Act Like a Radiologist: Radiology Report Generation across Anatomical Regions

Qi Chen[1*], Yutong Xie[1*], Biao Wu[2], Xiaomin Chen[3], James Ang[4]
Minh-Son To[1,4,5], Xiaojun Chang[2], and Qi Wu[1†]

[1] Australian Institute for Machine Learning, University of Adelaide
[2] University of Technology Sydney, [3] South China University of Technology
[4] Royal Adelaide Hospital, [5] Flinders University

This part provides more discussions and experimental details to supplement the main submission. We organise the supplementary into the following sections.

- In Section A, we provide more discussions, including the effect of different feature extractors (Section A.1), the impact of hyper-parameter λ (Section A.2), and whether feeding image tags into the model would cause information leakage (Section A.3).
- In Section B, we show some examples on our private datasets.
- In Section C, we depict details of our general knowledge base.
- In Section D, we provide more implementation details.
- In Section E, we show more quantitative results.

A More Discussions

In this part, we provide more discussions, including the effect of different feature extractors in Section A.1, the impact of hyper-parameter λ in Section A.2, and whether feeding image tags into the model would cause information leakage in Section A.3.

A.1 Effect of Feature Extractors

In our X-RGen framework, the tokeniser for knowledge word embeddings is initialised using MedClip [15]. It, trained extensively on a vast corpus of clinical text, offers a robust choice for such feature extraction. Meanwhile, within the cross-region analysis phase, the text encoder is initialised with MedClip as well. To empirically assess the contributions of the two medical-specific pre-training models, we modified our X-RGen, substituting these two pre-training feature extractors with a generic BERT pre-training [5]. For a fair comparison, we set all the batch sizes to 96. As shown in Table 1a, when initialised with this general-domain BERT, our X-RGen model experiences a performance degradation of approximately 22% in CIDEr (declining from 0.324 to 0.302) and a 4% decrease in B4 (from 0.104 to 0.100). The results demonstrate the significance of medical-specific initialisation. Nevertheless, even without it, our X-RGen significantly outperforms the base model. This suggests that the performance gains of the X-RGen framework are attributed not only to medical-aware initialisation but also to the cross-region analysis and medical interpretation phases we introduced.

	B4	CIDEr
Base	0.095	0.276
X-RGen with BERT init.	0.100	0.302
X-RGen	0.104	0.327

(a) Effect of different feature extractors

λ	B4	CIDEr
0.5	0.108	0.317
1.0	0.110	0.330
1.5	0.101	0.272

(b) Impact of λ

	B4	CIDEr
R2Gen [3]	0.096	0.280
R2Gen [3] with tags	0.097	0.284

(c) Information leakage from tags

Table 1: We test (a) the effect of different feature extractors. "X-RGen with BERT init." means we initialise all text encoders in X-RGen with a generic BERT pre-training model; (b) Impact of hyper-parameter λ in Eq. (7); (c) whether feeding image tags $c(\cdot)$ into the model would cause information leakage. All results are on IU-Xray (chest).

A.2 Impact of Hyper-parameter λ in Eq. (7)

As shown in Table 1b, when the value of λ is small, such as $\lambda = 0.5$, the performance of our X-RGen is suboptimal. The reason lies in the insufficient enhancement of the recognition across various anatomical regions and the semantic alignment between different modalities (*i.e.*, images and reports). As we increase the value of λ, the performance of X-RGen reaches its peak at $\lambda = 1.0$. However, beyond that point, the performance starts to degrade. To balance these two terms, we set the weighting parameter λ to a value of 1.0 in all our experiments.

A.3 Risk of Information Leakage from Tag $c(x)$

To examine the absence of information leakage, we feed the tag $c(x)$ of each input image x into the existing well-known R2Gen method and observe the impact of the performance. As shown in Table 1c, the inclusion of input tags does not lead to much-improved performance for R2Gen [3] (*i.e.*, B4: 0.096 $\rightarrow$ 0.097; CIDEr: 0.280 $\rightarrow$ 0.284). It implies that the presence of input tags $c(\cdot)$ does not result in information leakage. On the contrary, they can be considered as medical-related priors, but need a well-designed approach (*e.g.*, the medical interpretation phase in our X-RGen) to unleash their inherent potential.

B Examples on Private Datasets

In experiments, we construct a merged dataset that contains paired data w.r.t. six anatomical regions, including chest, abdomen, knee, hip, wrist and shoulder. Due to the lack of existing datasets, we collect private image-report pairs on all six anatomical regions. Anonymous Human Research Ethics Committee provides ethics approval for private data used in this study. For each region, we have $3,000$ patients and the ratio of train/val/test is 70%/15%/15%. Notably, for a fair comparison with previous works, we use chest pairs on IU-Xray [4], a publicly recognised dataset, rather than our private ones. It consists of $3,955$ fully de-identified radiology reports, each paired with frontal and/or lateral chest X-ray images. Following [3, 8], we remove cases that contain only a single image and then divide the dataset into train, validation, and test sets with 2069/296/590 pairs, respectively. Here, we provide some samples on the other five private datasets in Figure 1.

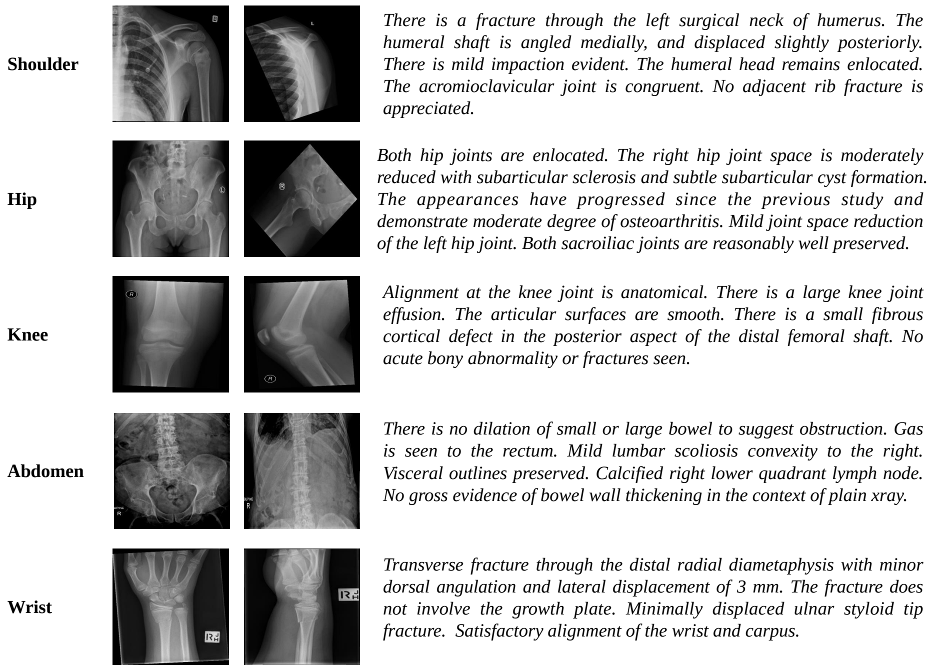

Fig. 1: Examples on the private datasets. Each example contains a frontal image (first column) and another image (second column) with the corresponding radiology report.

C Details of Knowledge Base

Here, we used different colours to highlight shared topics across the six anatomical regions. The results show that there are many topics commonly used, even across different regions. This finding indicates that our knowledge set has a relatively general scope. Topics on our general knowledge set $\mathfrak{S}$ include:

{abdomen, acetabular, acromioclavicular, acute, airspace disease, anatomical, angulation, atelectasis, bilateral, bone, bony, bowel, calcification, calcinosis, cardiomediastinal, cardiomegaly, carpal, cast, change, changes, cicatrix, clavicle, colon, compartment, complication, consolidation, contours, cuff, degenerative, dislocation, displacement, distal, dorsal, edema, effusion, emphysema, enlocated, evidence, faecal, femoral, femur, fracture, fractures, gas, glenohumeral, glenoid, head, healing, hernia, hip, humeral, humerus, hypoinflation, identified, inferior, intact, interval, joint, knee, lateral, lesion, limits, loading, loops, lucency, lumbar, lung, material, medical device, mild, moderate, nonspecific, normal, obstruction, opacity, other, patella, patellar, pelvic, pelvis, periprosthetic, plate, pleural, pneumonia, pneumothorax, projection, prosthesis, proximal, pubic, quadrant, radial, radio-carpal, radius, rectum, replacement, ring, sacroiliac, satisfactory, scaphoid, sclerosis, scoliosis, shoulder, situ, soft, space, stomach, styloid, subacromial, sub-

diaphragmatic, supine, suprapatellar, surgical, swelling, symphysis, thickening, tissue, tissues, transverse, tuberosity, ulnar, visualised, wrist}

Topics on each anatomical region namely $\mathcal{G}$ and we highlight the overlapped topics across different body parts in various colours:

- Chest = {*airspace disease, atelectasis, calcinosis, cardiomegaly, cicatrix, edema, effusion, emphysema, fractures, hernia, hypoinflation, lesion, medical device, normal, opacity, other, pneumonia, pneumothorax, scoliosis, thickening*}
- Abdomen = {*abdomen, bowel, cardiomediastinal, colon, consolidation, contours, degenerative, evidence, faecal, gas, limits, loading, loops, lumbar, lung, material, moderate, nonspecific, obstruction, pleural, projection, quadrant, rectum, stomach, subdiaphragmatic, supine, surgical, tissue*}
- Knee = {*acute, alignment, anatomical, changes, compartment, complication, degenerative, dislocation, effusion, evidence, femoral, fracture, gas, joint, knee, lateral, lucency, mild, moderate, patella, patellar, prosthesis, proximal, replacement, satisfactory, situ, soft, suprapatellar, swelling, tissue, tissues*}
- Hip = {*acetabular, acute, alignment, bilateral, bone, bony, degenerative, enlocated, femoral, femur, fracture, fractures, hip, identified, intact, joint, lucency, mild, moderate, pelvic, pelvis, periprosthetic, proximal, pubic, ring, sacroiliac, sclerosis, symphysis*}
- Wrist = {*acute, alignment, anatomical, angulation, bony, carpal, cast, degenerative, displacement, distal, dorsal, fracture, healing, intact, interval, lateral, mild, plate, radial, radio-carpal, radius, scaphoid, styloid, swelling, tissue, transverse, ulnar, wrist*}
- Shoulder = {*acromioclavicular, acute, alignment, bony, calcification, change, clavicle, cuff, degenerative, dislocation, fracture, fractures, glenohumeral, glenoid, head, humeral, humerus, identified, inferior, intact, joint, lateral, proximal, shoulder, space, subacromial, tissue, tuberosity, visualised*}

D More Implementation Details

Considering the domain disparity between medical and generic texts, we use the tokeniser and text encoder from MedClip [15] to embed the report. The knowledge aggregation network consists of a three-layer Transformer [6]. For a fair comparison, following the setting of previous works, we configure the dimensions of input images to 224×224 and incorporate data augmentation techniques, such as random cropping and flipping, to expand the X-ray training dataset. We limit the maximum epochs to 100 and use the Adam optimiser [7] with a weight decay parameter of 1e-4. The learning rates are set at 5e-5 for the image encoder and 1e-4 for the remaining trainable parameters. Besides, based on the findings from our ablation study, we empirically set the hyper-parameter λ to 1.0. Our experiments are conducted using A100 GPUs.

E More Quantitative Results

To assess the quality of the generated captions, we use four widely used NLG evaluation metrics, *i.e.*, BLEU (B1∼B4) [10], ROUGE [9], METEOR [1] and

CIDEr [12]. As shown in Table 2, we report the average scores of all the above evaluation metrics. The results exhibit that regardless of bs = 96 or 192, our X-RGen consistently outperforms R2Gen in terms of all the average scores (except for ROUGE-L), which demonstrates its effectiveness in generating accurate and high-quality radiology reports. Specifically, when comparing R2Gen to our X-RGen in both the specialised and generalist settings, the improvements of R2Gen are 2.1%, −0.4%, −2.6%, −2.9%, 5.6%, −2.2% and 8.9% for BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE-L and CIDEr, respectively[3]. In contrast, our X-RGen achieves even larger improvements in these evaluation metrics about 8.3%, 7.4%, 6.7%, 6.8%, 6.9%, −0.6% and 22.7% separately. Moreover, we also report the values of all the evaluation metrics on these six datasets from Tables 3 to 8.

Table 2: Average results on the six datasets compared with the recent specialised models. † means we optimise the model on our merged training dataset while the "bs" is the training batch size. All evaluations are conducted on the test set, and a higher value indicates better performance.

	BLEU-1 (Ave)	BLEU-2 (Ave)	BLEU-3 (Ave)	BLEU-4 (Ave)	METEOR (Ave)	ROUGE-L (Ave)	CIDEr (Ave)
specialised models							
Transformer [11]	0.368	0.223	0.147	0.100	0.134	0.305	0.230
R2Gen [3]	0.374	0.229	0.149	0.101	0.141	**0.312**	0.257
R2GenCMN [2]	0.371	0.229	0.150	0.101	0.138	0.307	0.255
MSAT [14]	0.393	0.237	0.151	0.100	0.139	0.302	0.232
X-RGen (ours)	0.370	0.227	0.150	0.103	0.144	**0.312**	0.269
generalist models							
R2Gen† (bs=16)	0.345	0.200	0.126	0.082	0.133	0.289	0.222
R2Gen† (bs=96)	0.382	0.228	0.145	0.096	0.149	0.301	0.280
R2Gen† (bs=192)	0.369	0.225	0.145	0.098	0.146	0.305	0.274
X-RGen (ours, bs=16)	0.363	0.217	0.140	0.095	0.144	0.296	0.264
X-RGen (ours, bs=96)	0.383	0.231	0.151	0.104	0.149	0.306	0.327
X-RGen (ours, bs=192)	**0.401**	**0.244**	**0.160**	**0.110**	**0.154**	0.310	**0.330**

[3] For a fair comparison, we compare the highest results for both R2Gen and ours.

Table 3: Comparison with the recent specialised models on Chest (IU-Xray). † means we optimise the model on our merged training dataset while the "bs" is the training batch size. All evaluations are conducted on the test set, and a higher value indicates better performance.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
specialised models							
Transformer [11]	0.459	0.298	0.215	0.162	0.188	0.362	0.511
R2Gen [3]	0.470	0.304	0.219	0.165	0.187	0.371	0.430
R2GenCMN [2]	0.475	0.309	0.222	0.170	0.191	0.375	0.641
MSAT [14]	0.481	0.316	0.226	0.171	0.190	0.372	0.394
DCL [8]	-	-	-	0.163	0.193	**0.383**	0.586
METransformer [13]	**0.483**	**0.322**	**0.228**	0.172	0.192	0.380	0.435
X-RGen (ours)	0.441	0.285	0.208	0.163	0.184	0.361	0.609
generalist models							
R2Gen† (bs=16)	0.306	0.175	0.117	0.084	0.134	0.316	0.289
R2Gen† (bs=96)	0.433	0.275	0.196	0.147	0.184	0.355	0.470
R2Gen† (bs=192)	0.349	0.217	0.153	0.114	0.154	0.332	0.359
X-RGen (ours, bs=16)	0.444	0.287	0.202	0.152	0.190	0.365	0.509
X-RGen (ours, bs=96)	0.454	0.290	0.210	0.161	0.187	0.361	**0.700**
X-RGen (ours, bs=192)	0.466	0.306	0.225	**0.177**	**0.199**	0.367	0.602

Table 4: Comparison with the recent specialised models on Abdomen. † means we optimise the model on our merged training dataset while the "bs" is the training batch size. All evaluations are conducted on the test set, and a higher value indicates better performance.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
specialised models							
Transformer [11]	0.409	0.247	0.161	0.108	0.142	0.314	0.261
R2Gen [3]	0.389	0.241	0.156	0.105	0.143	0.309	0.248
R2GenCMN [2]	0.361	0.231	0.151	0.102	0.135	0.310	0.161
MSAT [14]	0.410	0.246	0.157	0.105	0.140	0.286	0.275
X-RGen (ours)	0.373	0.228	0.154	0.106	0.137	0.314	0.196
generalist models							
R2Gen† (bs=16)	0.386	0.238	0.154	0.104	0.144	0.297	0.280
R2Gen† (bs=96)	0.407	0.244	0.150	0.097	0.155	0.297	0.271
R2Gen† (bs=192)	0.397	0.240	0.151	0.100	0.153	0.296	0.271
X-RGen (ours, bs=16)	0.395	0.243	0.159	0.108	0.152	0.305	0.276
X-RGen (ours, bs=96)	0.409	0.252	0.162	0.110	0.159	0.313	0.292
X-RGen (ours, bs=192)	**0.432**	**0.269**	**0.175**	**0.118**	**0.161**	**0.322**	**0.327**

Table 5: Comparison with the recent specialised models on Knee. † means we optimise the model on our merged training dataset while the "bs" is the training batch size. All evaluations are conducted on the test set, and a higher value indicates better performance.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
specialised models							
Transformer [11]	0.304	0.177	0.116	0.078	0.115	0.288	0.169
R2Gen [3]	0.308	0.191	0.121	0.077	0.130	0.300	0.193
R2GenCMN [2]	0.329	0.201	0.130	0.083	0.120	0.284	0.164
MSAT [14]	**0.366**	0.203	0.128	0.082	0.134	0.282	0.135
X-RGen (ours)	0.339	0.207	0.133	0.087	0.135	0.295	0.175
generalist models							
R2Gen† (bs=16)	0.321	0.170	0.100	0.064	0.119	0.255	0.154
R2Gen† (bs=96)	0.343	0.197	0.120	0.075	0.134	0.284	0.181
R2Gen† (bs=192)	0.333	0.207	0.134	0.089	**0.139**	**0.308**	0.204
X-RGen (ours, bs=16)	0.315	0.180	0.111	0.071	0.124	0.276	0.166
X-RGen (ours, bs=96)	0.331	0.193	0.120	0.077	0.130	0.277	0.188
X-RGen (ours, bs=192)	0.359	**0.219**	**0.141**	**0.093**	**0.139**	0.291	**0.242**

Table 6: Comparison with the recent specialised models on Hip. † means we optimise the model on our merged training dataset while the "bs" is the training batch size. All evaluations are conducted on the test set, and a higher value indicates better performance.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
specialised models							
Transformer [11]	0.334	0.193	0.118	0.077	0.116	0.264	0.137
R2Gen [3]	0.358	0.211	0.131	0.082	0.131	0.288	0.210
R2GenCMN [2]	0.362	0.214	0.133	0.083	0.133	0.286	0.220
MSAT [14]	0.362	**0.218**	0.131	0.081	0.125	0.282	0.235
X-RGen (ours)	0.356	0.216	**0.135**	**0.086**	0.138	**0.294**	0.192
generalist models							
R2Gen† (bs=16)	0.351	0.199	0.120	0.074	0.132	0.275	0.203
R2Gen† (bs=96)	0.361	0.209	0.126	0.080	0.137	0.281	0.226
R2Gen† (bs=192)	**0.367**	0.214	0.133	**0.086**	**0.139**	0.285	0.238
X-RGen (ours, bs=16)	0.332	0.187	0.113	0.073	0.129	0.263	0.184
X-RGen (ours, bs=96)	0.366	0.211	0.130	0.084	0.137	0.281	**0.257**
X-RGen (ours, bs=192)	**0.367**	0.206	0.122	0.076	0.133	0.277	0.215

Table 7: Comparison with the recent specialised models on Wrist. † means we optimise the model on our merged training dataset while the "bs" is the training batch size. All evaluations are conducted on the test set, and a higher value indicates better performance.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
specialised models							
Transformer [11]	0.339	0.203	0.133	0.086	0.120	0.301	0.129
R2Gen [3]	0.359	0.214	0.139	0.093	0.135	0.299	0.288
R2GenCMN [2]	0.351	0.210	0.134	0.087	0.129	0.290	0.212
MSAT [14]	0.374	0.216	0.134	0.081	0.124	0.295	0.180
X-RGen (ours)	0.358	0.214	0.137	0.089	0.142	0.302	0.243
generalist models							
R2Gen† (bs=16)	0.351	0.207	0.133	0.085	0.136	0.293	0.217
R2Gen† (bs=96)	0.375	0.215	0.133	0.084	0.144	0.291	0.258
R2Gen† (bs=192)	0.389	**0.238**	**0.154**	**0.102**	0.148	**0.312**	0.296
X-RGen (ours, bs=16)	0.342	0.199	0.124	0.079	0.133	0.280	0.229
X-RGen (ours, bs=96)	0.368	0.217	0.138	0.090	0.144	0.298	0.255
X-RGen (ours, bs=192)	**0.390**	0.232	0.148	0.097	**0.149**	0.299	**0.305**

Table 8: Comparison with the recent specialised models on Shoulder. † means we optimise the model on our merged training dataset while the "bs" is the training batch size. All evaluations are conducted on the test set, and a higher value indicates better performance.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
specialised models							
Transformer [11]	0.363	0.219	0.138	0.088	0.123	0.301	0.192
R2Gen [3]	0.358	0.213	0.130	0.082	0.122	0.307	0.174
R2GenCMN [2]	0.348	0.210	0.129	0.082	0.119	0.297	0.134
MSAT [14]	0.364	0.221	0.131	0.080	0.123	0.297	0.173
X-RGen (ours)	0.353	0.211	0.133	0.088	0.129	0.304	0.197
generalist models							
R2Gen† (bs=16)	0.355	0.212	0.131	0.082	0.132	0.299	0.186
R2Gen† (bs=96)	0.374	0.225	0.142	0.095	0.142	0.297	0.274
R2Gen† (bs=192)	0.380	0.231	0.145	**0.096**	**0.144**	0.299	0.277
X-RGen (ours, bs=16)	0.350	0.207	0.128	0.084	0.133	0.288	0.220
X-RGen (ours, bs=96)	0.369	0.225	0.145	0.099	0.139	**0.304**	0.272
X-RGen (ours, bs=192)	**0.389**	**0.234**	**0.146**	**0.096**	0.141	0.302	**0.287**

References

1. Banerjee, S., Lavie, A.: 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. pp. 65–72 (2005) 4
2. Chen, Z., Shen, Y., Song, Y., Wan, X.: Cross-modal memory networks for radiology report generation. ACL-IJCNLP pp. 5904–5914 (2022) 5, 6, 7, 8
3. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. EMNLP pp. 1439–1449 (2020) 2, 5, 6, 7, 8
4. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association pp. 304–310 (2016) 2
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. NAACL pp. 4171–4186 (2019) 1
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. Int. Conf. Learn. Represent. (2021) 4
7. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 4
8. Li, M., Lin, B., Chen, Z., Lin, H., Liang, X., Chang, X.: Dynamic graph enhanced contrastive learning for chest x-ray report generation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3334–3343 (2023) 2, 6
9. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004) 4
10. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: ACL. pp. 311–318 (2002) 4
11. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Adv. Neural Inform. Process. Syst. **30** (2017) 5, 6, 7, 8
12. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4566–4575 (2015) 5
13. Wang, Z., Liu, L., Wang, L., Zhou, L.: Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 11558–11567 (2023) 6
14. Wang, Z., Tang, M., Wang, L., Li, X., Zhou, L.: A medical semantic-assisted transformer for radiographic report generation. In: MICCAI. pp. 655–664 (2022) 5, 6, 7, 8
15. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. arXiv preprint arXiv:2210.10163 (2022) 1, 4