InVERGe: Intelligent Visual Encoder for Bridging Modalities in Report Generation

Ankan Deria[1], Komal Kumar[1], Snehashis Chakraborty[1], Dwarikanath Mahapatra[2], Sudipta Roy[1,*]
[1]Artificial Intelligence & Data Science, Jio Institute, Navi Mumbai-410206, India.
[2]Inception Institute of Artificial Intelligence, UAE.
{ankanderia01, dmahapatra}@gmail.com
{Komal2.Kumar, Snehashis1.C, sudipta1.roy}@jioinstitute.edu.in

Abstract

Medical image captioning plays an important role in modern healthcare, improving clinical report generation and aiding radiologists in detecting abnormalities and reducing misdiagnosis. The complex visual and textual data biases make this task more challenging. Recent advancements in transformer-based models have significantly improved the generation of radiology reports from medical images. However, these models require substantial computational resources for training and have been observed to produce unnatural language outputs when trained solely on raw image-text pairs. Our aim is to generate more detailed reports specific to images and to explain the reasoning behind the generated text through image-text alignment. Given the high computational demands of end-to-end model training, we introduce a two-step training methodology with an Intelligent Visual Encoder for Bridging Modalities in Report Generation (InVERGe) model. This model incorporates a lightweight transformer known as the Cross-Modal Query Fusion Layer (CMQFL), which utilizes the output from a frozen encoder to identify the most relevant text-grounded image embedding. This layer bridges the gap between the encoder and decoder, significantly reducing the workload on the decoder and enhancing the alignment between vision and language. Our experimental results, conducted using the MIMIC-CXR, Indiana University chest X-ray images, and CDD-CESM breast images datasets, demonstrate the effectiveness of our approach. Code: https://github.com/labsroy007/InVERGe

1. Introduction

The growing volume of medical imaging data presents a significant challenge to radiologists, who are under pressure to analyze and report results promptly. To address this challenge, automated medical report generation has emerged as a valuable solution, reducing workload, reducing diagnostic errors and streamlining clinic workflow. In this paper, the main objective is to create clear reports about the image's content. This work typically follows an encoder-decoder architecture, with an image encoder extracting features from the image and a large language model (LLM) decoder converting these features into text output. When applying conventional image captioning models directly to medical report generation, performance often suffers. From the previous methods, we also observe that single vanilla Vision Transformer (ViT) [32], initially trained for tasks such as natural image classification or convolutional neural network (CNN) struggles to extract whole information from the medical images due to slight differences between medical images which make it a challenging task. Also in medical imaging [10, 28, 41, 42], capturing pixel-level details such as colour is unnecessary but region, intensity and other details are important. That's why we need a powerful ViT, trained on medical images that will extract high-quality features that can help the decoder generate reports. Therefore, we employ a Self-Supervised Joint-Embedding Predictive Architecture like - I-JEPA [5] to efficiently train our encoder

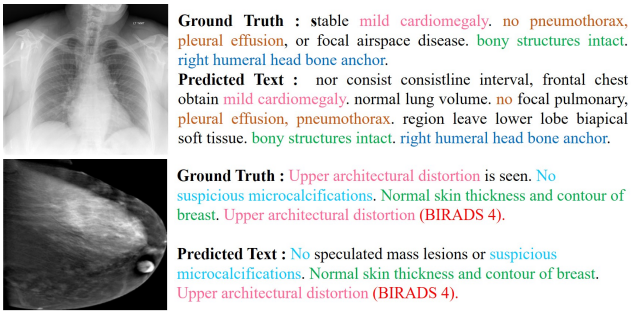

Figure 1. Illustration of two sets of sample reports produced by the InVERGe model, alongside their corresponding ground truth reports for comparison. The matched text is highlighted in the same color, underscoring the alignment between the predicted and actual reports.

in a self-supervised manner [27], obtaining high-level semantic image representations. It learns robust off-the-shelf representations without the use of hand-crafted view augmentation. The fundamental idea behind this architecture is to predict various target sections within the image.
Since the encoder has not seen texts during their unimodal pretraining and the decoder has also not seen the images during their unimodal pretraining, it becomes challenging to merge them and attain effective alignment between vision and language in our task. To bridge this modality gap requires tighter integration of visual and text representations, We incorporate an intermediary CMQFL layer to obtain text-grounded image embeddings which boosts the decoder to generate reports. Since end-to-end model training of a huge encoder-decoder model requires lots of computing power, we propose a this two-step training approach. In the first stage, we train the CMQFL layer to enhance visual representation. In the second stage, we fine-tune the decoder. Our model exhibits better performance than the existing state-of-the-art (SOTA) models. As a result, the entire network gradually enhances its performance, ultimately acquiring the capability to enhance image-based text generation, as demonstrated in Figure 1.

The key contributions of this work are summarized as :

- Our image-grounded text generation pre-training employs a self-supervised image representation task to enhance semantic depth. This involves predicting missing information using pixel reconstruction in an abstract representation space without reliance on external knowledge or transformations. The improvement in report generation is achieved without requiring extra features, annotations, external datasets, or task-specific knowledge.
- We introduce a CMQFL layer to enhance report generation by obtaining a text-grounded image embedding. This layer, trained jointly with the main model, selectively provides crucial image features alongside text, optimizing the process. The iterative approach aims to detect subtle disparities and produce concise, high-quality text-grounded image embeddings and also boost the decoder's ability to generate reports.
- To enhance method explainability, we have used a mechanism that validates image regions corresponding to the report. This involves plotting attention map features from the encoder alongside the CMQFL layer and also plotting the attention maps for individual words.

2. Related work

2.1. Image grounded text generation

This task involves generating descriptive sentences for a given image. However, medical report generation is more challenging than image captioning and reports are usually much longer than captions. Several approaches have been proposed in previous years [16, 35], including approaches using CNNs as image encoders and recurrent neural network (RNN) as a decoder. However, after the improvement of ViTs and transformers' attention mechanism, most of the models used ViT as the encoder and LLM as a decoder. These architectures incorporate the dual-encoder architecture [24, 39], the fusion-encoder architecture [34], the encoder-decoder architecture [14, 51] and more recently, the unified transformer architecture - Bootstrapping Language-Image Pre-training (BLIP) [30] and Beit [49]. Over the years, several pre-training objectives have been put forth, gradually coalescing around a select few proven approaches. These include image-text contrastive learning [29, 39, 55], image-text matching in Align before fuse [29] and Vlmo [8] and (masked) language modelling [29, 49, 57]. BLIP-2 [32] has a Q-Former that uses a frozen visual encoder and then enables zero-shot image-to-text generation through a frozen LLM.

2.2. Radiology Report Generation

Some approaches [11, 31, 43] use a robust CNN-based network that encodes images into visual features and a strong sequential network based on RNN that takes visual features as an initial state to generate image reports. Among these approaches, show attend and tell [53] paper gives the idea of putting some attention to the encoder output to get image grounded caption. Drawing inspiration from this framework, inspired by human intelligence, several papers applied the attention mechanism [35, 56]. The paper [20] employs segmentation models like UNet and TorchXRayVision to extract features from segmented regions. These features are then concatenated to form comprehensive image features. R2Gen [13] utilises a memory-based Transformer architecture, allowing it to remember important information from earlier in the report and uses a special method to include this memory in the report generation process. A recent publication introduced R2GenGPT [52], which is quite similar to the BLIP-2 model but trained on medical image datasets. X-REM [23] method which uses image text similarity loss to get important features from the image. M2Transformer [16] employs a Meshed-Memory Transformer architecture with region encoding to enhance image captioning performance by the integration of a priori knowledge. CvT2DistilGPT2 [36] illustrates that pre-trained models designed for conventional computer vision and natural language tasks can provide valuable support for generating radiology reports. CXR-RePaiR [19] addresses the problem of medically inconsistent information in reports using text-grounded image labels. It labels a report that has the highest cosine similarity in CLIP [39] text embeddings with CLIP image embedding. To calculate this similarity score, two pre-trained singular modality encoders are used. The HReMRG-MR [54] model

used reinforcement learning after the decoder to penalize the incorrectly predicted words. They adopted varying weighted hybrid rewards derived from their search solution and used them as training rewards. The PPKED [34] model consists of three main components: Posterior Knowledge Explorer (PoKE), Prior Knowledge Explorer (PrKE), and Multi-domain Knowledge Distiller (MKD). PoKE identifies explicit abnormal regions in the image using ResNet-152. PrKE examines relevant prior knowledge related to the image for that they use some pre-structure words for both normal and abnormal images. MKD can distil and integrate the subsequent and prior knowledge to create the report.

2.3. Multimodal Task

After the evolution of the LLMs, researchers used this for visual language modelling. LLMs have demonstrated the ability to master novel tasks. They exhibit distinct behaviours and remarkable emergent abilities, like GPT-3's [9] proficiency in few-shot learning, compared to smaller models like BERT [17] and GPT-2 [38]. Recent LLMs such as GPT-3 [9], PaLM [4, 15], LLaMA [44, 45], Vicuna [59], GPT-4 [2] demonstrate enhanced capacity when scaled in terms of model size or data. Several multimodal models have emerged in biomedical applications. Notably, Geneformer [60] focuses on context-specific predictions in low data networks biology application, BiomedGPT [58] combines medical images and literature but requires task-specific fine-tuning using a combination of language model and masked image infilling objectives, while Med-PaLM M [47] tackles multiple biomedical tasks without further fine-tuning for further downstream applications.

2.4. Masked Image Modelling

Masked image modelling (MIM) has made significant progress in parallel with masked language modelling (MLM) [9, 12, 22, 46] tasks in NLP, although initially in a less prominent position. Pioneering efforts, such as context encoder methods and Contrastive Predictive Coding (CPC) [22, 46], predict masked areas and missing pixels in images. Modern vision transformers such as ViT [18], and BEiT [7] have revived this approach with innovative design elements including pixel clustering, average colour prediction, and tokenization via dVAE networks. Recent I-JEPA [5] model introduces an effective method for learning semantic image representations, not pixel-level information, emphasizing simple representations without representational space prediction and scene augmentation for rapid convergence.

3. Method

We give an overall architecture overview of our model in Figure 3. Our model has three main components - an image encoder, a BERT-based CMQFL layer responsible for generating text-grounded image embeddings, and our decoder Vicuna, which draws its inspiration from the LLAMA model.Our model architecture operates in two distinct training stages. Initially, the pretraining stage involves training the encoder once for fine-tuning. In the first stage, we train the CMQFL layer with a frozen image encoder which is trained for different tasks and the frozen decoder performs pre-training using pairs of images and corresponding reports. Lastly, we finetune the LLM according to the output of the CMQFL layer. This comprehensive process culminates in the generation of the final report, bridging the gap between visual content and textual description effectively.

3.1. Encoder

For the visual encoder of our model, we fine-tuned the I-JEPA model using the NIH dataset. For that, we initialized the model's weights from [5]. In total, there are three parts: the context encoder (CE), the predictor (f_ϕ), and the target encoder (TE) of that model.

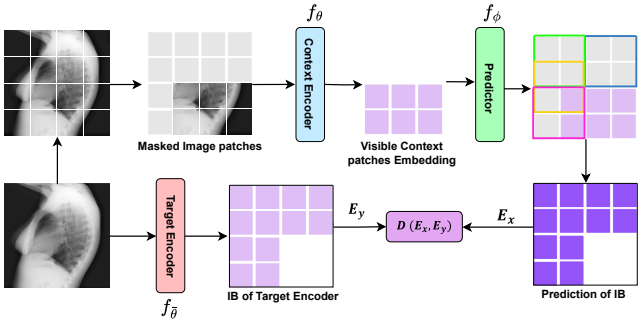

Figure 2. Visual representation of the initial step of training the encoder component of the InVERGe model.

In Figure 2, after masking the patches in block-wise, it uses the unmasked visible block to predict the originating of the interested blocks by the TE ($f_{\bar{\theta}}$). The CE (f_θ) is also a ViT that only processes the visible context patches. The TE ($f_{\bar{\theta}}$) is the same as the CE (f_θ) which is interested in some masked blocks. The predictor (f_ϕ) is a small ViT that takes the CE's output and predicts the representations of the interested blocks of the TE ($f_{\bar{\theta}}$) at a specific location. After the prediction of the predictor, the L_2 loss (i.e., $D(E_x, E_y)$) is computed between the interested block of the Target encoder and the Prediction of those blocks by the Predictor.

In this scenario, we employ gradient methods to fine-tune the parameters of both the predictor f_ϕ and the CE (f_θ). Simultaneously, the parameters of TE ($f_{\bar{\theta}}$) are continuously adjusted, achieved by applying an exponential moving average (EMA) technique to the parameters of the context encoder. Adopting an EMA strategy for target encoders is essential in achieving effective training results for joint embedding architectures (JEA) that incorporate ViT, as discussed in [5]. These updates also ensure the target

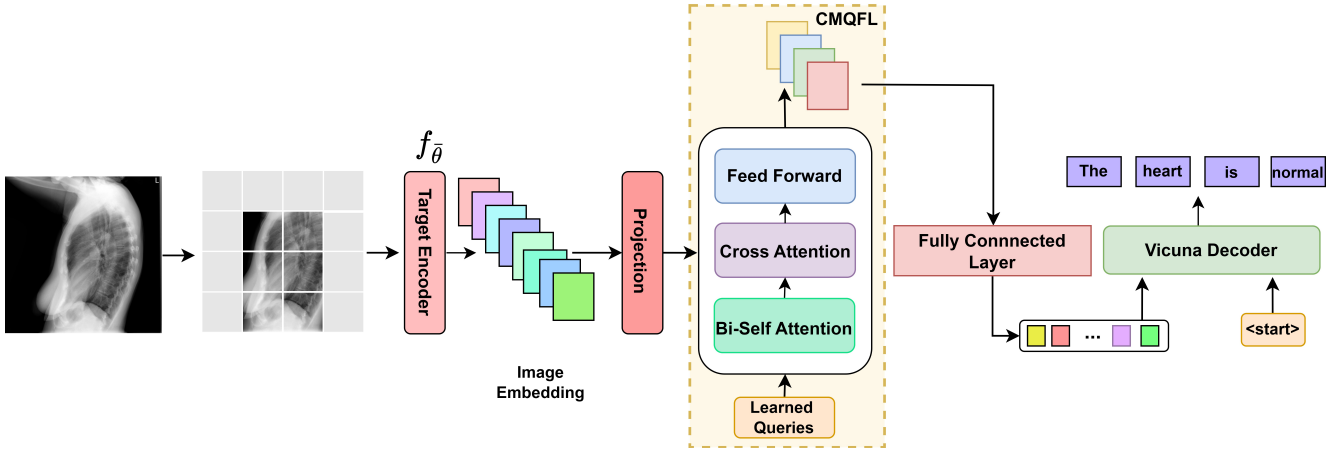

Figure 3. The architecture of the proposed InVERGe approach, includes our encoder trained on self-supervised methods, learnable CMQFL layers, and Vicuna decoder.Specifically, the encoder encodes high-level semantic features and the CMQFL layer uses that information and applies a query that will generate text-grounded image embeddings. Then the decoder uses that embedding to produce a report.

encoder captures high-level pixel information, thereby solidifying its pivotal role as an invaluable component within our model's encoder.

After completing the training of the entire model, we adopt the target encoder as our primary encoder for the proposed **InVERGe** model.

3.2. Text Grounded Image Embedding

We introduce a CMQFL layer using BERT architecture between the encoder and decoder. This CMQFL layer consists of learnable query embeddings, enabling interactions among queries through self-attention and with image features from the frozen encoder via cross-attention layers and producing the most useful text-grounded image embeddings for the decoder to generate the desired report. We initialize this CMQFL layer with pre-trained weights from $BERT_{large}$ with the cross-attention layers which are initialized in a randomised way. In total, this CMQFL layer, contains 32 queries, each with a dimension of 1024, resulting in a much smaller size (32×1024) compared to the frozen image features (e.g., 257×1280 in the encoder). This design, combined with our training three objectives as shown in Figure 4, encourages the query tokens to extract the most relevant visual insights from the image embedding for the text generation.

We pre-train text-grounded image embeddings using three key objectives: multimodal contrastive learning (MCL), masked language modelling (MLM), and enhancing multi-modality matching (MMM) through the implementation of batch negative mining.

Multimodal Contrastive Learning (MCL) : This objective function aims to optimize the alignment of image-text labels, maximizing shared information which is achieved by contrasting the similarity of positive pairs with that of negative pairs. Interactions between query embeddings and image embeddings occur through cross-attention mechanisms in the CMQFL layer, leading to the generation of text-grounded image embeddings (Z) that are more informative. Additionally, we consider the [CLS] token from the text-transformer, denoted as t_{cls}.

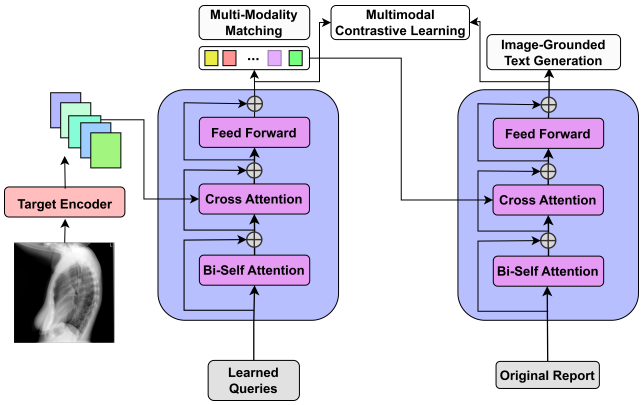

Figure 4. Model architecture of CMQFL layer and decoder at first stage training of CMQFL layer. We jointly train the model using three objective functions that enforce the query tokens to extract the most relevant visual information for the text.

Since Z encompasses multiple output embeddings, each corresponding to a different query, we evaluate the similarity between each query output and t_{cls} by calculating pairwise similarities. The highest similarity is then identified as the image-text similarity. This similarity is computed using the function $L_z(z) \cdot L_t(t_{\text{cls}})$, where L_z and L_t apply linear transformations to the embeddings, converting them into normalized 1024-dimensional representations.

For the image-to-text similarity ($\text{sim}(Q, T)$), we employ the following equation:

$$\text{sim}(Q, T) = \frac{\max(L_z(z) \cdot L_t(t_{\text{cls}}))}{\tau} \quad (1)$$

For the text-to-image similarity ($\text{sim}(T, Q)$), the equation is as follows:

$$\text{sim}(T, Q) = \frac{\max(L_t(t_{\text{cls}}) \cdot L_z(z))}{\tau} \tag{2}$$

where τ is the temperature parameter.

To create the target vector (Y_f), we assigned unique integers to each item within a batch to facilitate the loss calculation. The target vector is specifically designed to match the number of items in the batch, ensuring that each item has a distinct target value. This allows us to calculate losses efficiently and accurately during training. The objective function ($\mathcal{L}_{mcl}$) is defined as the cross-entropy loss (CE) between Y_f and sim :

$$\mathcal{L}_{mcl} = \frac{1}{2}\left[CE(Y_f, \text{sim}(Q, T)) + CE(Y_f, \text{sim}(T, Q))\right] \tag{3}$$

Masked Language Modelling (MLM) : Due to the CMQFL layer's architecture, direct interactions between the fixed image encoder and text tokens are not initially feasible. Therefore, the information necessary for text generation is initially extracted by the queries. Now we masked some text tokens and then passed them via self-attention layers. Queries can interact with each other via self-attention but all text tokens can interact with all the queries and previous text tokens and produce the masked tokens. Therefore, the queries are forced to extract visual features that can produce the masked tokens of the text. For that we use cross-entropy loss to improve its text generation ability, ensuring that it can efficiently and accurately generate text autoregressively.

$$\mathcal{L}_{\text{mlm}} = -\frac{1}{M}\sum_{j=1}^{M} \mathbb{1}_{\text{mask}}(j) \log(p_{ij}) \tag{4}$$

Here, $\mathcal{L}_{\text{mlm}}$ represents the cross-entropy loss for masked language modelling. In this equation, M is the number of masked tokens, and $\mathbb{1}_{\text{mask}}(j)$ is an indicator function that evaluates to 1 if token j is a masked token and 0 otherwise. The purpose of this loss function is to guide the model in minimizing the dissimilarity between its predictions and the actual text for the masked tokens, promoting accurate text generation. These variables are crucial in the computation of the MLM loss, which focuses on the masked tokens and is instrumental in training the model to enhance the accuracy of token predictions in an autoregressive manner. This loss function ensures that the model maximizes the likelihood of the text tokens, particularly the masked ones, during training, a vital aspect of autoregressive text generation.

Multi-Modality Matching (MMM) : This procedure is designed to determine whether an image and text pair is either positively matched or not. To achieve this, We leverage a bi-directional self-attention mask where all queries and texts can appear to each other. As a result, the query embedding, denoted as Q, effectively captures multimodal information. Then we pass each query embedding through a binary classifier to get the two-class probability score which indicates whether the pair is matched or not. Then we take an average of the probability scores of all the queries as the final matching score (p). To enhance the quality of positive-negative pairs, we take negative sample pairs from the batch. Here we utilize the hard negative mining strategy inspired by ALBEF [29] to construct close negative pairs.

$$\mathcal{L}_{\text{mmm}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{C} y_{ij} \log(p_{ij}) \tag{5}$$

In our approach, we employ the Cross-Entropy Loss ($\mathcal{L}_{\text{mmm}}$) to assess the dissimilarity between predicted and ground truth class probabilities. During the training process, we operate with a batch size denoted as B. Here, i represents the sample index, j indicates the class index, and C is set to 2, signifying the two classes for matched and unmatched pairs. The variables y_{ij} and p_{ij} correspond to the ground truth probability of class j in sample i and the predicted probability of class j in sample i, respectively.

The total pre-training objective for training the CMQFL layer of InVERGe consists of a combination of three distinct damage components:

$$\mathcal{L} = \mathcal{L}_{mcl} + \mathcal{L}_{\text{mlm}} + \mathcal{L}_{\text{mmm}} \tag{6}$$

3.3. Decoder

After training the CMQFL layer, it becomes adept at extracting relevant information from image features. To ensure consistency between the output of the CMQFL layer and the input of LLM, we employ a fully-connected layer for linear projection. This projection transforms the output embeddings from the CMQFL layer for the input of the decoder (i.e., Vicuna). These newly projected embeddings are then prepended to the input text embeddings, effectively acting as soft visual cues that condition the decoder on the visual context received by the CMQFL layer. Because the CMQFL layer is pre-trained to capture language-informative visual features, it acts as an effective information filter. This filter selectively delivers the most relevant visual information to the LLM while filtering out less relevant visual details. This not only eases the LLM's burden in learning the alignment between sight and language but also reduces the problem of catastrophic forgetting.

In the fine-tuning process of the decoder, we use the frozen CMQFL layer and exclusively employ the MLM loss given by the model. This architectural configuration ensures close integration of visual and textual information.

4. Experiments

4.1. Dataset Description

For this experiment, we train the encoder using the NIH dataset [50]. For generating the reports, we utilize three datasets: IU-Xray, MIMIC-CXR and CDD-CESM.

MIMIC-CXR: This extensive dataset [25] is commonly used for tasks involving generating reports. It includes 10 folders, comprising a total of 377,110 chest X-ray images and 227,835 corresponding reports. In our research, we specifically focus on the 'p10' folder, which contains 36,337 images.

IU-Xray: The IU-Xray dataset [1] serves as a well-recognized benchmark for evaluating the performance of radiology report generation techniques. It includes 7,470 chest X-ray images, each paired with a corresponding radiology report, totalling 3,955 reports. To prepare the data for our experiments, we initially filtered out entries lacking of findings section.

CDD-CESM: The Breast dataset [26] contains 1003 low-energy images along with corresponding subtracted CESM images, featuring CC and MLO views for both breasts. This dataset is derived from 326 female patients.

In each dataset, we split the dataset into training, validation, and testing sets, with proportions of 75%, 10%, and 15%, respectively.

4.2. Experimental Settings

Metrics: Our evaluation employs a set of well-established metrics, namely BLEU [37], METEOR [6] and ROUGE-L [33]. These metrics are computed using the standard evaluation toolkit. It's worth noting that BLEU and METEOR were initially developed for assessing machine translation quality, while ROUGE-L is specifically designed for evaluating the quality of textual summaries.

Implementation Details: For image feature extraction, we use the target encoder that operates on non-overlapping (14×14) patches, part of the whole trained encoder 3.1. The extracted features consist of 257 patches, each with a dimension of 1280. Then we employ a projection layer to reduce it to 1024 for the CMQFL layer input. The queries are learned from the CMQFL layer, to extract the most useful information from the image features. This vector is then transformed into the shape required for input into the decoder, which generates the final report. We first train the CMQFL layer while freezing all other parts of the model, and then we fine-tune the decoder for better results. During encoder training, we use the AdamW optimizer with a batch size of 4, and the learning rate is linearly increased from 1.0e-4 to 1.0e-3 during the first 10 epochs of pretraining and decays to 1.0e-6 following a cosine scheduler for last 20 epoch. We start with an initial weight decay of 0.04, which progressively increases to a final value of 0.4. Training is conducted at a resolution of (224×224) pixels. Additionally, we introduce momentum with an initial value of 0.996, which is linearly increased to 1.0 during pre-training.

For both tasks, we utilize the AdamW optimizer with mathematical parameters: β_1 set to 0.9, β_2 set to 0.999, and a weight decay of 0.05. Our training process involves 1000 warm-up steps for the IU dataset and 5000 for the MIMIC dataset, with a warm-up learning rate of $1e{-}4$. We employ a cosine learning rate decay strategy, beginning with an initial learning rate (lr) of $1e{-}3$ and gradually decreasing to a final lr of $1e-7$. Throughout both training and evaluation, we consistently use a batch size of 4 due to limited resources.

To train the model, we begin by pre-training the CMQFL layer for 20 epochs. Following this, we proceed to fine-tune the decoder with 25 additional epochs. The entire process takes place with the assistance of an NVIDIA RTX A4000 GPU, which has 16 GB of memory for computational tasks.

4.3. Quantitive Results

We conducted a comprehensive comparison of our model with nine SOTA radiology report generation approaches in Table 1. These methods encompass a wide range of techniques, including both classic and modern approaches such as Show-tell [48], AdaAtt [35], Att2in [40], Up-down [3], R2Gen [13], M2transformer [16], X-REM [23], BLIP-2 [32] and R2GenGPT [52]. The proposed InVERGe model demonstrates superior performance across nearly all metrics. In both the MIMIC-CXR and IU datasets, we achieve superior performance compared to the latest R2GenGPT medthod across all metrics except BLEU-4. The reason behind this improvement is the selection of robust encoder that generate high-level semantic features and the CMQFL layer that generates text-grounded image features.

4.4. Qualitative Results

In Figure 5, we present visualizations of individual words from the report, showcasing how our model not only generates reports but also connects specific words to distinct sections of the image. First, we retrieve the attention map from the last layer of the text decoder cross-attention layer. These attention maps capture the model's focus on different regions of the image corresponding to specific words in the input text. The attention maps are then processed to obtain a single map per word by taking the maximum attention score across all attention heads. Subsequently, these maps are reshaped to match the dimensions of the image.

5. Ablation studies

5.1. Variour Image Encoder

We employ different frozen encoders to extract optimal features from images. Our image encoder choices cover a range of models, including the standard ViT, the Masked

Table 1. Comparison of the proposed InVERGe and other SOTA methods on the MIMIC-CXR, IU and CDD-CESM datasets. A higher value indicates superior performance in all categories. The best performance is highlighted in **bold**.

DataSet	Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
MIMIC-CXR	Show-Tell [48]	0.279	0.159	0.101	0.076	0.121	0.251
	AdaAtt [35]	0.28	0.16	0.105	0.079	0.122	0.254
	Attn2in [40]	0.331	0.213	0.121	0.081	0.131	0.271
	Up-Down [3]	0.318	0.191	0.113	0.071	0.128	0.267
	R2Gen [13]	0.311	0.186	0.112	0.077	0.125	0.265
	M2 Transformer [16]	0.347	0.211	0.122	0.085	0.140	0.269
	X-REM [23]	0.314	0.188	0.112	0.069	0.121	0.266
	BLIP-2 [32]	0.377	0.221	0.125	0.088	0.152	0.274
	R2GenGPT (Deep) [52]	0.392	0.229	0.129	**0.101**	0.159	0.283
	Proposed (InVERGe)	**0.425**	**0.240**	**0.132**	0.100	**0.175**	**0.309**
IU X-Ray	Show-Tell [48]	0.341	0.203	0.140	0.079	0.123	0.321
	AdaAtt [35]	0.434	0.283	0.195	0.126	0.143	0.342
	Attn2in [40]	0.402	0.262	0.188	0.119	0.133	0.338
	Up-Down [3]	0.383	0.248	0.176	0.116	0.129	0.337
	R2Gen [13]	0.421	0.261	0.170	0.121	0.139	0.335
	M2 Transformer [16]	0.456	0.312	0.202	0.151	0.168	0.351
	X-REM [23]	0.426	0.263	0.171	0.119	0.135	0.341
	BLIP-2 [32]	0.476	0.273	0.210	0.168	0.181	0.372
	R2GenGPT(Deep) [52]	0.481	0.301	0.214	**0.169**	0.189	0.375
	Proposed (InVERGe)	**0.499**	**0.324**	**0.226**	0.168	**0.195**	**0.384**
CDD-CESM	Show-Tell [48]	0.284	0.165	0.109	0.079	0.153	0.264
	AdaAtt [35]	0.293	0.169	0.116	0.080	0.180	0.269
	Attn2in [40]	0.340	0.217	0.124	0.081	0.240	0.306
	Up-Down [3]	0.338	0.204	0.123	0.079	0.237	0.310
	R2Gen [13]	0.335	0.199	0.122	0.077	0.213	0.299
	M2 Transformer [16]	0.357	0.221	0.125	0.085	0.256	0.315
	X-REM [23]	0.333	0.197	0.119	0.074	0.210	0.297
	BLIP-2 [32]	0.382	0.235	0.139	0.102	0.301	0.342
	R2GenGPT(Deep) [52]	0.417	0.249	0.165	0.129	0.354	0.377
	Proposed (InVERGe)	**0.453**	**0.267**	**0.185**	**0.134**	**0.391**	**0.430**

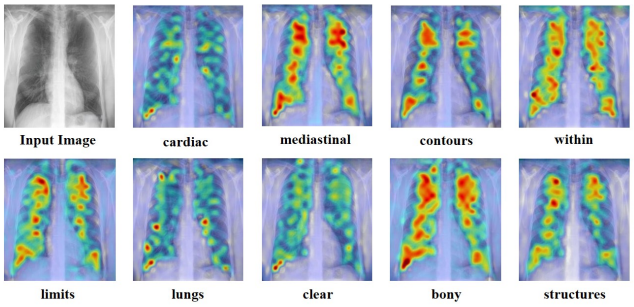

Figure 5. Visual representation of cross-attention maps illustrating word-level explainability.

Autoencoders ViT [21], and the encoder from the IJEPA architecture. These encoders produce image embeddings that are used as input features. The resulting feature sets were evaluated for model performance, as presented in Table 2. The table reveals that employing I-JEPA's encoder significantly enhances the performance of the base model, for instance, improving the BLEU-2 score from $0.161 \rightarrow 0.183$.

5.2. Effect of CMQFL Layer

After selecting an effective encoder, we noticed an improvement in the accuracy of our base model. However, an examination of the attention maps revealed that the model still faced challenges in identifying abnormal regions to generate accurate reports. To solve this problem, we introduced a CMQFL layer capable of detecting abnormal regions. In Figure 6 we discard 80 % low-value attention weights and then plot the top 20% attention, it is clear that the model is actively searching for potentially abnormal regions and using this information to generate high-quality reports.

Table 2. Baseline refers to one normal ViT encoder and decoder only. CPE (Context Pixel encoder) stands for our trained Model's Target Encoder. For this, we use the MIMIC-CXR dataset. Here we only use MLM objective function to check the performance.

Model	BLEU-2	METEOR	Rouge-L
Baseline	0.161	0.124	0.255
MAE + Decoder	0.178	0.060	0.208
CPE + Decoder	0.183	0.117	0.260
CPE + CMQFL + Decoder	**0.227**	**0.163**	**0.290**

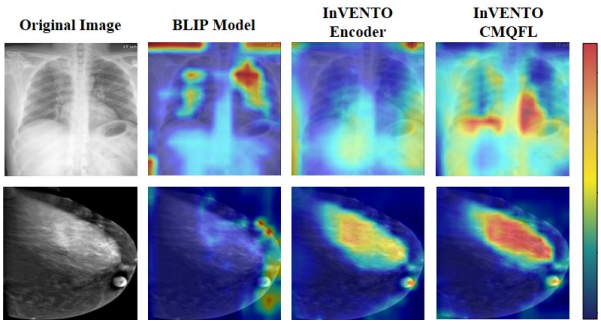

Figure 6. Qualitative results of attention maps generated by BLIP and our InVERGe model's Encoder and CMQFL Layer.

The initial stage of representation learning involves the pre-training of CMQFL layers, which capture visual features relevant to textual content. This process lightens the burden on the LLM when it comes to achieving vision-language alignment. In the absence of a representation learning stage, the CMQFL layer relies entirely on generative learning of the decoder from vision to language to bridge the representation gap of image and text embedding. As shown in Figure 6 and Table 2, it's evident that the impact of the CMQFL layer on the generative learning of the decoder is significant. Without this layer, both types of models exhibit notably lower performance. At the CMQFL layer, key image features are identified as the most relevant features, including cardiomegaly, effusion, atelectasis, consolidation, opacities etc. These features are then directed to the corresponding anomalous regions which clearly highlights the performance of the CMQFL layer in capturing anomalous visual regions. As illustrated in Figure 6 chest X-ray, reveals mild left and right pleural effusion, demonstrating the model's capability to accurately identify and describe specific conditions. These visual results are verified and confirmed by an in-house clinician, ensuring their accuracy and reliability. We can clearly see from Table 2 that there is a significant increase in accuracy after adding the CMQFL layer BLEU-2 score increases from $0.183 \rightarrow 0.227$.

5.3. Objective Functions

In our training approach of the CMQFL layer, we employ a combination of three objective functions to enhance the learning process, as discussed in Section 3.2. These objective functions substantially improve the model's performance. To elaborate, in Table 3, we utilize the Masked Language Model (MLM) loss as the initial training objective for the CMQFL layer. This foundational step allows the model to grasp linguistic and contextual understanding. Subsequently, we augment the training method by introducing two additional objectives: Multimodal Contrastive Learning (MCL) and Multi-Modality Matching (MMM). By adding these objectives, the model's overall capabilities are noticeably enhanced.The addition of MCL focuses on strengthening the alignment between images and text, maximizing their common information. This contrastive approach encourages the model to distinguish between positive image-text pairs and negative pairs, resulting in a stronger understanding of the relationship between visual and text data. At the same time, the introduction of MMM further refines the model's cross-modal capabilities, helping to fine-tune the CMQFL layer by emphasizing alignment between images and text through matching objectives.

Table 3. Evaluation after adding additional objective to train the CMQFL layer. For this, we use MIMIC-CXR dataset.

Model	BLEU-1	BLEU-2	METEOR	Rouge-L
MLM	0.410	0.227	0.163	0.290
MLM+MMM	0.416	0.231	0.166	0.30
MLM+MMM+MCL (InVERGe)	**0.425**	**0.24**	**0.175**	**0.309**

The combined effect of these three objective functions leads to a substantial improvement in the performance of the pre-trained model, enhancing its ability to understand and leverage both text and image data effectively.

6. Conclusion

In this research, we present a novel, high-performing visual-language model that significantly advances the alignment of texts with corresponding visual features. By employing a two-stage training procedure, focusing initially on the CMQFL layer and then fine-tuning the Vicuna decoder, the model demonstrates exceptional performance. The introduction of an advanced encoder for extracting detailed visual features enhances the model's ability to generate reports without requiring additional annotations or external task-specific knowledge. The CMQFL layer, with its three objectives, contributes to creating small yet highly informative image embeddings, promoting a more grounded vision and language representation. This approach not only improves the accuracy of the model within a shorter training period but also surpasses previous SOTA models on publicly available datasets, delivering detailed radiology reports and marking a significant advancement in the field.

In our work, since the decoder component of our model is pre-trained on natural language, we require the integration of an LLM specifically trained on a wide range of medical datasets to increase its effectiveness in generating comprehensive and contextually relevant reports. In a feature we use open-source high-performance medical LLM for getting better results.

References

[1] Indiana university - chest x-rays (xml reports). https://openi.nlm.nih.gov/faq.php. 6

[2] Corporate Ai. Gpt-4 by admin march 27th, 2023 no comments 19 min read gpt-4 technical report. 3

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 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, 2018. 6, 7

[4] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023. 3

[5] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023. 1, 3

[6] Satanjeev Banerjee and Alon Lavie. 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, 2005. 6

[7] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 3

[8] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912, 2022. 2

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3

[10] Snehashis Chakraborty, Komal Kumar, Balakrishna Pailla Reddy, Tanushree Meena, and Sudipta Roy. An explainable ai based clinical assistance model for identifying patients with the onset of sepsis. In *2023 IEEE 24th International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 297–302. IEEE, 2023. 1

[11] Fuhai Chen, Rongrong Ji, Chengpeng Dai, Xuri Ge, Shengchuang Zhang, Xiaojing Ma, and Yue Gao. Factored attention and embedding for unstructured-view topic-related ultrasound report generation. *arXiv preprint arXiv:2203.06458*, 2022. 2

[12] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020. 3

[13] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*, 2020. 2, 6, 7

[14] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021. 2

[15] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 3

[16] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10578–10587, 2020. 2, 6, 7

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[19] Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y Ng, and Pranav Rajpurkar. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Machine Learning for Health*, pages 209–219. PMLR, 2021. 2

[20] Esin Darici Haritaoglu, Aleksandr Timashov, Matthew Tan, and Kathy Yu. Chest x-ray report generation from chest-x ray images. 2023. 2

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 7

[22] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*, pages 4182–4192. PMLR, 2020. 3

[23] Jaehwan Jeong, Katherine Tian, Andrew Li, Sina Hartung, Fardad Behzadi, Juan Calle, David Osayande, Michael Pohlen, Subathra Adithan, and Pranav Rajpurkar. Multimodal image-text matching improves retrieval-based chest x-ray report generation. *arXiv preprint arXiv:2303.17579*, 2023. 2, 6, 7

[24] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2

[25] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr,

a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019. https://physionet.org/content/mimic-cxr/2.0.0. 6

[26] R. Khaled, M. Helal, O. Alfarghaly, O. Mokhtar, A. Elkorany, H. El Kassas, and A. Fahmy. Categorized digital database for low energy and subtracted contrast enhanced spectral mammography images [dataset]. The Cancer Imaging Archive, 2021. 6

[27] Komal Kumar, Snehashis Chakraborty, and Sudipta Roy. Self-supervised diffusion model for anomaly segmentation in medical imaging. In *International Conference on Pattern Recognition and Machine Intelligence*, Kolkata, India, 2023. 2

[28] Komal Kumar, Balakrishna Pailla, Kalyan Tadepalli, and Sudipta Roy. Robust msfm learning network for classification and weakly supervised localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2442–2451, 2023. 1

[29] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2, 5

[30] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 2

[31] Jun Li, Shibo Li, Ying Hu, and Huiren Tao. A self-guided framework for radiology report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 588–598. Springer, 2022. 2

[32] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1, 2, 6, 7

[33] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 6

[34] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 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, 2021. 2, 3

[35] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383, 2017. 2, 6, 7

[36] Aaron Nicolson, Jason Dowling, and Bevan Koopman. Improving chest x-ray report generation by leveraging warm starting. *Artificial Intelligence in Medicine*, 144:102633, 2023. 2

[37] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002. 6

[38] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 3

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[40] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024, 2017. 6, 7

[41] S Roy, T Meena, and SJ Lim. Demystifying supervised learning in healthcare 4.0: a new reality of transforming diagnostic medicine. diagnostics 12 (10): 2549, 2022. 1

[42] Sudipta Roy, Debojyoti Pal, and Tanushree Meena. Explainable artificial intelligence to increase transparency for revolutionizing healthcare ecosystem and the road ahead. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 13(1):4, 2023. 1

[43] Vivek Tiwari, Krutika Bapat, Kushashwa R Shrimali, Saurabh K Singh, Basant Tiwari, Swati Jain, and Hemant Kumar Sharma. Automatic generation of chest x-ray medical imaging reports using lstm-cnn. In *Proceedings of the international conference on data science, machine learning and artificial intelligence*, pages 80–85, 2021. 2

[44] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3

[45] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 3

[46] Trieu H Trinh, Minh-Thang Luong, and Quoc V Le. Selfie: Self-supervised pretraining for image embedding. *arXiv preprint arXiv:1906.02940*, 2019. 3

[47] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *arXiv preprint arXiv:2307.14334*, 2023. 3

[48] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 6, 7

[49] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 2

[50] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 3462–3471, 2017. https://nihcc.app.box.com/v/ChestXray-NIHCC. 6

[51] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 2

[52] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology*, 1(3):100033, 2023. 2, 6, 7

[53] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015. 2

[54] Wenting Xu, Zhenghua Xu, Junyang Chen, Chang Qi, and Thomas Lukasiewicz. Hybrid reinforced medical report generation with m-linear attention and repetition penalty. *arXiv preprint arXiv:2210.13729*, 2022. 2

[55] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 2

[56] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016. 2

[57] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2

[58] Kai Zhang, Jun Yu, Zhiling Yan, Yixin Liu, Eashan Adhikarla, Sunyang Fu, Xun Chen, Chen Chen, Yuyin Zhou, Xiang Li, et al. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv preprint arXiv:2305.17100*, 2023. 3

[59] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023. 3

[60] Yuxuan Zheng and George F Gao. Geneformer: a deep learning model for exploring gene networks. *Science China Life Sciences*, pages 1–3, 2023. 3