

# Unify, Align and Refine: Multi-Level Semantic Alignment for Radiology Report Generation

Yaowei Li<sup>1,2</sup>, Bang Yang<sup>1,2</sup>, Xuxin Cheng<sup>1</sup>, Zhihong Zhu<sup>1</sup>, Hongxiang Li<sup>1</sup>, Yuexian Zou<sup>1†</sup>

<sup>1</sup>School of Electronic and Computer Engineering, Peking University

<sup>2</sup>Peng Cheng Laboratory

{ywl, chengxx, zhihongzhu, lihongxiang}@stu.pku.edu.cn; {yangbang, zouyx}@pku.edu.cn

## Abstract

Automatic radiology report generation has attracted enormous research interest due to its practical value in reducing the workload of radiologists. However, simultaneously establishing global correspondences between the image (e.g., Chest X-ray) and its related report and local alignments between image patches and keywords remains challenging. To this end, we propose an **Unify, Align and then Refine (UAR)** approach to learn multi-level cross-modal alignments and introduce three novel modules: *Latent Space Unifier (LSU)*, *Cross-modal Representation Aligner (CRA)* and *Text-to-Image Refiner (TIR)*. Specifically, LSU unifies multimodal data into discrete tokens, making it flexible to learn common knowledge among modalities with a shared network. The modality-agnostic CRA learns discriminative features via a set of orthonormal basis and a dual-gate mechanism first and then globally aligns visual and textual representations under a triplet contrastive loss. TIR boosts token-level local alignment via calibrating text-to-image attention with a learnable mask. Additionally, we design a two-stage training procedure to make UAR gradually grasp cross-modal alignments at different levels, which imitates radiologists' workflow: writing sentence by sentence first and then checking word by word. Extensive experiments and analyses on IU-Xray and MIMIC-CXR benchmark datasets demonstrate the superiority of our UAR against varied state-of-the-art methods.

## 1. Introduction

Automatic radiology report generation, as a potential intelligent assistant to relieve radiologists from the heavy workload, has attracted a surge of research interests in recent years [19, 30, 18, 24, 7, 35, 61, 67]. Mainstream methods adopt the *de facto* encoder-decoder framework, where a medical image (e.g., chest X-ray) is first encoded as latent representations via convolutional neural networks

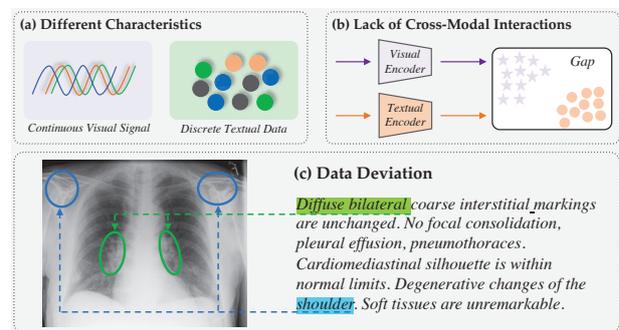


Figure 1. Challenges of modeling cross-modal alignments in radiology report generation. At the encoding stage, it is intractable to globally align visual and textual semantics due to (a) their different characteristics (i.e., *continuous* signals vs. *discrete* data) and (b) the lack of cross-modal interactions. (c) During decoding, capturing the *fine-grained alignment* between keywords and image patches is difficult because of the data deviation problem.

(CNNs) [50, 15], and then further decoded into a radiology report comprised of natural language sentences by recurrent neural networks (RNNs) [16, 10] or fully-attentive networks like Transformer [55]. The key problems in this task are twofold: 1) how to obtain comprehensive information associated with the input medical image and 2) how to accurately establish cross-modal alignments (CMA), e.g., matching generated words with their corresponding image regions.

This paper targets at improving CMA in automatic radiology report generation, which however, is hindered by three factors shown in Figure 1. First of all, *continuous* visual signals and *discrete* text data have the very different characteristics, resulting in semantic inconsistency and modality-independent encoding pipelines in common practices. Secondly, as illustrated in Figure 1 (b), the vision and text modalities are usually encoded via different backbones without cross-modal interactions [7, 67, 43], leading to disparities in the representation space. Thirdly, as shown in Figure 1 (c), there exists the so-called “data deviation” problem,

where important details in the report (*e.g.*, abnormalities) are sparse. As a result, the above three factors pose challenges to learning *global* and *local* CMA between radiographs and related reports. To this end, although there has emerged progresses for improving either global CMA [7, 61, 67, 43] or local CMA [19, 60, 68, 8], how to exploit multi-level CMA to enhance radiology report generation is underexplored.

Considering the above issues, we propose a Unify, Align and Refine (UAR) framework to facilitate multi-level CMA for generating faithful and believable radiology reports. Within UAR, we design a Latent Space Unifier (LSU) to tokenize images into discrete visual tokens with a discrete variational autoencoder (dVAE) [44]. By doing so, LSU unifies vision and text modalities into discrete tokens, making it flexible to design a shared network that seamlessly processes both modalities to learn common knowledge [59, 66]. Next, we introduce a modality-agnostic Cross-modal Representation Aligner (CRA) to learn global CMA. In implementation, CRA not only learns discriminative visual and textual features based on a set of orthonormal basis and a dual-gate mechanism, but also globally aligns both type of features under the supervision of a triplet contrastive loss [48]. What's more, we improve local CMA by integrating Transformer [55] with our proposed Text-to-Image Refiner (TIR). Different from the vanilla attention mechanism in Transformer, TIR additionally contains a learnable mask to re-calibrate text-to-image attention activations. TIR constrains the learnable mask with an auxiliary loss to focus word prediction on useful visual information. Last but not least, we design a two-stage training procedure to make the full model grasp CMA at different levels gradually.

We conduct experiments on two widely-adopted radiology report generation benchmarks, *i.e.*, IU-Xray [12] and MIMIC-CXR [20]. The results demonstrate that our UAR outperforms state-of-the-art methods by a large margin, *e.g.*, with up to 1.9% and 15% absolute improvements in terms of BLEU-4 [42] and CIDEr [56] on IU-Xray, respectively. Ablation study is also carried out to understand the effect of each module of our approach.

In brief, our contributions are three-fold:

- To facilitate multi-level cross-modal alignments, we propose a Unify, Align and Refine framework with three modules: Latent Space Unifier (LSU), Cross-modal Representation Aligner (CRA), and Text-to-Image Refiner (TIR).
- LSU unifies vision and text modalities into discrete tokens, based on which CRA learns to align the global semantics of both modalities whereas TIR encourages the token-level text-to-image alignment.
- Extensive experiments and analyzes on IU-Xray and MIMIC-CXR datasets validate the superiority and effectiveness of our approach, which sets state-of-the-art performance and generates radiology reports accurately.

## 2. Related Work

### 2.1. Visual Captioning

As a member of cross-modal task [28, 9, 26, 29, 27, 49, 65], neural-network-based visual captioning methods [57, 58, 21, 64, 6] are rapidly emerging due to the boom in AI-generated content [69, 46, 53, 25, 54]. Mainstream works use CNNs to extract global [13, 41, 13], grid [63, 39, 45] or region [22, 1, 40] features, and utilize RNNs [58, 13] with varied attention mechanisms [51, 1, 17, 40] to capture cross-modal interactions and generate descriptions in an auto-regressive manner. In recent years, a growing number of methods leverage the potential of the fully-attentive Transformer [55] in capturing long-range dependencies to boost captioning performance [11, 71, 33, 14]. However, these methods are designed to generate a one-sentence description, which is brief and could lack details. Although several works [23, 31] use hierarchical models to generate long paragraphs, they may fail to capture accurate concepts and keywords, which limits application of such models in radiology report generation.

### 2.2. Chest X-ray Report Generation

To adapt the conventional visual captioning framework to radiology report generation, various improvements has been proposed. HRGR [30] designs a retrieval-generation method, which decides to either generate a new sentence or retrieve an existing template for different topics. CMAS [18] utilizes two writers to describe the abnormal and normal regions respectively. PPKED [35] incorporates the posterior-and-prior knowledge for alleviating the data deviation problem. CA [36] compares the current input image with normal images to obtain discriminative abnormal features. Furthermore, for generating more coherent and consistent reports, reinforcement learning [30, 37, 18, 43] is employed to directly optimize desired metrics, graph-based methods [70, 24] are used to construct relationships between diseases, and curriculum learning [34] as a training strategy simulates the writing pattern of “easy first, then difficult”.

Moreover, a portion of works [19] focus on improving cross-modal alignments (CMA) in radiology report generation. CoAtt [19] utilizes semantic information produced by an auxiliary tag prediction task to improve fine-grained CMA. Self-boosting [61] extracts the text-correlated visual feature via coupling the image-text matching branch. TieNet [60] employ a multi-level attention for highlighting the meaningful words and regions. R2GenCMN [7], CMM+RL [43] and JPG [67] treat trainable memory networks as the intermediary of vision and text modalities to enhance global CMA. However, these methods achieve CMA implicitly, which could result in insufficient and inaccurate semantic matching. By contrast, our approach uses explicit constraints to gradually grasp multi-level CMA.

### 3. Method

#### 3.1. Overview

As shown in Figure 2, our proposed UAR is built upon the encoder-decoder framework and contains three novel modules: Latent Space Unifier (LSU), Cross-modal Representation Aligner (CRA), and Text-to-Image Refiner (TIR). Given a radiograph  $I$ , our UAR aims to generate a descriptive report  $R = \{r_1, r_2, \dots, r_T\}$  of  $T$  words automatically. During training, we formulate our approach as follows:

$$\mathbf{E}^{(I)}, \mathbf{E}^{(R)} = \text{LSU}(I, R); \quad (1)$$

$$\mathbf{F}^{(I)}, \mathbf{F}^{(R)} = \text{CRA}(\mathbf{E}^{(I)}, \mathbf{E}^{(R)}); \quad (2)$$

$$h_t = \text{Transformer}_{\text{TIR}}(\mathbf{F}^{(I)}, \mathbf{F}_{<t}^{(R)}); \quad (3)$$

$$p_\theta(r_t | R_{<t}, I) = \text{softmax}(\text{Head}(h_t)). \quad (4)$$

Specifically,  $\text{LSU}(\cdot)$  unifies  $I$  and  $R$  into discrete tokens and produces their embeddings  $\mathbf{E}^{(I)}$  and  $\mathbf{E}^{(R)}$  respectively.  $\text{CRA}(\cdot)$  takes embeddings as the input and outputs discriminative features  $\mathbf{F}^{(I)}$  and  $\mathbf{F}^{(R)}$ , which will be used to contrast and align with each other to enhance global cross-modal alignment (as introduced next). Next,  $\text{Transformer}_{\text{TIR}}(\cdot)$ , a standard Transformer [55] integrated with our proposed TIR to improve fine-grained cross-modal alignment, calculates the hidden state at the  $t$ -th time step ( $h_t$ ) based on the visual features  $\mathbf{F}^{(I)}$  and the preceding textual features  $\mathbf{F}_{0:t-1}^{(R)}$ . Finally, the hidden state  $h_t$  is mapped into the distribution over the vocabulary by  $\text{Head}(\cdot)$  (*i.e.*, a fully-connected layer and a softmax function). The basic training objective of our UAR for language modeling is defined as:

$$\mathcal{L}_{CE} = - \sum_{t=1}^T \log p_\theta(r = r_t | R_{<t}, I) \quad (5)$$

In the following, we will elaborate on our proposed LSU, CRA, and TIR, followed by the introduction of a two-stage training procedure.

#### 3.2. Latent Space Unifier (LSU)

As shown in Figure 2 (a), LSU unifies image and text modalities into discrete tokens and obtain their embeddings.

For extracting visual embeddings  $\mathbf{E}^{(I)}$  of the input radiograph  $I$ , we follow [44] and [4] to exploit a discrete variational autoencoder (dVAE), which involves three components, *i.e.*, a codebook  $\mathcal{V}_I$ , an encoder that maps an image into distributions over the codebook, and a decoder that reconstructs the image from its discrete tokens sampled from the distribution. Please see [44] for more details of dVAE. Here, we only adopt the codebook and the encoder of dVAE and discard its decoder. Concretely, the encoder of dVAE compresses  $I \in \mathbb{R}^{H \times W \times C}$  into  $\mathbf{D} \in \mathbb{R}^{L \times |\mathcal{V}_I|}$ , where  $L = HW/M^2$  denotes the number of visual tokens,  $M$  is

the downsampling factor, and  $|\mathcal{V}_I|$  the size of the codebook. We apply the argmax operation on  $\mathbf{D}$  to obtain  $L$  visual tokens, and then use a trainable look-up matrix  $\mathbf{W}_I \in \mathbb{R}^{|\mathcal{V}_I| \times d}$  to attain  $\mathbf{E}^{(I)} \in \mathbb{R}^{L \times d}$ .

As the radiology report  $R$  is already discrete data, we only need a vocabulary  $\mathcal{V}_R$  (*e.g.*, constructed from the training corpus) and a trainable look-up matrix  $\mathbf{W}_R \in \mathbb{R}^{|\mathcal{V}_R| \times d}$  for extracting text embeddings  $\mathbf{E}^{(R)} \in \mathbb{R}^{T \times d}$ .

It is noteworthy that although the dVAE we use is pre-trained on non-medical images, it produces reasonable reconstruction results for X-ray images (see Figure 6).

#### 3.3. Cross-modal Representation Aligner (CRA)

As shown in Figure 2 (b), CRA is modality-agnostic. It processes visual and textual embeddings with the same set of orthonormal basis and a dual-gate mechanism to model cross-modal interactions, and enhances global semantic alignment under a triplet contrastive loss.

Specifically, we construct a set of orthonormal basis  $\mathbf{B} \in \mathbb{R}^{2,048 \times d}$  via Gram-Schmidt Orthogonalization [47]. Similar to Layer Normalization [2], we adjust  $\mathbf{B}$  as follows:

$$\hat{\mathbf{B}} = \gamma \odot \mathbf{B} + \beta, \quad (6)$$

where  $\odot$  is element-wise multiplication,  $\gamma$  and  $\beta$  are gain and bias parameters of the same dimension as  $\mathbf{B}$ . Next, we use  $\hat{\mathbf{B}}$  and the attention mechanism to process embeddings from different modalities:

$$\begin{aligned} \tilde{\mathbf{F}}^{(*)} &= \text{Attention}(\mathbf{E}^{(*)}, \hat{\mathbf{B}}, \hat{\mathbf{B}}), \\ \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) &= \text{softmax}(\mathbf{A})(\mathbf{V}\mathbf{W}_V), \quad (7) \\ \mathbf{A} &= (\mathbf{Q}\mathbf{W}_Q)(\mathbf{K}\mathbf{W}_K)^\top / \sqrt{d}, \end{aligned}$$

where  $* \in \{I, R\}$ ,  $\mathbf{A}$  denotes attention activations, and all  $\mathbf{W}$  are trainable parameters. In practice, we extend the above attention mechanism into a multi-head version as in [55]. By representing features with the same basis  $\hat{\mathbf{B}}$ , it could benefit feature alignment [62, 52]. Next, inspired by LSTM [16], we define a gate mechanism to control the information flow:

$$G(\mathbf{X}, \mathbf{Y}) = \sigma(\mathbf{X}\mathbf{W}_1 + \mathbf{Y}\mathbf{W}_2), \quad (8)$$

where  $\sigma$  is the sigmoid function,  $\mathbf{W}_1$  and  $\mathbf{W}_2$  are trainable weights. We introduce an input gate  $G_I(\cdot, \cdot)$  and a forget gate  $G_F(\cdot, \cdot)$  to adaptively fuse  $\mathbf{E}^{(*)}$  and  $\tilde{\mathbf{F}}^{(*)}$  to produce the final features  $\mathbf{F}^{(*)}$  as follows:

$$\begin{aligned} \mathbf{F}^{(*)} &= G_I(\mathbf{E}^{(*)}, \tilde{\mathbf{F}}^{(*)}) \odot \tanh(\mathbf{E}^{(*)} + \tilde{\mathbf{F}}^{(*)}) \\ &+ G_F(\mathbf{E}^{(*)}, \tilde{\mathbf{F}}^{(*)}) \odot \mathbf{E}^{(*)} + \tilde{\mathbf{F}}^{(*)}. \quad (9) \end{aligned}$$

**Triplet Contrastive Loss** In order to achieve global cross-modal alignment with explicit supervision signals, we intro-

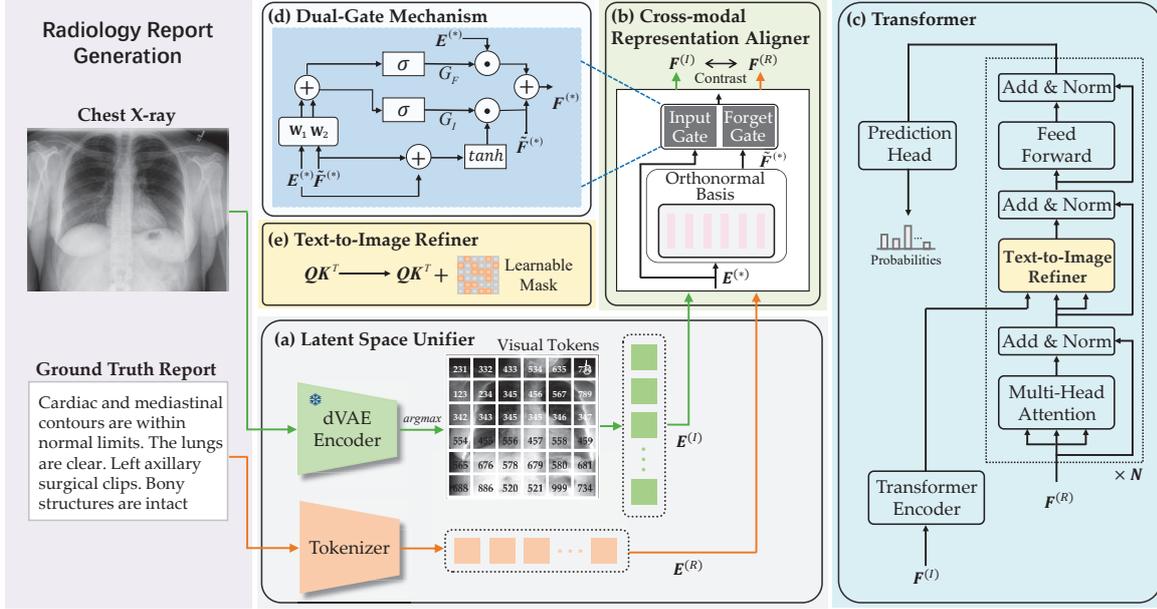


Figure 2. Overview of our UAR framework. In addition to the widely-adopted Transformer shown in (c), it comprises three novel modules to boost multi-level cross-modal alignments: (a) Latent Space Unifier (Section 3.2), (b) Cross-modal Representation Aligner (Section 3.3), and (c, e) Text-to-Image Refiner (Section 3.4). In (d), we illustrate the flow chart of the dual-gate mechanism shown in (b).

duce the following training objective:

$$\mathcal{L}_{Global} = \text{ReLU} \left( \alpha - \langle \mathbf{F}^{(I)}, \mathbf{F}^{(R)} \rangle + \langle \mathbf{F}^{(I)}, \mathbf{F}_{-}^{(R)} \rangle \right) + \text{ReLU} \left( \alpha - \langle \mathbf{F}^{(I)}, \mathbf{F}^{(R)} \rangle + \langle \mathbf{F}_{-}^{(I)}, \mathbf{F}^{(R)} \rangle \right) \quad (10)$$

where  $\langle \cdot, \cdot \rangle$  denotes cosine similarity,  $\alpha$  is the margin value, and  $\mathbf{F}_{-}^{(R)}$  and  $\mathbf{F}_{-}^{(I)}$  indicate hard negative samples that have a high cosine similarity with the anchors  $\mathbf{F}^{(I)}$  and  $\mathbf{F}^{(R)}$ .

### 3.4. Text-to-Image Refiner (TIR)

TIR aims to enhance the correspondence between the radiology report  $R$  and its associated image  $I$  at the token level. This is achieved by affects text-to-image attention patterns with an additional learnable mask, as shown in Figure 2 (c) and (e). Specifically, we modify the definition of  $\mathbf{A}$  (Eq. 7) in the vanilla text-to-image attention of the Transformer decoder as follows:

$$\mathbf{A} = ((\mathbf{Q}\mathbf{W}_Q)(\mathbf{K}\mathbf{W}_K)^\top + k \cdot \sigma(\mathbf{M})) / \sqrt{d} \quad (11)$$

where  $k$  is a large scaling constant (e.g., 1,000) and  $\sigma(\mathbf{M}) \in \mathbb{R}^{T \times L}$  denotes the learnable mask that impacts the activations between  $T$  textual tokens and  $L$  visual tokens. To focus each text token on proper visual information, we constrain the mask with the following objective:

$$\mathcal{L}_{Mask} = \sum_{i=1}^T \sum_{j=1}^L (1 - \sigma(M_{ij})). \quad (12)$$

### 3.5. Two-Stage Training

So far, we have introduced three training objectives of our UAR approach:  $\mathcal{L}_{CE}$  (Eq. 5) for language modeling,  $\mathcal{L}_{Global}$  (Eq. 10) for enhancing global cross-modal alignment, and  $\mathcal{L}_{Mask}$  (Eq. 12) that re-calibrates attention activations to boost local cross-modal alignment. The overall loss function of our approach is formulated as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{CE} + \lambda_2 \mathcal{L}_{global} + \lambda_3 \mathcal{L}_{Mask}. \quad (13)$$

To learn cross-modal alignment incrementally and stably, we split the training process into two stages. In the first stage,  $\{\lambda_1, \lambda_2, \lambda_3\} = \{1, 1, 0\}$ . The model does not include a learnable mask in the text-to-image attention and is forced to learn coarse-grained semantic alignment besides language modeling. In the second stage,  $\{\lambda_1, \lambda_2, \lambda_3\} = \{1, 1, 1\}$ , i.e., the model additionally emphasizes the learning of fine-grained semantic alignment.

## 4. Experiments

### 4.1. Configurations

**Datasets** Our evaluation is performed on two publicly available radiology report generation benchmark datasets, namely IU-Xray [12] and MIMIC-CXR [20].

• **IU-Xray**<sup>1</sup>, established by Indiana University, is a commonly-used dataset that comprises 7,470 X-ray images

<sup>1</sup><https://openi.nlm.nih.gov/faq/>

DATASET	METHOD	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
IU-XRAY [12]	CoATT [19]	0.455	0.288	0.205	0.154	-	0.369	0.277
	HRGR [30]	0.438	0.298	0.208	0.151	-	0.322	0.343
	CMAS-RL [18]	0.464	0.301	0.210	0.154	-	0.362	0.275
	SENTSAT+KG [70]	0.441	0.291	0.203	0.417	-	0.304	0.304
	R2GEN [8]	0.470	0.304	0.219	0.165	-	0.371	-
	CMCL [34]	0.473	0.305	0.217	0.162	0.186	0.378	-
	PPKED [35]	0.483	0.315	0.224	0.168	0.190	0.376	0.351
	R2GENCMN* [7]	0.470	0.304	0.222	0.170	0.191	0.358	0.344
	JPG [67]	0.479	0.319	0.222	0.174	0.193	0.377	-
	CMM+RL [43]	0.494	0.321	0.235	0.181	0.201	0.384	-
UAR (Ours)	0.530	0.365	0.263	0.200	0.218	0.405	0.501	
MIMIC-CXR [20]	ST [58]	0.299	0.184	0.121	0.084	0.124	0.263	-
	ATT2IN [45]	0.325	0.203	0.136	0.096	0.134	0.276	-
	ADAATT [39]	0.299	0.185	0.124	0.088	0.118	0.266	-
	TOPDOWN [1]	0.317	0.195	0.130	0.092	0.128	0.267	-
	R2GEN [8]	0.353	0.218	0.145	0.103	0.142	0.270	-
	CMCL [34]	0.344	0.217	0.140	0.097	0.133	0.281	-
	PPKED [35]	0.360	0.224	0.149	0.106	0.149	0.284	0.237
	R2GENCMN* [7]	0.348	0.206	0.135	0.094	0.136	0.266	0.158
	CMM+RL [43]	0.381	0.232	0.155	0.109	0.151	0.287	-
	UAR (Ours)	0.363	0.229	0.158	0.107	0.157	0.289	0.246

Table 1. Comparison with state-of-the-art methods on IU-Xray and MIMIC-CXR datasets. Optimal and suboptimal performance is highlighted. \*: Our re-implementations with the official code. **Ours UAR achieves competitive if not the best performance on all metrics.**

along with 3,955 corresponding reports. We use the same 70%-10%-20% training-validation-testing splits as previous works [7, 35, 67]. We keep words that appear more than 3 times, resulting a vocabulary of size around 1K.

- **MIMIC-CXR**<sup>2</sup>, provided by Beth Israel Deaconess Medical Center, is a recently released large chest X-ray dataset that contains 473,057 radiographs and 206,563 corresponding reports. We adopt the official splits and retain words that appear more than 10 times. We also preprocess the reports by tokenizing, converting them to lowercase, and removing non-alphabetic tokens. This results in a vocabulary of approximately 4,000 words.

**Metrics** We report four widespread automatic metrics: BLEU [42], METEOR [3], ROUGE-L [32] and CIDEr [56], and compute them with Microsoft COCO Evaluation Server [5]. BLEU [42] and METEOR [3] is originally proposed for machine translation evaluation. ROUGE-L [32] is initially introduced for summarization. CIDEr [56] is proposed for image captioning. Higher is better for all metrics.

**Settings** Our baseline model comprises a pre-trained ResNet101 [15] and a randomly initialized Transformer-Base [55] with 3 layers. For our UAR model, we replace ResNet101 with a pre-trained dVAE, whose details are left to Appendix. Following previous works [30, 35, 67], we use frontal and lateral X-ray images as input on IU-Xray, and only frontal X-ray images on MIMIC-CXR. To save compu-

tion cost, we resize images to  $128 \times 128$  and randomly crop a region of size  $112 \times 112$  at the training phase; we directly resize the image to  $112 \times 112$  at the inference phase. We use the AdamW optimizer [38] to train the model. We employ distinct training stage arrangements and learning rate schedules, along with varying batch sizes, for different datasets. More implementation details are given in Appendix.

## 4.2. Comparison with State-of-the-Art Methods

We compare our UAR model with state-of-the-art (SOTA) methods of different backbones, including CNN-RNN [58, 39, 1, 19] and CNN-Transformer [8, 7, 35, 67]. Besides, we also consider SOTA methods using different technologies, including reinforcement learning [45, 30, 18, 43], knowledge graph [70], and curriculum learning [34]. As shown in Table 1, our approach achieves SOTA performance on the IU-Xray dataset, with up to 1.9% and 15% absolute gains in terms of BLEU-4 and CIDEr metrics. As for the MIMIC-CXR dataset, our UAR obtains competitive results compared to the strongest competitor, *i.e.*, CMM+RL [43], which utilizes reinforcement learning to directly optimize metric scores. Notably, as our approach mainly involves network designing, incorporating existing technologies like reinforcement learning and curriculum learning may bring further improvements. Besides, the superiority of our approach against CoATT [19], R2GenCMN [7], and JPG [67] show the necessity of capturing multi-level cross-modal alignments.

<sup>2</sup><https://physionet.org/content/mimic-cxr/>

SECTION	MODEL	LSU	CRA		TIR	TWO-STAGE TRAINING	DATASET: IU-XRAY [12]						
			Integration	$\mathcal{L}_{global}$			BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
3.2	BASE (a)	✓					0.424	0.268	0.189	0.144	0.171	0.339	0.271
							0.475	0.319	0.233	0.177	0.218	0.399	0.295
3.3	(b)		✓				0.446	0.283	0.196	0.141	0.174	0.360	0.330
				✓			0.426	0.274	0.192	0.143	0.184	0.356	0.409
				✓	✓		0.471	0.302	0.213	0.158	0.187	0.390	0.393
3.4	(e)	✓	✓	✓			0.538	0.365	0.262	0.193	0.218	0.365	0.410
			✓	✓	✓	✓	0.505	0.332	0.244	0.182	0.206	0.401	0.460
3.5	UAR (Ours)	✓	✓	✓	✓	✓	0.530	0.365	0.263	0.200	0.218	0.405	0.501

Table 2. Ablation studies on the proposed Latent Space Unifier (LSU), Cross-modal Representation Aligner (CRA), Text-to-Image Refiner (TIR), and the two-stage training procedure. In CRA, “Integration” means using the orthogonal subspace and the dual-gate mechanism simultaneously. As we can observe, **our UAR model significantly outperforms the base model on all metrics**.

	BL-1	BL-2	BL-3	BL-4	MTOR	RG-L	CIDEr
<b>Ours</b>	0.530	0.365	0.263	0.200	0.218	0.405	0.501
Uniform	0.441	0.284	0.207	0.158	0.179	0.378	0.428
Normal	0.467	0.303	0.210	0.155	0.197	0.371	0.427

Table 3. Effect of using subspaces subject to **orthogonal (default)**, uniform, and normal distributions in the proposed CRA.

	BL-1	BL-2	BL-3	BL-4	MTOR	RG-L	CIDEr
<b>Ours</b>	0.530	0.365	0.263	0.200	0.218	0.405	0.501
No Gate	0.439	0.296	0.208	0.155	0.193	0.373	0.405
Addition	0.487	0.318	0.225	0.167	0.214	0.386	0.376

Table 4. Effect of different fusing mechanisms in the proposed CRA, including **dual-gate (default)**, “no gate”:  $F^{(*)} = \tilde{F}^{(*)}$ , and “addition”:  $F^{(*)} = E^{(*)} + \tilde{F}^{(*)}$ .

## 5. Ablation Study and Analysis

In this section, we delve deeper into each module of our UAR with both quantitative and qualitative results.

### 5.1. Effect of Latent Space Unifier

**Quantitative Analysis** As shown in Table 2 show, model (a) surpasses the base model on all metrics, *e.g.*, 2.3% improvements on BLEU-4. Meanwhile, model (a) is also a strong competitor even compared with the SOTA methods listed in Table 1 on IU-Xray. The performance boost is possibly because LSU reconciles the distributions of visual and textual data and thus prompts the implicit learning of cross-modal alignments. To verify our hypothesis, we calculate the *alignment score*<sup>3</sup> of different models. As shown in Figure 4 (a), we can observe that “Base + LSU” achieves higher alignment score than the base model, *i.e.*, 49% vs. 36%. which proves the potential of LSU in implicitly aligning features of different modalities.

**Qualitative Analysis** We visualize the heatmaps of pairwise cosine similarity among all data samples in Figure 4 (b-d). As we can see in (b), samples similar to the query sample are relatively rare in the base model. Considering that radiology reports contain inherent patterns, high correlations shall be

<sup>3</sup>Alignment score is defined as the portion of radiograph-report pairs whose cosine similarity of features is larger than 0.5 after min-max normalization.

observed frequently. In (c), integrating the base model with LSU improves this situation to some extent via unifying multimodal characteristics. Furthermore, Figure 3 shows retrieval results of different models given the same image as the query. We can see that compared with the base model, “Base + LSU” gives higher cosine similarity between the query image and the ground-truth report, *i.e.*, 0.4440  $\rightarrow$  0.6060. These results further confirm that LSU can benefit global semantic alignments.

### 5.2. Effect of Cross-modal Representation Aligner

**Quantitative Analysis** By comparing models (b,c,d) with the base model in Table 2, we can observe that both the feature integration and the triplet contrastive loss of CRA can bring performance boosts. Specifically, the feature integration encourages the model to generate fluent and faithful reports and thus model (b) achieves higher BLEU and ROUGE-L scores than the base model. Meanwhile, the triplet contrastive loss promotes the generation of diverse sentences that are semantically consistent with ground truth report, resulting in higher CIDEr score of model (c) than the base model. By combining complementary merits of models (b) and (c), model (d) performs better. Moreover, compared with model (d), incorporating LSU and CRA (*i.e.*, model (e)) brings further improvements, which can be attributed to the better cross-modal alignment ability shown in Figure 4

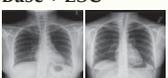
Image Query		Report Retrieval	
 <b>Ground Truth:</b> heart size is normal and the lungs are clear.		Rank 1 0.8020  Rank 2 0.7758  Rank 565 ★0.4440  <b>Ground Truth</b>	Rank 1 0.6683  Rank 2 0.6678  Rank 22 ★0.6060  <b>Ground Truth</b>
Image Retrieval		Base + LSU	
 <b>Base</b>  <b>Base + LSU</b>  <b>Base + LSU + CRA</b>		Rank 1 ★0.7342  <b>Ground Truth</b>	Rank 2 0.7333  <b>Lungs are clear. Heart size normal. No pneumothorax.</b>
		Rank 3 0.7299  <b>The cardiac contours are normal. The lungs are clear. Thoracic spondylosis.</b>	
		<b>Base</b>	<b>Base + LSU + CRA</b>

Figure 3. Image retrieval and report retrieval results of different models. “Base”, “Base + LSU”, and “Base + LSU + CRA” correspond to the base model, model (a) and model (e) in Table 2, respectively. In report retrieval, we highlight **accurate**, **reasonable**, and **wrong** phrases. We also foreground rankings and cosine similarities. As we can observe, the model is more likely to retrieve the ground-truth report by gradually integrating the base model with our proposed LSU and CRA, *i.e.*, **our approach effectively enhances global semantic alignments**.

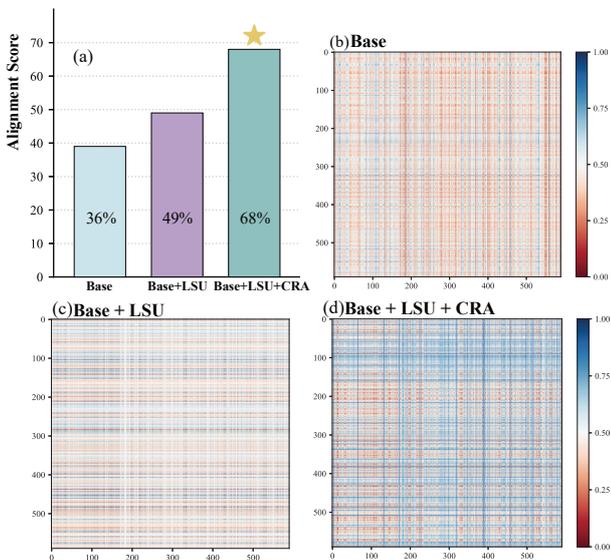


Figure 4. (a): Alignment scores of different models. (b-d): Heatmaps of pairwise cosine similarity among all data samples.

(a), where model (e) (*i.e.*, “Base + LSU + CRA”) gets the highest alignment score (68%).

As the feature integration of CRA is composed of two

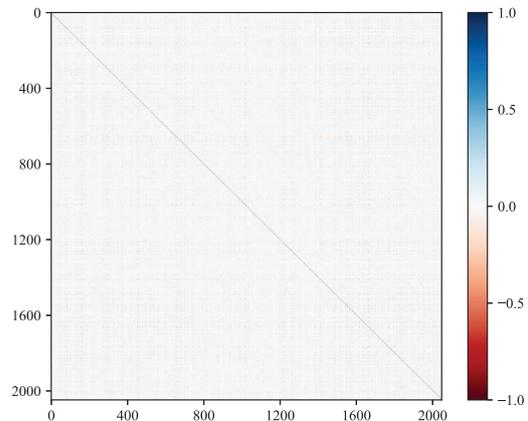


Figure 5. Visualization of Gram matrix for the memory matrix trained in R2GenCMN [7].

parts: the orthogonal subspace<sup>4</sup> and the two-gate mechanism, we investigate different variants that can be substituted with our proposals in Table 3 and Table 4. We can see that our proposed orthogonal subspace outperforms the uniform and normal ones, and the dual-gate mechanism is superior in feature fusion than all other variants.

**Qualitative Analysis** Here, we first probe into the SOTA

<sup>4</sup>Since we use the scaled mechanism like Layer Normalization [2] to adjust the basis, the orthonormal subspace evolved into orthogonal.

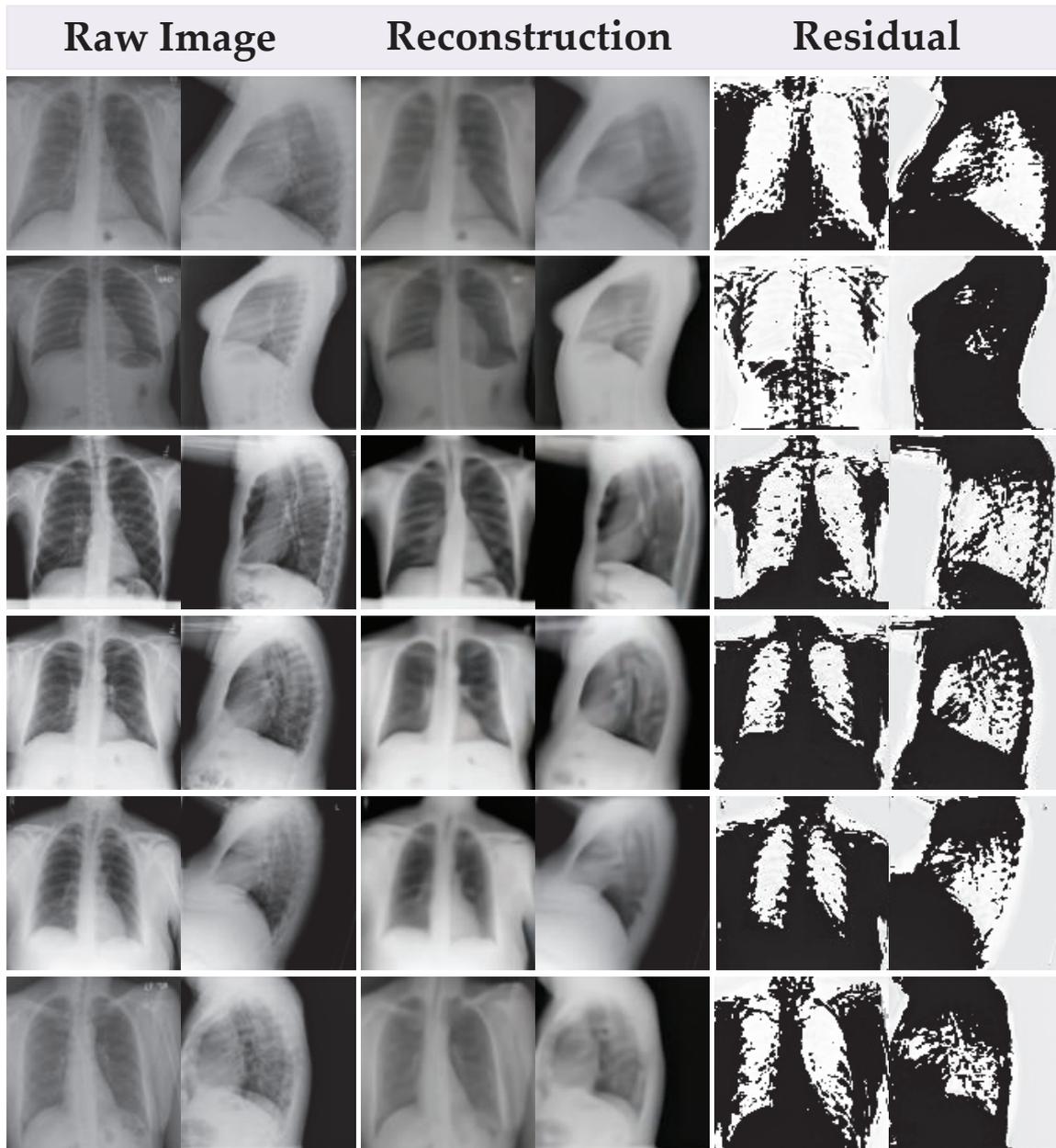


Figure 6. Chest X-ray reconstructions of pre-trained dVAE on IU-Xray dataset [12]. From left to right, we visualize the raw radiograph, the reconstruction result and the residual image between the two in turn.

R2GenCMN model [7] that enhance global semantic alignments with a memory matrix. We visualize the Gram matrix of the pre-trained memory matrix in Figure 5. Surprisingly, almost all values in Gram matrix equal to 0, demonstrating that it is approximately an orthogonal subspace. This visualization result justifies our subspace design in CRA.

Likewise, we show the similarity heatmap of “Base + LSU + CRA” in Figure 4 (d), whose pattern is quite different

to that of (b) and (c). In Figure 3, we can observe that “Base + LSU + CRA” performs the best in retrieval. On the one hand, retrieved reports have high similarity to the query image, and they are semantically consistent with the ground-truth report, *e.g.*, *lung are clear* and *the cardiac contours are normal*. On the other hand, the ground-truth report can be retrieved precisely, *i.e.*, ranking first with a similarity of 0.7342.

In brief, the above quantitative and qualitative analyses

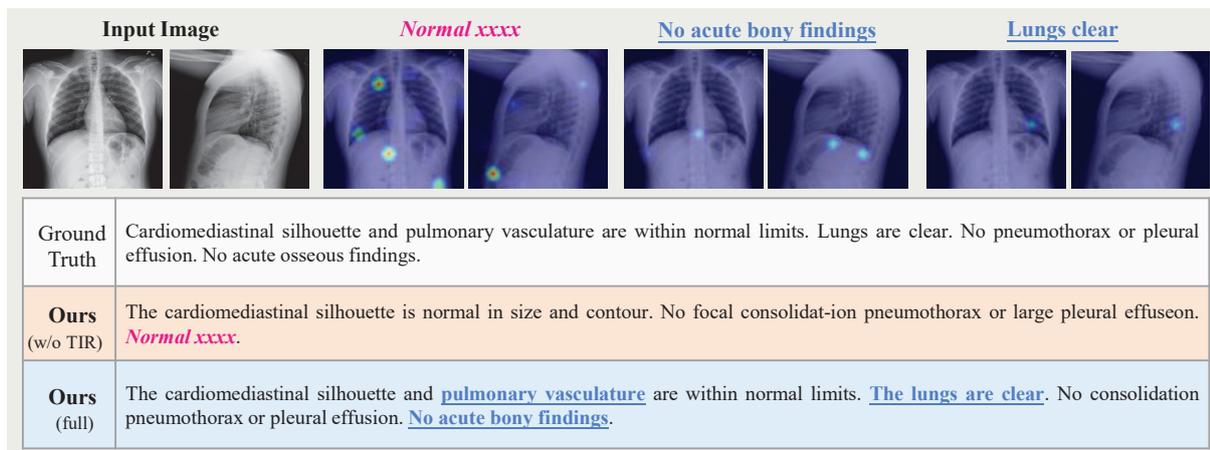


Figure 7. An example of radiology reports generated by our model without or with the proposed Text-to-Image Refiner (TIR). We highlight *incomplete* and *detailed* descriptions in the report. We also show the text-to-image attention heatmaps of specific descriptions at the top. **Our full model learns accurate correspondences between image regions and keywords.**

verify that our proposed LSU and CRA can effectively enhances global semantic alignments.

### 5.3. Effect of Text-to-Image Refiner

**Quantitative Analysis** By comparing model (e), (f), and our UAR model in Table 2, we have the following observations. 1) TIR can effectively improves the CIDEr metric, *e.g.*, 5% absolute improvement for model (f) vs. model (e), showing that TIR is beneficial for generating accurate keywords in the radiology report. Moreover, excluding two-stage training leads to obvious performance degradation, *i.e.*, our UAR model vs. model (f). This indicates the instability of multi-objective optimization and proves the effectiveness of our two-stage training procedure.

**Qualitative Analysis** In Figure 7, we illustrate the effect of TIR on radiology report generation. As we can see, the model without TIR outputs an useless description “Normal xxxx”, where “xxxx” is anonymized information used to protect patient privacy and results in an incomplete sentence. By contrast, our TIR remedies this content omission and interprets the visual content in detail. Additionally, text-to-image attention heatmaps demonstrate that our TIR is capable of focusing word prediction on proper visual information.

In all, TIR can capture accurate fine-grained correspondences between image regions and keywords. Our full model can generate more informative and meaningful radiology reports by enhancing multi-level semantic alignments.

### 5.4. Case Study of dVAE reconstructions

As shown in Figure 6, we present several reconstructions for chest X-ray images. We observe that pre-trained dVAE can generate reasonable and consistent reconstruction results,

which demonstrates that the extracted visual tokens remain high-level semantic information, *i.e.*, different tokens may represent different organs or regions. In other words, the unification of visual and textual characteristics facilitates the design of a shared network to flexibly achieve cross-modal alignment. Furthermore, most of the low-frequency structure are retained and the high-frequency information is ignored in reconstruction X-ray images. Despite the low-frequency structure already contains enough semantic information, the high-frequency information may be conducive to the judgement of specific abnormalities, *e.g.*, bone structure, which will remain in future research.

## 6. Conclusion

In this paper, we make the first attempt to align visual and textual semantics at different levels with explicit constraints in automatic radiology report generation. To this end, we propose the UAR approach to unify multimodal data into discrete tokens, align visual and textual representations globally, and refine fine-grained text-to-image correspondences. Extensive experiments and analyses on two benchmark datasets, *i.e.*, IU-Xray and MIMIC-CXR, demonstrate that our approach can generate more informative and meaningful radiology reports by boosting multi-level cross-modal alignments and thus achieves state-of-the-art performance.

## Acknowledgments

This paper was partially supported by NSFC (No: 62176008) and Shenzhen Science & Technology Research Program (No: GXWD20201231165807007-20200814115301001).

## References

- [1] 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. 2, 5
- [2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 3, 7
- [3] Satantjeev 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. 5
- [4] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*, 2023. 3
- [5] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 5
- [6] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3193–3203, 2021. 2
- [7] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 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, 2021. 1, 2, 5, 7, 8
- [8] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449, 2020. 2, 5
- [9] Xuxin Cheng, Zhihong Zhu, Hongxiang Li, Yaowei Li, and Yuexian Zou. Ssvm: Saliency-based self-training for video-music retrieval. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 2
- [10] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NeurIPS Workshop on Deep Learning*, 2014. 1
- [11] 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
- [12] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016. 2, 4, 5, 6, 8
- [13] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015. 2
- [14] Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lin Liang, Zhe Gan, Lijuan Wang, Yezhou Yang, and Zicheng Liu. Injecting semantic concepts into end-to-end image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18009–18019, 2022. 2
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 5
- [16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 1, 3
- [17] Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. Recurrent fusion network for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 499–515, 2018. 2
- [18] Baoyu Jing, Zeya Wang, and Eric Xing. Show, describe and conclude: On exploiting the structure information of chest x-ray reports. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6570–6580, 2019. 1, 2, 5
- [19] Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2577–2586, 2018. 1, 2, 5
- [20] 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. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019. 2, 4, 5
- [21] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Denscap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4565–4574, 2016. 2
- [22] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 2
- [23] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–325, 2017. 2
- [24] Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6666–6673, 2019. 1, 2

- [25] Dongxu Li, Junnan Li, and Steven CH Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *arXiv preprint arXiv:2305.14720*, 2023. 2
- [26] Hongxiang Li, Meng Cao, Xuxin Cheng, Yaowei Li, Zhihong Zhu, and Yuexian Zou. G2I: Semantically aligned and uniform video grounding via geodesic and game theory. *arXiv preprint arXiv:2307.14277*, 2023. 2
- [27] Hongxiang Li, Meng Cao, Xuxin Cheng, Zhihong Zhu, Yaowei Li, and Yuexian Zou. Generating templated caption for video grounding. *arXiv preprint arXiv:2301.05997*, 2023. 2
- [28] 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. 2
- [29] Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. Decap: Decoding clip latents for zero-shot captioning via text-only training. *arXiv preprint arXiv:2303.03032*, 2023. 2
- [30] Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Hybrid retrieval-generation reinforced agent for medical image report generation. *Advances in neural information processing systems*, 31, 2018. 1, 2, 5
- [31] Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan, and Eric P Xing. Recurrent topic-transition gan for visual paragraph generation. In *Proceedings of the IEEE international conference on computer vision*, pages 3362–3371, 2017. 2
- [32] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 5
- [33] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. Swinbert: End-to-end transformers with sparse attention for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17949–17958, 2022. 2
- [34] Fenglin Liu, Shen Ge, and Xian Wu. Competence-based multimodal curriculum learning for medical 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 3001–3012, 2021. 2, 5
- [35] 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. 1, 2, 5
- [36] Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun. Contrastive attention for automatic chest x-ray report generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 269–280, 2021. 2
- [37] Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation. In *Machine Learning for Healthcare Conference*, pages 249–269. PMLR, 2019. 2
- [38] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5
- [39] 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, 5
- [40] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7219–7228, 2018. 2
- [41] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 2
- [42] 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. 2, 5
- [43] Han Qin and Yan Song. Reinforced cross-modal alignment for radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 448–458, 2022. 1, 2, 5
- [44] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2, 3
- [45] 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. 2, 5
- [46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2
- [47] Erhard Schmidt. Zur theorie der linearen und nichtlinearen integralgleichungen. *Mathematische Annalen*, 63(4):433–476, 1907. 3
- [48] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 2
- [49] Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14974–14983, 2023. 2
- [50] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, 2015. 1

- [51] Jingkuan Song, Lianli Gao, Zhao Guo, Wu Liu, Dongxiang Zhang, and Heng Tao Shen. Hierarchical lstm with adjusted temporal attention for video captioning. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2737–2743, 2017. 2
- [52] Gilbert Strang. The discrete cosine transform. *SIAM review*, 41(1):135–147, 1999. 3
- [53] 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. 2
- [54] 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. 2
- [55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1, 2, 3, 5
- [56] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 2, 5
- [57] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. Translating videos to natural language using deep recurrent neural networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1494–1504, 2015. 2
- [58] 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. 2, 5
- [59] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, 2022. 2
- [60] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9049–9058, 2018. 2
- [61] Zhanyu Wang, Luping Zhou, Lei Wang, and Xiu Li. A self-boosting framework for automated radiographic report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2433–2442, 2021. 1, 2
- [62] Yiyang Wu and William Y Zou. Orthogonal frequency division multiplexing: A multi-carrier modulation scheme. *IEEE Transactions on Consumer Electronics*, 41(3):392–399, 1995. 3
- [63] 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
- [64] Bang Yang, Yuexian Zou, Fenglin Liu, and Can Zhang. Non-autoregressive coarse-to-fine video captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3119–3127, 2021. 2
- [65] Shuzhou Yang, Moxuan Ding, Yanmin Wu, Zihan Li, and Jian Zhang. Implicit neural representation for cooperative low-light image enhancement. *arXiv preprint arXiv:2303.11722*, 2023. 2
- [66] Haoxuan You, Luwei Zhou, Bin Xiao, Noel Codella, Yu Cheng, Ruochen Xu, Shih-Fu Chang, and Lu Yuan. Learning visual representation from modality-shared contrastive language-image pre-training. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*, pages 69–87. Springer, 2022. 2
- [67] Jingyi You, Dongyuan Li, Manabu Okumura, and Kenji Suzuki. Jpg-jointly learn to align: Automated disease prediction and radiology report generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5989–6001, 2022. 1, 2, 5
- [68] Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, pages 721–729. Springer, 2019. 2
- [69] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations*, 2022. 2
- [70] Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. When radiology report generation meets knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12910–12917, 2020. 2, 5
- [71] Luwei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13041–13049, 2020. 2