Energy-Based Controllable Radiology Report Generation with Medical Knowledge

Zeyi Hou[1], Ruixin Yan[2], Ziye Yan[3], Ning Lang[2], and Xiuzhuang Zhou[1] ✉

[1] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China
xiuzhuang.zhou@bupt.edu.cn
[2] Department of Radiology, Peking University Third Hospital, Beijing, China
[3] Perception Vision Medical Technologies Co Ltd, Guangzhou, China

Abstract. Automated generation of radiology reports from chest X-rays has the potential to substantially reduce the workload of radiologists. Recent advances in report generation using deep learning algorithms have achieved significant results, benefiting from the incorporation of medical knowledge. However, incorporation of additional knowledge or constraints in existing models often require either altering network structures or task-specific fine-tuning. In this paper, we propose an energy-based controllable report generation method, named ECRG. Specifically, our method directly utilizes diverse off-the-shelf medical expert models or knowledge to design energy functions, which are integrated into pre-trained report generation models during the inference stage, without any alterations to the network structure or fine-tuning. We also propose an acceleration algorithm to improve the efficiency of sampling the complex multi-modal distribution of report generation. ECRG is model-agnostic and can be readily used for other pre-trained report generation models. Two cases are presented on the design of energy functions tailored to medical expert systems and knowledge. The experiments on widely used datasets Chest ImaGenome v1.0.0 and MIMIC-CXR demonstrate the effectiveness of our proposed approach.

Keywords: Radiology report generation · Chest X-ray · Energy based model · Controllable generation

1 Introduction

Chest radiography (chest X-ray, CXR) is currently the most prevalent medical imaging examination, serving as a pivotal tool in clinical diagnoses [19] and epidemiological research [18]. Writing a comprehensive and accurate radiology report proves to be a challenging and time-consuming task in practice. As a result, automated interpretation of chest X-rays through deep learning models has garnered considerable interest for its potential to substantially alleviate the workload of radiologists and enhance clinical efficiency.

Automatic radiology report generation is a challenging task, as it essentially involves converting complex visual input of chest X-rays into long text output

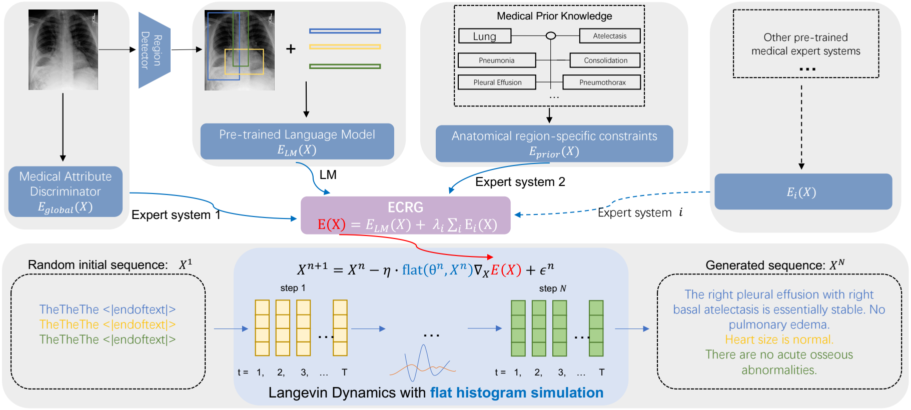

Region Detector
+
Medical Attribute Discriminator
$E_{global}(X)$
Pre-trained Language Model
$E_{LM}(X)$
LM
Medical Prior Knowledge
Lung
Atelectasis
Pneumonia
Consolidation
Pleural Effusion
Pneumothorax
...
Anatomical region-specific constraints
$E_{prior}(X)$
Expert system 2
Other pre-trained medical expert systems
...
$E_i(X)$
Expert system 1
ECRG
$E(X) = E_{LM}(X) + \lambda_i \sum_i E_i(X)$
Expert system i
Random initial sequence: X^1
TheTheThe <|endoftext|>
TheTheThe <|endoftext|>
TheTheThe <|endoftext|>
$X^{n+1} = X^n - \eta \cdot \mathrm{flat}(\theta^n, X^n)\nabla_X E(X) + \epsilon^n$
step 1
step N
t = 1, 2, 3, ... T
...
t = 1, 2, 3, ... T
Langevin Dynamics with flat histogram simulation
Generated sequence: X^N
The right pleural effusion with right basal atelectasis is essentially stable. No pulmonary edema.
Heart size is normal.
There are no acute osseous abnormalities.

Fig. 1. Overview of the proposed ECRG framework for radiology report generation.

that accurately describes the medical observations. Recently, extensive works [3,11,24,26,28] has been devoted to generating informative and semantically coherent reports using deep learning algorithms. These approaches typically follow the encoder-decoder paradigm and existing methods can be categorized into two main concepts. The first involves designing improved network structures or training losses for the image encoder and text decoder to enhance cross-modal feature matching. For example, R2Gen [3] introduced a relational memory to record key information and integrate it into the Transformer decoder. AlignTransformer [26] proposed an align hierarchical attention module and a multi-grained Transformer structure to address issues related to data bias and long sequences. RGRG [22] employed an object detector to extract anatomical regions and produced corresponding localized visual features, which were then leveraged to generate sentences delineating anatomy-specific pathological observations. However, such approaches necessitate modifications to the network structures, leading to limited flexibility and transparency. The second concept involves incorporating additional medical knowledge into the process of image feature extraction or text generation. For instance, PPKED [11] integrated disease labels and medical knowledge graphs into the encoder process to mitigate visual and textual data biases. [28] introduced a graph convolutional neural network pre-constructed on multiple disease findings to assist report generation. Despite advancements in medical knowledge-guided report generation, these methods still require the customization of intricate network architectures and retraining.

In this paper, we propose an Energy-based Controllable Radiology Report Generation (ECRG) method to cope with the above issues. As illustrated in Fig. 1, ECRG circumvents the training process and flexibly integrates various off-the-shelf pre-trained medical expert systems into the report generation pipeline using the energy-based framework. The contributions of our work are as follows: 1)Inspired by constrained text generation [14,17], we propose ECRG framework to bypass the training process and utilize medical knowledge to control report generation during the inference phase. 2)We present two cases illustrating the use of pre-trained expert models and medical prior knowledge to design energy functions aimed at improving radiology report generation within the ECRG framework. 3)Lastly, we devise an acceleration algorithm based on the idea of flat histogram simulation [6,9,23] to improve the efficiency of report generation. Experimental results on large public datasets Chest ImaGenome v1.0.0 [25] and MIMIC-CXR [7,8] confirm the effectiveness of our proposed approach.

2 Method

2.1 Energy-based Controllable Report Generation

Given a radiology image I, the task of radiology report generation aims to automatically generate a descriptive report $R = \{X_1, X_2, ..., X_N\}$, where X_i denote the sentences in the report. Typically, this task is achieved with an encoder-decoder framework: 1)*encoding*: The radiology image I is first encoded to an image embedding $e_I = Encoder\left(I\right)$. 2)*decoding*: Then, a report generation language model P_{LM} computes the probability distribution of the sequences $p_{LM}\left(X\right) = LM\left(X, e_I\right)$. The final radiology report is obtained from the probability distribution of the sequences through different search strategies (*e.g.*, greedy search, beam search, sampling, etc.).

We draw inspiration from the constrained text generation [14,17] and use an energy-based framework during the decoding process to enhance the generated reports. Let p_{expt} denote the sequence distribution satisfying a certain desired attribute, which is governed by a specific medical expert model P_{expt}. Generating radiology reports (sequences) using the report generation model P_{LM}, enhanced by various expert systems, can be viewed as drawing samples from the following distribution that defined by an energy-based model:

$$p\left(X\right) = e^{\log\left(p_{LM}(X)\cdot\prod_i \lambda_i p^i_{expt}(X)\right)}/Z \tag{1}$$

where $\lambda_i \geq 0$ denote the weight of the i-th expert model constraint, and Z is a normalization constant. Intuitively, the Boltzmann distribution in Eqn. (1) can be seen as a revised report generation distribution, which corrects the pre-trained distribution $p_{LM}\left(X\right)$ through a weighted multiplication of the desired attribute distributions $p^i_{expt}\left(X\right)$. Here, the energy function can be written as:

$$E\left(X\right) = -\log\left(p_{LM}\left(X\right)\right) - \log\left(\prod_i \lambda_i p^i_{expt}\left(X\right)\right) = E_{LM}\left(X\right) + \sum_i E_i\left(X\right) \tag{2}$$

whose energy values are linear combinations of scores from various independent energy functions. In this way, we can flexibly and controllably incorporate arbitrary energy functions $E_i\left(X\right)$ that are defined by pre-trained black-box expert systems P^i_{expt} (*e.g.* pathological discriminators or medical prior knowledge) into the off-the-shelf report generation model P_{LM}, without any structural change or task-specific fine-tuning. Furthermore, the energy function is defined over a sequence consisting of continuous vectors $X = \{x_1, ..., x_T\}$, where $x_t\left(v\right) \in \mathbb{R}^V$ denote the logits of token v in position t, and V is the size of the vocabulary.

In general, each expert model scores a desired attribute of the generated reports. These attributes can be medical diagnosis information, disease distribution, and pathological relationships, etc. We view the product of these medical expert models as a probabilistic energy model, allowing a flexible combination of various heterogeneous attributes and refining the report generation distribution during the inference phase. To ensure applicability, the pre-trained expert model in the proposed ECRG framework requires accurately reflecting the satisfaction degree of the relevant medical attributes through the score, which is used to calculate the energy function in Eqn. (2) and thus affects the quality of the generated report.

2.2 Heuristic Constraints with Medical Knowledge

The proposed framework is model-agnostic and can be transferred to other pre-trained report generation models. As an example, we utilize the recently proposed anatomical region-guided report generation model RGRG [22] as P_{LM}. As shown in Fig. 1, given a frontal chest X-ray image, RGRG [22] initially employs an object detector to extract anatomical regions and produce corresponding localized visual features. These features are then leveraged to generate sentences delineating anatomy-specific pathological observations.

Fusion of Multi-grained Medical Image Information Generating sentences that describe specific anatomical regions enhances the transparency and interpretability of report generation, while discerning certain pathological observations necessitates holistic consideration of the global X-ray image. For instance, *"Cardiomegaly"* is a significant pathological observation, with positive cases accounting for 17.2% of the MIMIC-CXR dataset. Diagnosing this disease necessitates the calculation of the cardiothoracic ratio, thus requiring the integration of multi-granularity features extracted from global and local X-ray images to generate accurate descriptions of the anatomical region *"Cardiac Silhouette"*. We use a pre-trained chest X-ray image classifier D_l from the TorchXRayVision library [4] to acquire the identification result of label l, $D_l(I) \in [0,1]$. This result containing global image information is then utilized to construct an energy function $E_{global}(X)$, which governs the medical semantics of the generated descriptions of the corresponding region. The expression of the fused energy function is:

$$\begin{aligned} E_{fuse}(X) &= E_{LM}(X) + E_{global}(X) = E_{local}(X) + E_{global}(X) \\ &= -\sum\nolimits_{t=1}^{T}\sum\nolimits_{v\in V}^{T} p_{LM}(v \mid X_{<t}, I_l) \log softmax(X_t(v)) \\ &\quad - [2 \cdot D_l(I) - 1] \cdot \varphi_{sim}(X \mid X^*) \end{aligned} \tag{3}$$

where $p_{LM}(\cdot \mid X_{<t}, I_l)$ denotes the distribution of the next token when providing the region-based report generation model P_{LM} with previous tokens $X_{<t}$. Intuitively, the energy function $E_{LM}(X)$ employs negative cross-entropy to align the distribution of generated sequences X_t with that represented by the pre-trained report generation model $p_{LM}(\cdot \mid X_{<t}, I_l)$. $\varphi_{sim}(X \mid X^*)$ represents the semantic similarity between the generated sequence X and the reference sequences

describing specific diseases. This is achieved by computing the BERTScore [27] between the two sequences. The linear combination of E_{LM} and E_{global} comprehensively incorporates local and global features of CXR images into the report generation process. Unlike other methods that introduce medical knowledge with structural modifications or fine-tuning, our proposed approach only requires designing energy functions based on attributes of the pre-trained expert systems.

Fusion of Anatomical Region-based Prior Knowledge The energy function $E_{LM}\left(X\right)$ in Eqn. (3) implicitly ensures the quality of the generated sequences, leveraging the pre-trained report generation model. However, the maximum likelihood estimation (MLE) employed in training such autoregressive language models, along with the aforementioned modifications to the underlying probability distribution, may cause text degradation, such as inconsistency with the corresponding anatomical region, information loss, etc. For example, the generated description of the *"Left Lung"* region by the RGRG model may only involve the observation *"Pneumothorax"* while ignoring other diseases. Hence, we devise an energy function that imposes soft constraints on presence/absence of keywords (or sequences) W_r related to the corresponding anatomical region r (according to medical prior knowledge in Fig. 1), prompting the generation of relevant descriptions and suppressing irrelevant ones to alleviate text degradation:

$$E_{prior}\left(X\right) = \textstyle\sum_j \pm f_{match}^{n-gram}\left(X, W_r^j\right) \tag{4}$$

where + and − control the presence and absence of keywords (sequences). $f_{match}^{n-gram}\left(\cdot,\cdot\right)$ represents a differentiable matching loss EISL [12] based on n-gram similarity, where $f_{match}^{1-gram}\left(\cdot,\cdot\right)$ is suitable for keywords and $f_{match}^{n-gram}\left(\cdot,\cdot\right), n > 1$ is suitable for sequences of length n.

The overall energy function of this anatomical region can be expressed as:

$$E\left(X\right) = \lambda_1 E_{LM}\left(X\right) + \lambda_2 E_{global}\left(X\right) + \lambda_3 E_{prior}\left(X\right) \tag{5}$$

2.3 Accelerate Report Generation

Generating enhanced radiology reports can be interpreted as sampling from the energy-based model $E\left(X\right)$ in Eqn. (5). Given the differentiability of the aforementioned energy functions, Langevin dynamics can be employed for sampling by forming a Markov chain: $X^{(k+1)} \leftarrow X^k - \frac{\eta}{2}\bigtriangledown_X E\left(X^k\right) + \epsilon_k$ as in [17], where X^k is the sample at iteration k, η denotes the step size and $\epsilon_k \in \mathcal{N}\left(0,\sigma\right)$ is the random noise. By iteratively applying this update rule, a sequence of samples that adhere to the target report distribution can be progressively generated. However, when dealing with the complex multi-modal distribution $p\left(X\right) = e^{-E(X)}$ of sequences X in controllable report generation, crossing the energy barriers becomes challenging, leading to low efficiency and suboptimal sampling.

We adopt the principles of flat histogram simulations [6,9,23] form physics to expedite the process of sampling reports. The sample space of sequence X is partitioned into m disjoint subspaces $\{s_1, s_2, ..., s_m\}$ with equal-size energy levels $s_i = \{X : E_{i-1} < E\left(X\right) < E_i\}$, according to the energy function $E\left(X\right)$.

We introduce an auxiliary variable

$$\phi_{\theta}\left(E\left(X\right)\right)=\sum\nolimits_{i=1}^{m}\left(\theta\left(i-1\right)e^{\left(\log\theta\left(i\right)-\log\theta\left(i-1\right)\right)\frac{E\left(X\right)-E_{i-1}}{\triangle E\left(X\right)}}\right)\mathbb{1}_{E\left(X\right)\in s_{i}} \tag{6}$$

parameterized by θ to simulate from a falattend density $\tilde{p}\left(X\right)=\frac{p\left(X\right)}{\phi_{\theta}\left(E\left(X\right)\right)}$, where $\mathbb{1}_{\left(\cdot\right)}$ denotes the indicator function. $\phi_{\theta}\left(E\left(X\right)\right)$ monitors the *spectral density* of each energy subspaces during sampling, thus facilitates traveling across energy barriers in the rugged energy landscape of $E\left(X\right)$, accelerating controllable report generation. Here, the sequence X can be efficiently sampled with the Markov chain:

$$X^{k+1}=X^{k}-\eta_{k+1}\left[1+\frac{\log\theta_{k}\left(In\left(X^{k}\right)\right)-\log\theta_{k}\left(\left(In\left(X^{k}\right)-1\right)\vee1\right)}{\triangle E\left(X\right)}\right]\bigtriangledown_{X}E\left(X\right)+\delta \tag{7}$$

where $In\left(X^{k}\right)$ denotes the index that $E\left(X^{k}\right)$ belongs to. η_{k+1} is the learning rate and $\delta=\sqrt{2\eta_{k+1}}\epsilon_{k+1}$, $\epsilon_{k+1}\in\mathcal{N}\left(0,\sigma\right)$ is the random noise. At the end of the iteration k, we update the *spectral density* for subspace s_i according to $\theta_{k+1}\left(i\right)=\theta_{k}\left(i\right)+\epsilon_{k+1}\theta_{k}\left(In\left(X_{k+1}\right)\right)\left(\mathbb{1}_{i=In\left(X_{k+1}\right)}-\theta_{k}\left(i\right)\right)$

3 Experiments and Results

3.1 Datasets and Experimental Settings

We conduct experiments on the open-source Chest ImaGenome v1.0.0 [25] dataset, which is derived from the most widely used public dateset MIMIC-CXR [7,8], which consists 77,110 chest X-ray images corresponding to 227,835 free-text radiology reports. The Chest ImaGenome dataset extends the MIMIC-CXR dataset with automatically constructed scene graphs. Each scene graph includes a frontal MIMIC-CXR image and corresponding bounding box coordinates of 29 anatomical regions. Additionally, the sentences in the *findings* section of the free-text radiology reports are assigned to the anatomical region they describe, enabling region-guided radiology report generation.

As previously mentioned, our method is model agnostic and we adopt RGRG [22] as the baseline model. For fair comparison, we employ the exact same data processing method, network structure, and pre-trained model as the baseline method. We use the model *"xrv.models.DenseNet"* available in the TorchXRayVision library [4] as the pathological discriminator D_l, which loads the pre-trained weights *"densenet121-res224-all"*. We use the pre-trained uncased base version of DistilBERT [20] (*"distilbert-base-uncased"*), obtained from Huggingface, to compute BERTScore [27] for semantic similarity measure of sequences. Since BERTScore can capture sentences with similar semantics but different expressions, a simple reference sentence list suffices for the description of certain diseases, *e.g.*, we use [*"There is cardiomegaly.", " heart size is enlarged." "cardiac silhouette is enlarged."*] for *"Cardiomegaly"*. The weights of the linear combinations of energy functions defined by different expert systems in Eqn. (5) are set to 0.5, 0.3, and 0.2. For report generation sampling, the length of the generated

Table 1. Natural language generation (NLG) metrics and clinical efficacy (CE) metrics micro-averaged over 14 observations for the report generation task.

Method	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	P_{14}	R_{14}	$F_{1,14}$
RGRG[22]	0.373	0.249	0.175	0.126	0.168	0.264	0.461	0.475	0.447
ECRG(Multi)	0.366	0.242	0.167	0.123	**0.173**	0.260	**0.464**	0.481	0.471
ECRG(Prior)	**0.382**	**0.255**	**0.178**	**0.128**	0.161	**0.269**	0.456	0.513	0.482
ECRG(Full)	0.379	0.253	0.175	0.123	0.164	0.266	0.460	**0.519**	**0.488**

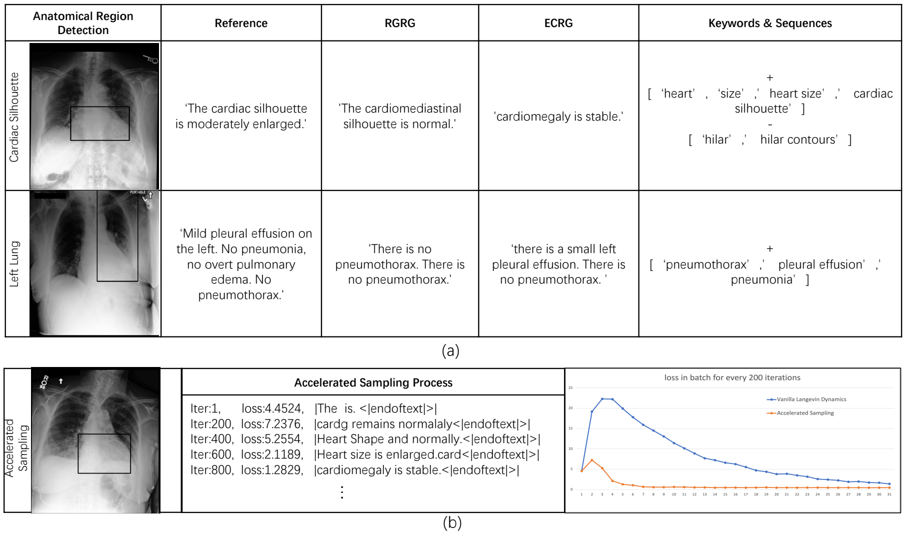
sequence X is fixed to 10, and p_{LM} is used to produce the continuation of X by greedy search until the end of the sentence. We run the Markov chain for 2,000 steps in advance to determine the upper and lower bounds of the energy. The energy space is then partitioned into 50 subregions, and the learning rate η is set to 0.01. The batch size for Langevin dynamics is set to 32, and the sample with the minimum loss is selected as the final generated sequence.

We compute widely used Natural Language Generation (NLG) metrics: BLEU [15], METEOR [1], and ROUGE-L [10] to assess the quality of the generated reports. The NLG metrics measure textual similarity between the generated reports and the reference reports, but are ill-suited to capture the clinical correctness of generated reports [2,13,16]. Further, we also report the clinical efficacy (CE) metrics to measure the diagnostic accuracy of generated reports. The CE metrics compare the presence status of 14 clinical observations (extracted from reports by CheXbert [21]) between the generated reports and the reference reports. Since the data with different clinical observation labels in the MIMIC-CXR dataset are unbalanced, we compute the CE scores by micro-averaging over 14 observations following [5,22]. Higher micro-averaged scores imply better performance for major disease categories.

3.2 Results and Analysis

The comparison of the quality of reports generated by different methods are shown in Table 1. RGRG [22] is the baseline model, ECRG(Multi) and ECRG(Prior) respectively represent adding the multi-grained image information fusion module, as well as the region-specific information fusion module to the baseline model for ablation. ECRG(Full) denotes the full model described in Eqn. (5).

We can observe that ECRG(Full) achieves better or comparable performance on NLG metrics compared to the baseline model, indicating that the energy-based control method can ensure the quality of generated reports. ECRG(Multi) has a slight decrease on the n-gram (*i.e.*, word overlap) based metrics (*i.e.*, BLEU and ROUGE-L) due to the competing constraint that sacrifice words matching during the decoding process. Different from the n-gram metrics, the METEOR metric takes similarity matches between words into account, which correlates better with semantic judgments of report quality. ECRG(Multi) shows supe-



Anatomical Region Detection	Reference	RGRG	ECRG	Keywords & Sequences
Cardiac Silhouette	'The cardiac silhouette is moderately enlarged.'	'The cardiomediastinal silhouette is normal.'	'cardiomegaly is stable.'	+ ['heart' , 'size' ,' heart size' ,' cardiac silhouette'] - ['hilar' ,' hilar contours']
Left Lung	'Mild pleural effusion on the left. No pneumonia, no overt pulmonary edema. No pneumothorax.'	'There is no pneumothorax. There is no pneumothorax.'	'there is a small left pleural effusion. There is no pneumothorax. '	+ ['pneumothorax' ,' pleural effusion' ,' pneumonia']

Accelerated Sampling Process

```
Iter:1,     loss:4.4524,  |The  is. <|endoftext|>|
Iter:200,  loss:7.2376,  |cardg remains normalaly<|endoftext|>|
Iter:400,  loss:5.2554,  |Heart Shape and normally.<|endoftext|>|
Iter:600,  loss:2.1189,  |Heart size is enlarged.card<|endoftext|>|
Iter:800,  loss:1.2829,  |cardiomegaly is stable.<|endoftext|>|
                    ⋮
```

Fig. 2. (a) Generated reports for anatomical regions with different methods from chest X-ray images. (b) The process of accelerated sampling, as well as the loss comparison of different sampling methods during the Markov chain transition

rior performance on METEOR than the baseline method. This is because the image-level information improves the diagnostic accuracy of medical observation, while the similarity constraint steer the decoding procedure to generate sentences that are semantically closer to the reference report. Additionally, ECRG(Prior) achieves better NLG scores due to the soft constraints on keywords or sequences related to corresponding anatomical region. As for the CE metrics that measure the diagnostic accuracy of generated reports, ECRG(Multi) achieves higher scores because the pre-trained expert system can make more accurate judgments about diseases based on global image features. ECRG(Prior) significantly improves the recall of diseases with anatomical region-specific soft constraints, which is significant because missed diagnoses are often unacceptable in the medical field. The full model integrating these two modules also achieves better recall and f1 score, as well as comparable precision to the baseline model. which illustrates that our proposed energy-based controllable report generation framework can effectively improve the diagnostic ability of generated reports.

The visualization of the generated report is shown in Fig. 2. The first row shows a report generated for the *"Cardiac Silhouette"* region. RGRG [22] generates corresponding reports based solely on local region features and fails to diagnose *"Cardiomegaly"*. In contrast, the proposed ECRG method comprehensively considers global and local image information, accurately diagnosing "Cardiomegaly" and generating correlated descriptions. The second row shows a report generated for the *"Left Lung"* region. ECRG introduces soft constraints with anatomical region-specific keywords/sequences [*pneumothorax,pleural effu-*

sion,pneumonia], reducing the misdiagnosis rate of diseases. Additionally, Fig. 2 also illustrates the process of accelerated sampling, along with the loss comparison of different sampling methods during the Markov chain transition. Our proposed acceleration algorithm can produce reasonable samples in about 800 iterations, while the vanilla Langevin dynamics requires approximately 5,000 iterations to generate comparable reports.

4 Conclusion

In this paper, we have devised an energy-based controllable radiology report generation method, which is model-agnostic and can be transferred to other pre-trained report generation models. Unlike state-of-the-art alternatives, we directly leverage off-the-shelf medical expert models or knowledge to formulate energy functions, which are then incorporated into pre-trained langeuage models for report generation during the reference stage, without any modifications to the network structures or fine-tuning. We also devised an acceleration algorithm to efficiently sample complex multi-modal distributions in controllable report generation. Regarding the limitations of ECRG, first, it is necessary to design reasonable energy functions, whose values accurately reflect the properties of the pre-trained models. Additionally, although the ECRG framework based on the energy model can flexibly combine various heterogeneous energy functions through linear combination, different energy functions may inhibit each other. Addressing these problems warrants in-depth research and discussion in the future.

Acknowledgements This work is partially supported by the National Natural Science Foundation of China under grants 61972046, the Beijing Natural Science Foundation under grant Z190020, and the Proof of Concept Program of Zhongguancun Science City and Peking University Third Hospital under grant HDCXZHKC2022202.

Disclosure of Interests The authors have no competing interests.

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)
2. Boag, W., Hsu, T.M.H., McDermott, M., Berner, G., Alesentzer, E., Szolovits, P.: Baselines for chest x-ray report generation. In: Machine learning for health workshop. pp. 126–140. PMLR (2020)
3. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1439–1449 (2020)

4. Cohen, J.P., Viviano, J.D., Bertin, P., Morrison, P., Torabian, P., Guarrera, M., Lungren, M.P., Chaudhari, A., Brooks, R., Hashir, M., Bertrand, H.: TorchXRayVision: A library of chest X-ray datasets and models. In: Medical Imaging with Deep Learning (2022), https://github.com/mlmed/torchxrayvision
5. Dalla Serra, F., Wang, C., Deligianni, F., Dalton, J., O'Neil, A.Q.: Finding-aware anatomical tokens for chest x-ray automated reporting. In: International Workshop on Machine Learning in Medical Imaging. pp. 413–423. Springer (2023)
6. Deng, W., Lin, G., Liang, F.: A contour stochastic gradient langevin dynamics algorithm for simulations of multi-modal distributions. Advances in neural information processing systems **33**, 15725–15736 (2020)
7. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)
8. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
9. Liang, F.: An overview of stochastic approximation monte carlo. Wiley Interdisciplinary Reviews: Computational Statistics **6**(4), 240–254 (2014)
10. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
11. Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y.: Exploring and distilling posterior and prior knowledge for radiology report generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13753–13762 (2021)
12. Liu, G., Yang, Z., Tao, T., Liang, X., Bao, J., Li, Z., He, X., Cui, S., Hu, Z.: Don't take it literally: An edit-invariant sequence loss for text generation. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 2055–2078 (2022)
13. Liu, G., Hsu, T.M.H., McDermott, M., Boag, W., Weng, W.H., Szolovits, P., Ghassemi, M.: Clinically accurate chest x-ray report generation. In: Machine Learning for Healthcare Conference. pp. 249–269. PMLR (2019)
14. Mireshghallah, F., Goyal, K., Berg-Kirkpatrick, T.: Mix and match: Learning-free controllable text generationusing energy language models. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 401–415 (2022)
15. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
16. Pino, P., Parra, D., Messina, P., Besa, C., Uribe, S.: Inspecting state of the art performance and nlp metrics in image-based medical report generation. arXiv preprint arXiv:2011.09257 (2020)
17. Qin, L., Welleck, S., Khashabi, D., Choi, Y.: Cold decoding: Energy-based constrained text generation with langevin dynamics. Advances in Neural Information Processing Systems **35**, 9538–9551 (2022)
18. Rahman, T., Khandakar, A., Qiblawey, Y., Tahir, A., Kiranyaz, S., Kashem, S.B.A., Islam, M.T., Al Maadeed, S., Zughaier, S.M., Khan, M.S., et al.: Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. Computers in biology and medicine **132**, 104319 (2021)
19. Raoof, S., Feigin, D., Sung, A., Raoof, S., Irugulpati, L., Rosenow III, E.C.: Interpretation of plain chest roentgenogram. Chest **141**(2), 545–558 (2012)

20. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
21. Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A.Y., Lungren, M.: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1500–1519 (2020)
22. Tanida, T., Müller, P., Kaissis, G., Rueckert, D.: Interactive and explainable region-guided radiology report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7433–7442 (2023)
23. Wang, F., Landau, D.: Efficient, multiple-range random walk algorithm to calculate the density of states. Physical Review Letters **86**(10), 2050–2053 (2001)
24. Wang, Z., Tang, M., Wang, L., Li, X., Zhou, L.: A medical semantic-assisted transformer for radiographic report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 655–664. Springer (2022)
25. Wu, J.T., Agu, N.N., Lourentzou, I., Sharma, A., Paguio, J.A., Yao, J.S., Dee, E.C., Mitchell, W., Kashyap, S., Giovannini, A., et al.: Chest imagenome dataset (version 1.0. 0). PhysioNet **5**, 18 (2021)
26. You, D., Liu, F., Ge, S., Xie, X., Zhang, J., Wu, X.: Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. pp. 72–82. Springer (2021)
27. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. In: International Conference on Learning Representations (2019)
28. Zhang, Y., Wang, X., Xu, Z., Yu, Q., Yuille, A., Xu, D.: When radiology report generation meets knowledge graph. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12910–12917 (2020)