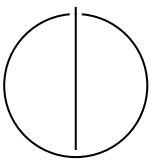
SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY — INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics

Vision-Language Foundation Models for Multi-Modal Medical Data

Muhammed Göktepe



SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY — INFORMATICS

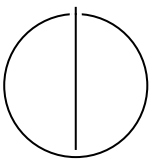
TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics

Vision-Language Foundation Models for Multi-Modal Medical Data

Vision-Language-Foundation-Modelle für Multimodale Medizinische Daten

Author:	Muhammed Göktepe
Examiner:	Prof. Dr. Christian Wachinger
Supervisor:	Fabian Bongratz
Submission Date:	16.03.2026



I confirm that this master's thesis is my own work and I have documented all sources and material used.

Munich, 16.03.2026 Muhammed Göktepe

Acknowledgments

Abstract

Contents

1 Introduction

1.1 Section

Citation test [**latex**].

Acronyms must be added in `main.tex` and are referenced using macros. The first occurrence is automatically replaced with the long version of the acronym, while all subsequent usages use the abbreviation.

E.g. `\ac{TUM}, \ac{TUM}` ⇒ Technical University of Munich (TUM), TUM

For more details, see the documentation of the `acronym` package[1].

1.1.1 Subsection

See Table 1.1, Figure 1.1, Figure 1.2, Figure 1.3.

Table 1.1: An example for a simple table.

A	B	C	D
1	2	1	2
2	3	2	3

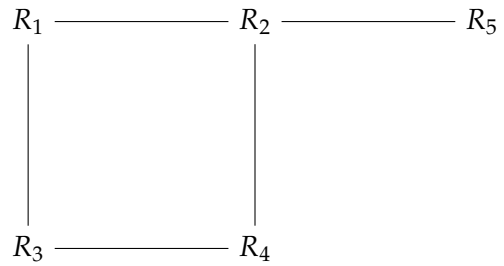


Figure 1.1: An example for a simple drawing.

[1] `https://ctan.org/pkg/acronym`

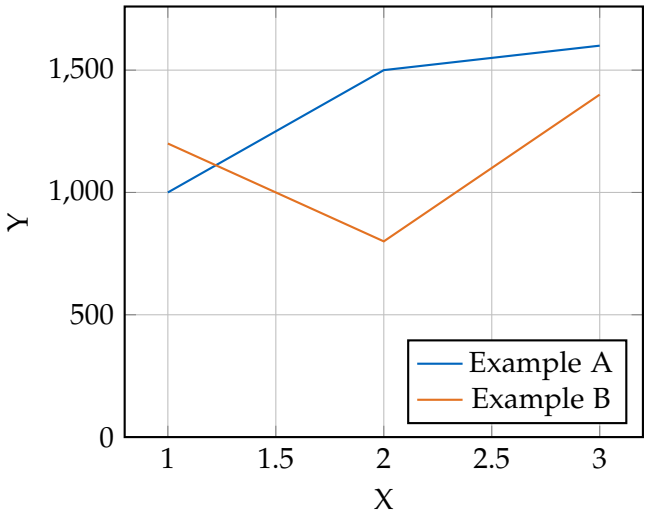


Figure 1.2: An example for a simple plot.

```
SELECT * FROM tbl WHERE tbl.str = "str"
```

Figure 1.3: An example for a source code listing.

2 Related Work

2.1 Section

3 Methodology

3.1 LLM-based Report Decomposition

To enable fine-grained, organ-specific radiology report generation, we implement a preprocessing pipeline that decomposes full free-text radiology reports into structured, organ-specific findings. This is achieved using a Large Language Model (LLM) steered by a predefined anatomical schema.

3.1.1 Anatomical Schema Definition

We define a comprehensive schema covering major anatomical regions in chest CT scans. The schema targets the following structures:

- **Thorax**: Lung, Heart, Aorta, Trachea, Esophagus, Portal Vein, Rib.
- **Abdomen**: Liver, Gallbladder, Stomach, Pancreas, Spleen, Kidney, Adrenal, Colon, Small Bowel, Bladder.

This schema ensures that the model learns to associate specific image regions with their corresponding textual descriptions, rather than generating a monolithic report.

3.1.2 Decomposition Pipeline

We utilize the **Llama-3.3-70B-Instruct** model for the decomposition task. Due to the significant memory requirements of this 70-billion parameter model, we employ the **AWQ (Activation-aware Weight Quantization)** version [1] to fit the model within the memory constraints of our GPU infrastructure (e.g., 4×A100). The model is served using **vLLM** [2], a high-throughput and memory-efficient LLM serving engine.

The decomposition process involves the following steps:

1. **Input**: The raw "Findings" and "Impression" sections of the radiology report are concatenated.
2. **Prompting**: The LLM is prompted with a system instruction to act as a radiologist and extract organ-specific information according to our defined JSON schema.

The prompt explicitly requests valid JSON output and handles missing organs by assigning them a `null` value.

3. **Output Parsing**: The generated output is parsed to extract the JSON structure, separating findings and impressions for each organ.

The implementation uses a robust parsing mechanism to handle potential formatting irregularities in the LLM's output, ensuring high data quality for downstream training.

4 Experiments and Results

4.1 Dataset

We validate our methods using the **CT-RATE** dataset [3], a large-scale multimodal dataset for 3D computed tomography. CT-RATE is the first large-scale publicly available dataset that pairs full 3D medical volumes with free-text radiology reports, enabling the training of end-to-end 3D vision–language models for tasks such as report generation and text-conditional image synthesis [4].

4.1.1 Dataset Overview

The dataset consists of 25,692 3D chest CT volumes from 21,304 unique patients, paired with their corresponding free-text radiology reports. The data was collected from multiple sources and anonymized to ensure privacy. Key statistics of the dataset are summarized in Table 4.1 [3].

Table 4.1: Key statistics of the CT-RATE dataset used in our experiments.

Statistic	**Value**
Total CT Volumes	25,692
Unique Patients	21,304
Modality	3D Chest CT
Text Modality	Radiology Reports (Findings & Impression)

4.1.2 Preprocessing

Since CT volumes in the CT-RATE dataset vary significantly in all dimensions, we applied a standardized preprocessing pipeline to ensure a stable training input for our models. The steps, implemented using the MONAI library, are as follows:

1. **Intensity Scaling**: We apply distinct intensity scaling to normalize Hounsfield Units (HU) to the range $[0, 1]$. We clip intensities between -1150 and 350 HU,

which covers the relevant radiodensity range for lung and soft tissue structures while suppressing extreme outliers.

2. **Spatial Normalization**: To handle the variable volume sizes, we standardize the spatial dimensions to a fixed size of $112 \times 256 \times 352$ voxels. This is achieved by first applying *Spatial Padding* with constant zero values to ensure the volume is at least the target size.

3. **Dimension Reordering**: We ensure a channel-first data layout and transpose dimensions to match the input expectations of our 3D vision encoder.

- **Text Preprocessing**: To enable organ-level report generation, we employed an LLM-based decomposition strategy. We used a Large Language Model to decompose each full radiology report into organ-specific descriptions (e.g., "Lung", "Heart", "Liver"). This allows our model to learn fine-grained section-specific representations. We specifically filter for and utilize these decomposed sections, tokenizing them with a maximum sequence length of 128 tokens for our decoder.

Recent work has shown that using improved tokenization strategies can further enhance performance on this dataset [5]. We utilize the official train/validation/test splits provided by the authors to ensure comparable results with prior work.

4.2 Evaluation Metrics

To assess the performance of our radiology report generation system we evaluate generated reports with multiple complementary metrics. Metrics are grouped by the type of signal they capture:

1. **Surface-level (linguistic) metrics:** measure n-gram overlap and surface similarity to reference reports.

2. **Semantic / embedding-based metrics:** measure semantic similarity using contextual embeddings.

3. **Clinical / structured metrics:** evaluate clinical correctness by extracting findings, entities, relations or clinically-relevant labels.

4. **Human evaluation and statistical tests:** measure radiologist preference, clinical usefulness, and statistical significance.

4.2.1 Surface-level (Natural Language Generation) Metrics

These metrics are commonly used in image captioning and report generation benchmarks and capture lexical and syntactic similarity.

BLEU

BLEU [6] is an n-gram precision-based metric. For a candidate report c and a set of references $\{r\}$, BLEU is computed as

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \tag{4.1}$$

where p_n is the modified precision for n-grams, w_n are weights (we use uniform weights $w_n = 1/N$), and BP is the brevity penalty:

$$\text{BP} = \begin{cases} 1 & \text{if } c_{len} > r^*_{len} \\ \exp\left(1 - \frac{r^*_{len}}{c_{len}}\right) & \text{otherwise} \end{cases} \tag{4.2}$$

Here c_{len} is the candidate length and r^*_{len} is the effective reference length (e.g., closest reference length). We report BLEU-4.

ROUGE-L

ROUGE-L [7] uses the length of the Longest Common Subsequence (LCS) to capture sentence-level structure. Define

$$R_{LCS} = \frac{\text{LCS}(c, r)}{|r|}, \qquad P_{LCS} = \frac{\text{LCS}(c, r)}{|c|},$$

then the ROUGE-L F-measure is

$$F_{LCS} = \frac{(1 + \beta^2)\, R_{LCS}\, P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}}, \tag{4.3}$$

where β weights recall relative to precision. In practice we set $\beta = 1$ to report the balanced F_1 unless otherwise specified.

METEOR

METEOR [8] aligns unigrams and computes a weighted harmonic mean of precision and recall that emphasizes recall. The core F_{mean} is:

$$F_{\text{mean}} = \frac{(1 + \alpha)PR}{R + \alpha P}, \tag{4.4}$$

where typical METEOR uses $\alpha = 9$ (equivalently $F_{\text{mean}} = \frac{10PR}{R+9P}$). METEOR applies a fragmentation penalty:

$$\text{Penalty} = 0.5 \left(\frac{\#\text{chunks}}{\#\text{matches}} \right)^3, \tag{4.5}$$

and final score:

$$\text{METEOR} = F_{\text{mean}} \cdot (1 - \text{Penalty}). \tag{4.6}$$

CIDEr

CIDEr [9] uses TF–IDF weighting for n-grams and computes average cosine similarity between a candidate's and references' TF–IDF vectors. For n-grams of length n:

$$\text{CIDEr}_n(c, r) = \frac{1}{M} \sum_{i=1}^{M} \frac{\mathbf{g}^n(c) \cdot \mathbf{g}^n(r_i)}{\|\mathbf{g}^n(c)\| \, \|\mathbf{g}^n(r_i)\|}, \tag{4.7}$$

and final CIDEr is the (optionally IDF-normalized) average across $n = 1 \ldots 4$.

Notes: Surface metrics are sensitive to wording and may penalize clinically-correct paraphrases. Use them together with semantic and clinical metrics.

4.2.2 Semantic / Embedding-based Metrics

These metrics capture deeper (contextual) semantic similarity beyond n-gram overlap.

BERTScore

BERTScore computes pairwise token similarities using contextual token embeddings from a pre-trained transformer (e.g., BERT). For tokens in candidate and reference, token-wise cosine similarities are computed and aggregated into precision, recall and an F_1 score; optional IDF weighting can be applied [10].

BERT-based contextual similarity

Besides BERTScore, one can compute sentence-level embeddings (CLS / pooled embeddings) and cosine similarity, or use more specialized biomedical encoders (e.g., BioBERT, ClinicalBERT) to increase domain alignment.

4.2.3 Clinical / Structured Metrics

These metrics aim to measure clinical correctness and the presence/absence of findings.

CheXbert

CheXbert [11] is a BERT-based labeler trained on the output of the CheXpert rule-based labeler to extract findings from radiology reports with higher accuracy. We use it to detect the presence of 14 standard observations (e.g., Cardiomegaly, Edema, Pneumonia, No Finding). We compute the Micro-F_1 score across all 14 classes to measure the clinical accuracy of the generated reports compared to the ground truth.

RadGraph

RadGraph [12] converts radiology reports into knowledge graphs (entities and relations). Evaluation is done by measuring overlap between predicted and reference graphs; typical metrics are node (entity) Precision/Recall/F_1 and edge (relation) Precision/Recall/F_1. Reporting both entity-level and relation-level scores helps diagnose whether models miss or hallucinate links between findings and anatomy.

SRR-BERT / Structured-report metrics

SRR-BERT and similar methods fine-tune BERT-like encoders to score structured-report correctness; they are useful for fine-grained disease classification and structured-report components. Report the metric(s) as provided by the original implementation (precision/recall/F_1 or regression correlation).

GREEN and other factuality metrics

GREEN (Generative Radiology report Evaluation and Error Notation) and related metrics are designed to highlight factual errors and categorize them by severity (critical vs non-critical). Use them to count clinically-significant hallucinations and factual inconsistencies.

Composite clinical metrics (e.g., RadCliQ)

Composite metrics such as RadCliQ combine surface (e.g., BLEU) and clinical (e.g., CheXpert) signals, often with learned weights to better correlate with radiologist judgments. When using a learned composite, report how weights were obtained and validate correlation with human raters.

4.3 Results

4.3.1 Quantitative Results

Content goes here.

4.3.2 Qualitative Results

Content goes here.

5 Discussion

5.1 Section

Content goes here.

5.1.1 Subsection

More content goes here.

6 Conclusion and Future Work

6.1 Section

Content goes here.

6.1.1 Subsection

More content goes here.

6.2 Future Work

asdasdhashdas

Abbreviations

TUM Technical University of Munich

List of Figures

List of Tables

Bibliography

[1] J. Lin, J. Tang, H. Tang, S. Yang, X. Dang, and C. Gan. "AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration." In: *Proceedings of the 6th MLSys Conference*. 2023.

[2] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. Ragan-Kelley, J. E. Gonzalez, and I. Stoica. "Efficient Memory Management for Large Language Model Serving with PagedAttention." In: *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*. 2023.

[3] I. E. Hamamci, S. Er, F. Almas, A. G. Simsek, S. N. Esirgun, I. Dogan, M. F. Dasdelen, O. F. Durugol, B. Wittmann, T. Amiranashvili, E. Simsar, M. Simsar, E. B. Erdemir, A. Alanbay, A. Sekuboyina, B. Lafci, C. Bluethgen, M. K. Ozdemir, and B. Menze. *Generalist Foundation Models from a Multimodal Dataset for 3D Computed Tomography*. 2024. arXiv: 2403.17834 [cs.CV].

[4] I. E. Hamamci, S. Er, A. Sekuboyina, E. Simsar, A. Tezcan, A. G. Simsek, S. N. Esirgun, F. Almas, I. Doğan, M. F. Dasdelen, et al. "Generatect: Text-conditional generation of 3d chest ct volumes." In: *European Conference on Computer Vision*. Springer. 2024, pp. 126–143.

[5] I. E. Hamamci, S. Er, S. Shit, H. Reynaud, D. Yang, P. Guo, M. Edgar, D. Xu, B. Kainz, and B. Menze. *Better Tokens for Better 3D: Advancing Vision-Language Modeling in 3D Medical Imaging*. 2025. arXiv: 2510.20639 [cs.CV].

[6] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. "BLEU: a Method for Automatic Evaluation of Machine Translation." In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, PA, USA: Association for Computational Linguistics, 2002, pp. 311–318. DOI: 10.3115/1073083.1073135.

[7] C.-Y. Lin. "ROUGE: A Package for Automatic Evaluation of Summaries." In: *Text Summarization Branches Out (ACL Workshop)*. Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 74–81.

[8] S. Banerjee and A. 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. Ann Arbor, MI, USA: Association for Computational Linguistics, 2005, pp. 65–72.

[9] R. Vedantam, C. L. Zitnick, and D. Parikh. "CIDEr: Consensus-based Image Description Evaluation." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. arXiv:1411.5726. 2015, pp. 4566–4575.

[10] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. "BERTScore: Evaluating Text Generation with BERT." In: *arXiv preprint arXiv:1904.09675* (2019).

[11] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, and M. P. Lungren. *CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT*. 2020. arXiv: `2004.09167 [cs.CL]`.

[12] S. Jain, A. Agrawal, A. Saporta, S. Q. Truong, D. N. Duong, T. Bui, P. Chambon, Y. Zhang, M. P. Lungren, C. P. Langlotz, and P. Rajpurkar. "RadGraph: Extracting Clinical Entities and Relations from Radiology Reports." In: *arXiv preprint arXiv:2106.14463* (2021).