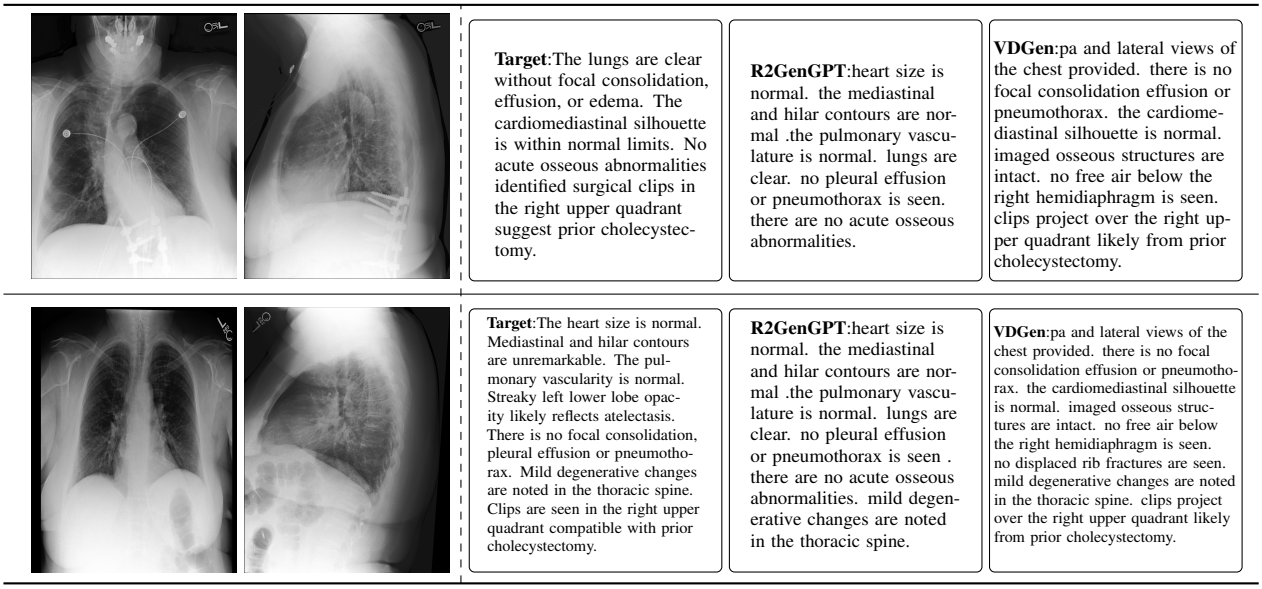

Figure 3: We compared the generated results with the popular LLM-based method. The text in the first rounded rectangle is the ground truth (i.e. target report) described by a professional radiologist. The text in the second rounded rectangle is generated by the popular LLM-based model (i.e. R2GenGPT). The content of the third rounded rectangle is the report generation result of VDGen.

Model	Precision↑	Recall↑	F1-score↑
R2GenCMN	0.334	0.275	0.278
SEBTSAT+KG	0.356	0.297	0.304
METransformer	0.364	0.309	0.311
DCL	**0.471**	0.352	0.373
KiUT	0.371	0.318	0.321
R2GenGPT	0.392	<u>0.387</u>	<u>0.389</u>
COMG	0.424	0.326	0.345
VDGen	<u>0.426</u>	**0.397**	**0.402**

Table 3: Clinical efficacy analysis result. We selected seven models most relevant to VDGen and compared their clinical benefits on the MIMIC-CXR dataset. We used the CE matrix to measure their clinical benefits. The bolded results are the best ones, and the underlined ones are the second best.

5 Conclusion

In this paper, we propose **VDGen**, a unified framework for calibrating visual and diagnostic representations in radiology report generation. VDGen incorporates two complementary modules. Vision Self-Equilibration(VSE) mitigates visual feature degradation through self-supervised contrastive learning, enhancing intra-modal discriminability. Disease Information Distillation(DID) leverages diagnostic reports as teacher signals to guide cross-modal representation learning, improving the model's ability to capture disease-specific semantics. These modules are integrated into an end-to-end generation pipeline with a LoRA-adapted large language model decoder. Experiments on MIMIC-CXR and IU-Xray show that VDGen improves the state-of-the-art by 1.0% on MIMIC-CXR. Ablation studies demonstrate that removing VSE and DID causes performance drops of 1.26% and 1.34%, respectively, highlighting the importance of both components. Overall, VDGen narrows the performance gap between natural and radiological domains and offers a scalable, interpretable solution for clinical report generation.

Limitations

Our study introduces a novel multimodal model, VDGen. This model aims to automatically adjust the model's attention regions via the Visual Self Equilibration(VSE) and extract disease information through the Disease Information Distillation (DID), thereby enhancing the performance of radiology report generation. However, this approach has certain limitations. Firstly, the model is tailored for the highly specialized task of radiology report generation, which may limit its adaptability in broader natural language generation tasks. Additionally, current evaluations are based on specific datasets and may not fully reflect the framework's applicability in diverse scenarios.

References

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.

Shenshen Bu, Taiji Li, Yuedong Yang, and Zhiming Dai. 2024. Instance-level expert knowledge and aggregate discriminative attention for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14194–14204.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2022. Cross-modal memory networks for radiology report generation. *arXiv preprint arXiv:2204.13258*.

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

Lin Chin-Yew. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out, 2004*.

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.

Tiancheng Gu, Dongnan Liu, Zhiyuan Li, and Weidong Cai. 2024. Complex organ mask guided radiology report generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7995–8004.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

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.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. In *European conference on computer vision*, pages 709–727. Springer.

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*, pages 2607–2615.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. 2019. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

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.

Fenglin Liu, Shen Ge, Yuexian Zou, and Xian Wu. 2022. Competence-based multimodal curriculum learning for medical report generation. *arXiv preprint arXiv:2206.14579*.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021a. 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.

Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Yuexian Zou, Ping Zhang, and Xu Sun. 2021b. Contrastive attention for automatic chest x-ray report generation. *arXiv preprint arXiv:2106.06965*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021c. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Farhad Nooralahzadeh, Nicolas Perez Gonzalez, Thomas Frauenfelder, Koji Fujimoto, and Michael Krauthammer. 2021. Progressive transformer-based generation of radiology reports. *arXiv preprint arXiv:2102.09777*.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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

Shijun Wang and Ronald M Summers. 2012. Machine learning and radiology. *Medical image analysis*, 16(5):933–951.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023a. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11558–11567.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023b. R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology*, 1(3):100033.

Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. 2020. When radiology report generation meets knowledge graph. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12910–12917.