A Task Settings

Table 1. Hyper-parameter configurations and evaluation metric in the experiments.

Name of hyperparameter	Value	Evaluation metric
Number of anatomy ($\mathcal{N}$)	10	**Clinical Efficacy**: evaluate the F-1 scores of clinical findings
Number of abnormalities ($\mathcal{K}$)	19	**RadRQI**: evaluate the F-1 scores of Top-50 abnormalities
Number of report formats ($\mathcal{M}$)	3	and their associated attributes (TopK), and the number
Number of dimensions ($\mathcal{D}$)	1,024	of abnormalities of which F-1 scores are not 0 (Hits)
Number of memory slots ($\mathcal{E}$)	32	**Uncertainty Accuracy**: evaluate the F-1 scores of
Number of patches ($\mathcal{HW}$)	16×16	abnormality with single uncertain diagnosis (Hard)
Number of degree ($\mathcal{C}$)	10	and multiple nested uncertain diagnosis (Soft)

Table 2. 19 abnormalities with uncertain diagnosis, 10 anatomical parts identified by RadGraph and Chest ImaGemone, and three report formats considered.

Anatomical Parts	Left lung, Right lung, Left hilar, Right hilar, Cardiac silhouette, Mediastinum, Upper mediastinum, Left apical zone, Right apical zone, Spine
Abnormalities	Atelectasis, Bone deformity, Calcification, Consolidation, Edema, Enlarged cardiac silhouette, Emphysema/COPD, Hernia, Granuloma, Lesion, Medical device, Fracture, Opacity, Other findings, Pleural effusion, Pneumothorax, Scarring, Thickening, Tube/line
Report Format	Diagnostic order, Reporting length, Level of detail

B Additional Results

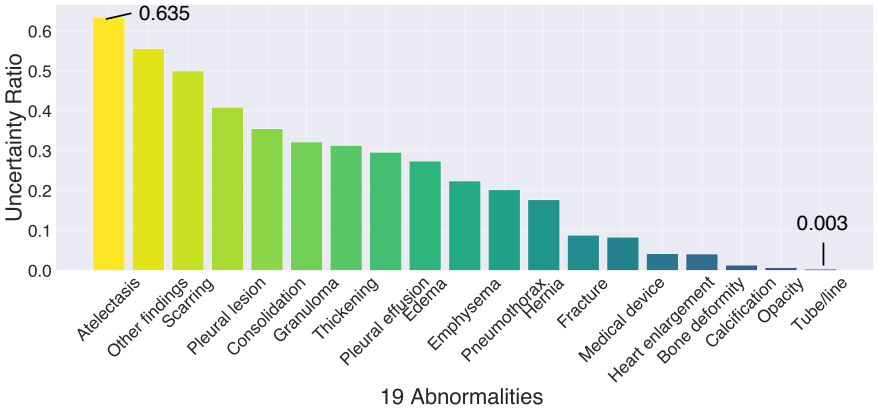
Table 3. Evaluation on report generation quality on MIMIC CXR. DiagDE shows comparable performances in NLG quality when high clinical accuracy is achieved.

Model	Natural Language Generation (NLG)		
	BLEU	ROUGE	CIDEr
Transformer	0.120	0.160	<u>0.135</u>
$\mathcal{M}^2$Transformer	**0.159**	**0.250**	0.100
R2Gen	0.120	0.161	**0.170**
R2Gen-CMN	<u>0.143</u>	<u>0.203</u>	0.132
WCL	0.122	0.152	0.031
XProNet	0.111	0.153	0.030
GIT	0.081	0.161	0.056
DiagUE (proposed)	0.140	0.169	0.109

Table 4. Performance on report generation on an additional IU Xray dataset (2,848).

Model	CE		RadRQI-F1		Uncertainty Acc.	
	(14)	(19)	TopK	Hits	Hard	Soft
R2Gen	0.289	0.310	<u>0.071</u>	10.0	<u>0.423</u>	<u>0.432</u>
R2Gen-CMN	0.290	0.323	0.056	10.0	0.368	0.375
WCL	<u>0.595</u>	<u>0.630</u>	0.065	<u>16.0</u>	0.349	0.362
XProNet	0.584	0.599	0.038	**17.0**	0.321	0.339
GIT	0.522	0.569	0.058	9.0	0.267	0.276
DiagUE (proposed)	**0.626**	**0.700**	**0.076**	13.0	**0.446**	**0.458**

Fig. 1. Uncertain ratio of 19 radiology abnormalities (# of uncertainty/ # of presence diagnosis) from MIMIC CXR, which underscore the necessity of uncertainty estimation.



C Ablation Study

Table 5. Performance on multi-label classification (AUC) of *abnormality* and *uncertainty* prediction. (Variability) corresponds to three diagnostic variations proposed.

Model (setting)	**Abnormality Prediction**	**Uncertainty Prediction**
FFN (Multi-Class)	0.591	0.573
FFN (Multi-Label)	0.681	0.690
FFN (Entropy)	0.779	0.561
FFN (Variability)	0.770	**0.724**
DiagUE (Multi-Class)	0.771	0.700
DiagUE (Multi-Label)	0.763	0.744
DiagUE (Variability)	**0.804**	**0.782**

Table 6. Ablation study on report generation by MIMIC CXR data. DiagUE$^\sharp$ corresponding to DiagUE w/ $(b, u, \mathcal{X})$ is the proposed approach.

Model	**CE**		**RadRQI-F1**		**Uncertainty**		**NLG**		
	(14)	(19)	TopK	Hits	Hard	Soft	BLEU	ROUGE	CIDEr
DiagUE w/ $(\emptyset)$	0.663	0.669	0.310	25.5	0.370	0.430	0.088	0.159	0.050
DiagUE w/ (b)	0.659	**0.697**	**0.322**	27.0	0.370	0.412	0.088	0.161	0.066
DiagUE w/ (b, u)	0.654	0.693	0.320	30.0	0.438	0.467	0.120	0.153	0.059
DiagUE w/ $(\mathcal{X})$	**0.666**	0.590	0.300	26.0	0.360	0.429	**0.145**	**0.199**	**0.109**
DiagUE $^\sharp$	0.664	0.688	0.319	**31.5**	**0.441**	**0.473**	0.140	0.169	**0.109**