Rad-ReStruct: A Novel VQA Benchmark and Method for Structured Radiology Reporting

Chantal Pellegrini*[1], Matthias Keicher*[1], Ege Özsoy[1], and Nassir Navab[1]

[1]Chair for Computer-Aided Medical Procedures and Augmented Reality, Technische Universität München, Germany

*Contributed equally.

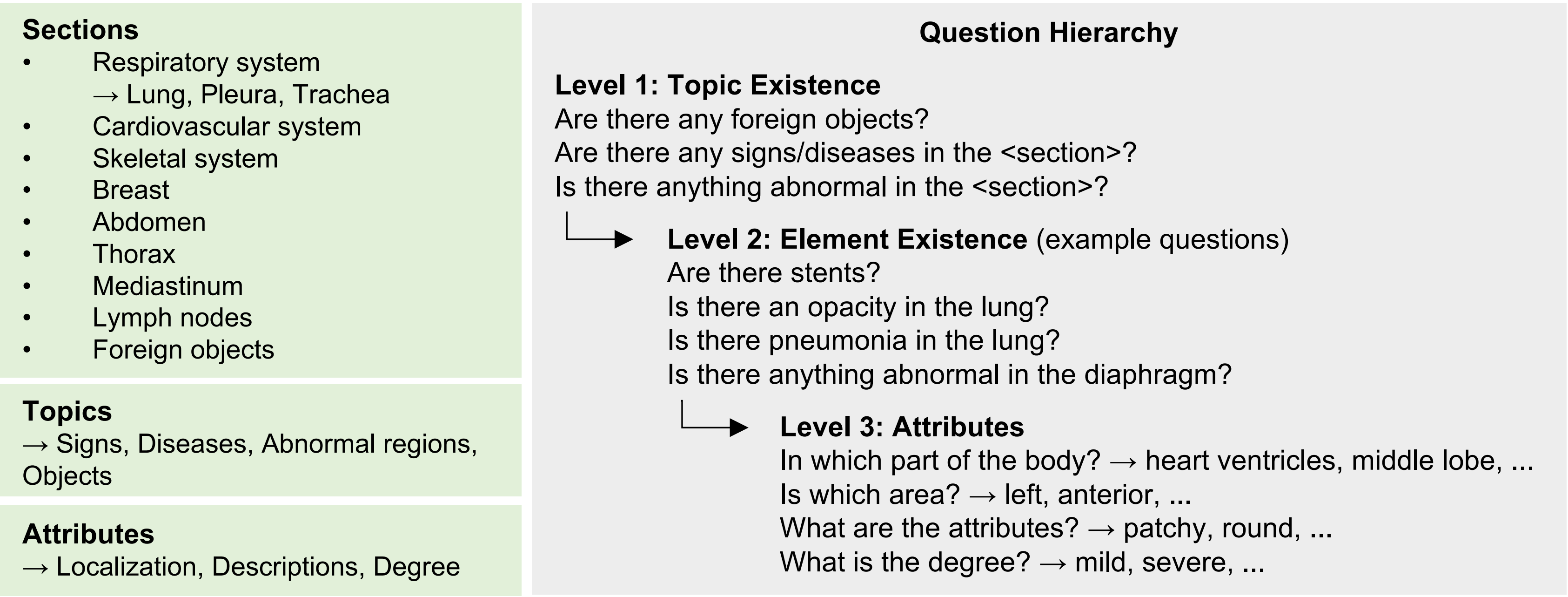
Motivation

Radiology reporting is an omnipresent and crucial task, but can be time-consuming and error-prone. Structured reporting offers a more efficient and accurate alternative to free-text reports and allows for a formalized and accurate evaluation of automated generation. Yet, there's a research gap in automating this process with no benchmark for method evaluation. **We present Rad-ReStruct, the first structured reporting benchmark** enabling the development of automated structured reporting methods. We further propose **hi-VQA**, a hierarchical VQA approach for integrating past questions and answers for accurate report generation.

Rad-ReStruct Dataset

Large-scale free-text reporting datasets exist, however free-text can be ambiguous, completeness can't be ensured and objective evaluation is difficult. Rad-ReStruct provides a structured, detailed and multi-level report template covering multiple sections in Chest X-Ray reports such as the respiratory, cardiovascular and skeletal system. The hierarchical structure is modeled after clinical structured reporting templates. By providing this first public structured reporting dataset and benchmark, we enable further research in automated structured reporting.

- Our reports provide an accurate summary of findings derived from expert-labeled MeSH and RadLex codes from the IU-XRay dataset [1]
- Rad-ReStruct encompasses **3720 images** and **3597 structured reports** with over **180k question-answer pairs**
- We provide a standardized, multi-level evaluation, which enforces consistency



Sections

- Respiratory system
  → Lung, Pleura, Trachea
- Cardiovascular system
- Skeletal system
- Breast
- Abdomen
- Thorax
- Mediastinum
- Lymph nodes
- Foreign objects

Topics
→ Signs, Diseases, Abnormal regions, Objects

Attributes
→ Localization, Descriptions, Degree

Question Hierarchy

Level 1: Topic Existence
Are there any foreign objects?
Are there any signs/diseases in the <section>?
Is there anything abnormal in the <section>?

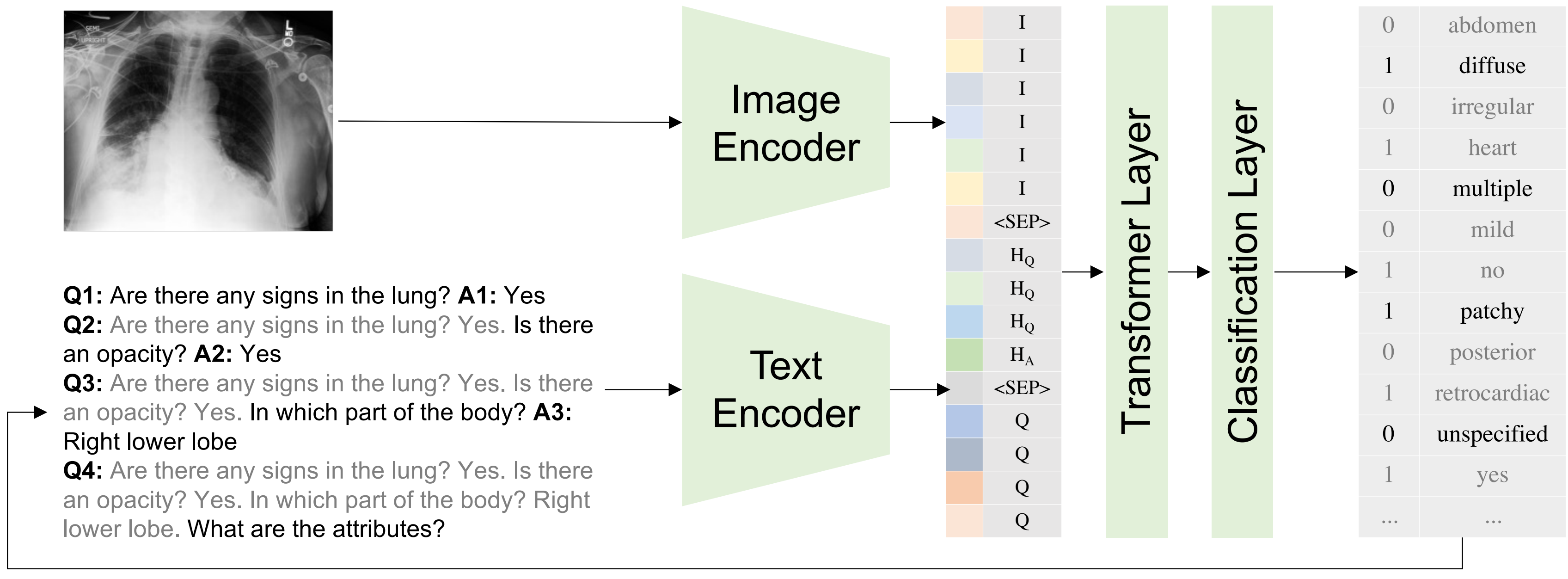
Level 2: Element Existence (example questions)
Are there stents?
Is there an opacity in the lung?
Is there pneumonia in the lung?
Is there anything abnormal in the diaphragm?

Level 3: Attributes
In which part of the body? → heart ventricles, middle lobe, ...
Is which area? → left, anterior, ...
What are the attributes? → patchy, round, ...
What is the degree? → mild, severe, ...

Structured Reporting using Hierarchical VQA

hi-VQA is a flexible automated reporting method for consistent report population. Our method populates the report iteratively from coarse to fine while incorporating the question context by including previous questions and answers:

- Model: We use EfficientNet-b5 as image encoder combined with pre-trained RadBERT [2] for text encoding, then a transformer-based fusion is applied, followed by a linear layer for multi-label classification.
- Answer Selection: We consider only valid answers for each question during both loss computation and prediction.
- Training and Evaluation: teacher forcing on question-answer level during training (gt question context) and autoregressive evaluation asking further questions depending on answer which automatically gives consistent report prediction



Q1: Are there any signs in the lung? **A1:** Yes
Q2: Are there any signs in the lung? Yes. Is there an opacity? **A2:** Yes
Q3: Are there any signs in the lung? Yes. Is there an opacity? Yes. In which part of the body? **A3:** Right lower lobe
Q4: Are there any signs in the lung? Yes. Is there an opacity? Yes. In which part of the body? Right lower lobe. What are the attributes?

Acknowledgement

The authors gratefully acknowledge the financial support by the Federal Ministry of Education and Research of Germany (BMBF) under project DIVA (FKZ 13GW0469C) and the Bavarian Research Foundation (BFS) under project PandeMIC (grant AZ-1429-20C).

Results

Rad-ReStruct

We evaluate hi-VQA on Rad-ReStruct to set a first baseline for our new benchmark. The key findings from our experiments on Rad-ReStruct include:

- Modeling the report generation as VQA task improves over a visual baseline, directly predicting the full classification vector for all questions.
- Leveraging history context enhances accuracy and precision, with a slight dip in recall. This is especially relevant for intricate attribute questions which gain meaning in context.
- Using the radiology-specific RadBERT [2] encoder improves over a general RoBERTa$_{BASE}$ model, highlighting the value of domain-specialized text understanding.

	domain-specific pretraining data	report acc	F1	prec	recall
Visual baseline	none (only general images)	31.3	30.7	65.6	31.2
hi-VQA - no history	radiologic reports	26.2	**31.9**	59.9	**34.1**
hi-VQA - RoBERTa$_{BASE}$	none (only general text/images)	26.2	31.6	67.9	32.4
hi-VQA	radiologic reports	**32.6**	31.7	**70.7**	32.1

A deeper dive into performance metrics across different reporting levels reveals impressive accuracy in identifying sub-topics (objects, diseases, signs, abnormalities), while predicting lower-level attributes proves to be a more difficult task.

	report acc	F1	prec	recall	#paths	avg #answers
Level 1 - Topic Existence	36.6	63.8	79.0	63.5	50	2
Level 2 - Element Existence (all)	33.7	72.2	86.0	72.3	432	2
- Diseases	52.4	74.6	83.7	74.9	206	2
- Signs	74.3	74.1	90.1	74.1	130	2
- Abnormal body regions	58.6	69.1	86.4	69.3	64	2
- Objects	90.4	67.8	87.6	67.1	32	2
Level 3 - Attributes	32.6	3.7	60.3	4.4	1988	4.2

VQARad

VQARad serves as a prominent benchmark in medical VQA, which we utilized to further validate our hi-VQA model:

- Our hi-VQA model, even without historical context, surpasses several existing methods, especially those devoid of domain-specific joint image-text pretraining.
- With the inclusion of historical data, hi-VQA performs competitively with leading methods, underscoring the benefit of jointly addressing queries for a single image.
- Again, the domain-specific text encoder proves to be crucial for understanding the questions better.

Our experiments reinforce the efficacy of hi-VQA in automated radiology reporting, spotlighting the benefits of historical context and domain-specific text encoding.

	domain-specific pretraining data	acc
MEVF	radiologic images	66.1
MMQ	none	67.0
MM-BERT	radiologic images and reports (joined PT)	72.0
CRPD	radiologic images	72.7
RepsNet	radiologic reports	73.5
M3AE	radiologic images and reports (joined PT)	77.0
hi-VQA - no history	radiologic reports	74.5
hi-VQA - RoBERTa$_{BASE}$	none (only general text/ images)	72.5
hi-VQA	radiologic reports	76.3

Conclusion

The introduction of Rad-ReStruct provides a much needed first benchmark for structured radiology reporting, emphasizing varied levels of clinical accuracy. Our hi-VQA model, with its history-centric approach, mirrors the workflow of structured reporting, offering interpretability and potential real-time adaptability for radiologists. Despite good performance on high-level questions, challenges remain at the attribute-specific levels. Nonetheless, the performance discrepancy between VQARad and Rad-ReStruct underscores the challenges of fine-grained structured reporting. We believe this work encourages the community to tackle these challenges.

References

[1] Dina Demner-Fushman et al. "Preparing a collection of radiology examinations for distribution and retrieval". In: *Journal of the American Medical Informatics Association* 23.2 (2016), pp. 304–310.

[2] An Yan et al. "RadBERT: Adapting transformer-based language models to radiology". In: *Radiology: Artificial Intelligence* 4.4 (2022), e210258.

Code: