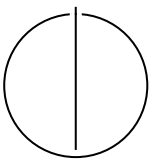
SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY — INFORMATICS

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

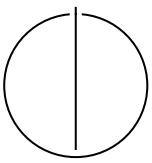
TECHNISCHE UNIVERSITÄT MÜNCHEN

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

As deep learning has continued to advance over recent years, it has been applied to various fields of science and engineering. One of the most prominent applications of deep learning is in the field of computer vision. Deep learning models have achieved state-of-the-art performance on various computer vision tasks, such as image classification, object detection, and image segmentation. However, deep learning models are data-hungry and require large amounts of labeled data to achieve high performance. This is especially true for medical imaging, where obtaining large-scale labeled datasets is challenging due to privacy concerns and the high cost of expert annotation. Traditional CNN-RNN-based frameworks suffered from a lack of localized reasoning and from weak integration of medical knowledge. This is partially solved by the introduction of the attention mechanism by transformers.

Despite these improvements, several limitations still remain. Attention mechanisms improve the alignment between image regions and text tokens. However, most existing methods rely on 2D image representations. They process single images or independent slices from volumetric scans. As a result, they fail to capture full three-dimensional context. In clinical practice, many important findings depend on volumetric information, such as lesion shape, spatial extent, and inter-slice continuity. These details are often lost when using only 2D inputs. Therefore, models trained on 2D data struggle to perform consistent localized reasoning across multiple slices.

There are several reasons why most studies focus on 2D approaches. First, large-scale 3D medical datasets are difficult to collect and share. Privacy regulations and storage requirements make this process challenging. Second, training 3D deep learning models requires significantly higher computational resources. This limits their practical adoption. Third, most existing pretrained vision and vision–language models are designed for 2D data. As a result, transfer learning for 3D medical imaging is less mature. These factors have led researchers to favor simpler 2D pipelines, even though they do not fully represent clinical imaging data.

Recent advances in vision–language models (VLMs) offer new opportunities to overcome these limitations. VLMs enable stronger alignment between visual features and natural language. They support better reasoning across modalities. However, most medical VLM studies still rely on 2D image inputs. The potential of volumetric reasoning in combination with large language models remains underexplored.

This thesis aims to address this gap by focusing on 3D vision–language modeling for radiology report generation. A volumetric vision encoder is used to preserve spatial relationships across slices. This allows the model to better capture anatomical structure and disease patterns. In addition, state-of-the-art large language models are used to improve medical text understanding and generation. These models help produce clearer, more structured, and clinically accurate reports. By combining 3D visual reasoning with advanced medical language understanding, this work aims to build a more reliable and clinically meaningful report-generation system.

2 Related Work

This chapter reviews the key areas of prior work that form the foundation for this thesis. Section 2.1 discusses the Vision Transformer architecture and its adaptation to medical imaging. Section 2.2 covers vision–language models, including both general-purpose and medical variants. Section 2.3 introduces large language models and parameter-efficient fine-tuning techniques. Section 2.4 surveys radiology report generation methods. Finally, Section 2.5 covers 3D medical image segmentation and organ-guided attention mechanisms.

2.1 Vision Transformers in Medical Imaging

The Vision Transformer (ViT) [1] introduced the idea of applying the transformer architecture, originally designed for natural language processing, to image recognition. ViT divides an input image into fixed-size non-overlapping patches, projects each patch into an embedding, and processes the resulting sequence through a standard transformer encoder. By treating image patches as tokens, ViT eliminates the need for convolutional inductive biases and achieves competitive or superior performance compared to CNNs on large-scale benchmarks such as ImageNet.

A key limitation of ViT is its reliance on large-scale supervised pretraining. To address this, He et al. [2] proposed Masked Autoencoders (MAE), a self-supervised pretraining method for ViT. MAE randomly masks a large proportion (e.g., 75%) of image patches during training and learns to reconstruct the missing pixels. This approach enables learning strong visual representations from unlabeled data, which is particularly valuable in the medical domain where labeled data is scarce. MAE-pretrained ViTs have shown strong transfer learning performance across a variety of downstream tasks.

In the medical imaging domain, transformers have been widely adopted. TransUNet [3] combines a CNN encoder with a transformer for medical image segmentation, showing that transformers can capture long-range dependencies that are crucial for understanding anatomical structures. UNETR [4] extends this approach to 3D medical image segmentation by using a ViT encoder to process volumetric data directly, avoiding the information loss associated with 2D slice-by-slice processing.

The core vision encoder used in this thesis is a 3D ViT initialized with MAE-pretrained weights and adapted from the BLIP framework [5], which is available through the LAVIS library [6]. Unlike standard 2D ViTs, this encoder processes volumetric CT inputs with 3D patch embeddings, preserving spatial relationships across all three dimensions.

2.2 Vision–Language Models

Vision–language models (VLMs) aim to bridge the gap between visual perception and natural language understanding. These models learn joint representations of images and text, enabling tasks such as image captioning, visual question answering, and cross-modal retrieval.

2.2.1 Contrastive Learning: CLIP

CLIP (Contrastive Language–Image Pre-training) [7] introduced a scalable approach to learning visual representations from natural language supervision. By training a vision encoder and a text encoder to maximize the similarity of matching image–text pairs while minimizing the similarity of non-matching pairs, CLIP learns transferable visual concepts. Trained on 400 million image–text pairs from the internet, CLIP demonstrates strong zero-shot transfer to a wide range of downstream visual tasks. Its contrastive learning paradigm has become a foundational building block for many subsequent VLMs.

2.2.2 Generative VLMs

While CLIP focuses on alignment, generative VLMs extend the paradigm to produce natural language outputs conditioned on visual inputs. BLIP [5] proposes a unified framework that jointly trains a contrastive alignment objective, an image–text matching objective, and a generative language modeling objective. This multi-task design enables BLIP to be effective for both understanding and generation tasks. The LAVIS library [6] provides a standardized implementation of BLIP and related models.

Flamingo [8] takes a different approach by interleaving visual tokens with text tokens in a frozen large language model. It introduces gated cross-attention layers to condition the LLM on visual features, demonstrating strong few-shot learning capabilities across diverse vision–language tasks.

LLaVA (Large Language-and-Vision Assistant) [9] simplifies the VLM architecture by using a linear projection layer to map visual features from a pretrained CLIP encoder into the embedding space of a large language model. Despite its architectural simplicity,

LLaVA achieves competitive performance through visual instruction tuning, where the model is trained on instruction-following data that combines images with text instructions.

2.2.3 Medical VLMs

Adapting VLMs to the medical domain presents unique challenges. Medical images have different characteristics from natural images, including higher resolution, domain-specific features, and the need for fine-grained understanding. BiomedCLIP [10] addresses this by pretraining a CLIP-style model on 15 million biomedical image–text pairs from PubMed, demonstrating that domain-specific pretraining significantly improves performance on biomedical tasks.

For 3D medical imaging, Hamamci et al. [11] introduced CT-CLIP, a contrastive learning framework for 3D computed tomography. Trained on the CT-RATE dataset, which pairs volumetric CT scans with radiology reports, CT-CLIP learns 3D visual representations aligned with clinical text. This work represents one of the first large-scale efforts to apply vision–language pretraining to volumetric medical data.

MedGemma [12] is a family of medical foundation models built on top of Google's Gemma architecture [13]. MedGemma extends the general-purpose Gemma language model with medical knowledge through continued pretraining on biomedical literature and clinical data. The instruction-tuned variant (MedGemma-4B-IT) demonstrates strong performance on medical question answering and report understanding tasks, making it well-suited as a language decoder for medical VLM architectures.

2.3 Large Language Models and Parameter-Efficient Fine-Tuning

Large language models (LLMs) such as the LLaMA family [14] and Gemma [13] have demonstrated remarkable capabilities in natural language understanding and generation. These models, typically containing billions of parameters, are pretrained on massive text corpora and can be adapted to specific downstream tasks through fine-tuning.

However, full fine-tuning of LLMs is computationally prohibitive for most research settings. Parameter-efficient fine-tuning (PEFT) methods address this challenge by updating only a small subset of model parameters while keeping the majority frozen.

2.3.1 LoRA

LoRA (Low-Rank Adaptation) [15] is one of the most widely adopted PEFT methods. LoRA introduces trainable low-rank decomposition matrices into the attention layers of a pretrained model. Instead of updating the full weight matrix $W \in \mathbb{R}^{d \times k}$, LoRA learns two smaller matrices $A \in \mathbb{R}^{d \times r}$ and $B \in \mathbb{R}^{r \times k}$ where $r \ll \min(d, k)$, such that the adapted weight becomes $W + AB$. This reduces the number of trainable parameters by orders of magnitude while maintaining competitive performance.

2.3.2 QLoRA and Quantization

QLoRA [16] further reduces the memory requirements by combining LoRA with 4-bit quantization. The pretrained model weights are quantized to 4-bit Normal Float (NF4) precision using the BitsAndBytes library, while the LoRA adapter weights are maintained in higher precision (e.g., BFloat16) for training. This enables fine-tuning of models with billions of parameters on consumer-grade hardware. A frozen, 4-bit quantized LLM decoder combined with trainable adapter layers or projectors has become a common paradigm for building resource-efficient VLMs.

2.4 Radiology Report Generation

Automatic radiology report generation is the task of producing clinically accurate free-text reports from medical images. This task is more challenging than generic image captioning because generated reports must be factually accurate, cover all relevant findings, and follow clinical conventions.

2.4.1 Traditional Approaches

Early approaches to radiology report generation adapted encoder–decoder architectures from image captioning. Jing et al. [17] proposed a co-attention mechanism combined with a hierarchical LSTM decoder to generate structured medical reports from chest X-rays. However, these CNN–RNN frameworks typically produce generic descriptions and struggle with capturing rare or subtle findings.

2.4.2 Transformer-Based Approaches

Transformer-based methods have significantly improved report generation quality. Chen et al. [18] introduced a memory-driven transformer that uses a relational memory module to record key information during generation, improving the coherence of long reports. In subsequent work, Chen et al. [19] proposed cross-modal memory networks

that maintain separate memory banks for visual and textual features, enabling better cross-modal reasoning.

Nicolson et al. [20] demonstrated that warm-starting the encoder and decoder from pretrained models substantially improves chest X-ray report generation, highlighting the importance of leveraging existing pretrained representations.

2.4.3 LLM-Based Approaches

More recent approaches leverage frozen large language models as report decoders. R2GenGPT [21] demonstrated that a frozen LLM combined with a trainable visual encoder and a lightweight projection layer can generate high-quality radiology reports. This paradigm benefits from the strong language understanding capabilities of pretrained LLMs while keeping training costs manageable.

The approach adopted in this thesis follows a similar paradigm: a trainable 3D vision encoder produces visual embeddings that are projected into the embedding space of a frozen MedGemma decoder via a linear projection layer, following the design principle introduced by LLaVA [9].

2.5 3D Medical Image Segmentation and Organ-Guided Attention

Accurate localization of anatomical structures is essential for generating organ-specific radiology reports. The availability of automated whole-body segmentation tools has enabled new approaches that leverage anatomical priors for medical image understanding.

2.5.1 TotalSegmentator

TotalSegmentator [22] is a widely used tool for automatic segmentation of 104 anatomical structures in CT images. Built on the nnU-Net framework, TotalSegmentator provides robust, out-of-the-box segmentation of organs, bones, vessels, and muscles. The segmentation masks produced by TotalSegmentator can serve as anatomical priors for downstream tasks, including region-of-interest (ROI) guided attention mechanisms.

2.5.2 Organ-Guided Attention

Incorporating anatomical localization into VLM architectures enables organ-specific reasoning. Rather than processing the entire image as a flat sequence of tokens, organ-guided attention mechanisms restrict the model's attention to regions corresponding

to specific anatomical structures. This approach draws upon the concept of learnable queries from architectures such as the Perceiver, where a fixed set of learned query vectors attend to a variable-length input through cross-attention.

In the context of multi-task learning [23], organ-specific feature extraction can be further enhanced by auxiliary classification objectives. By jointly training the visual encoder to perform disease classification alongside report generation, the model learns more discriminative organ-level representations. This auxiliary loss acts as an inductive bias that encourages the visual encoder to capture clinically relevant features.

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 [24] to fit the model within the memory constraints of our GPU infrastructure (e.g., 4×A100). The model is served using **vLLM** [25], 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 [11], 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 [26].

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 [11].

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 [27]. 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 [28] 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 [29] 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 [30] 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 [31] 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 $n = 1 \dots 4$.

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 [32].

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 [33] 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 [34] 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

List of Figures

List of Tables

Bibliography

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." In: *International Conference on Learning Representations (ICLR)*. 2021.

[2] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. "Masked Autoencoders Are Scalable Vision Learners." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 16000–16009.

[3] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou. "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation." In: *arXiv preprint arXiv:2102.04306* (2021).

[4] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu. "UNETR: Transformers for 3D Medical Image Segmentation." In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2022, pp. 574–584.

[5] J. Li, D. Li, C. Xiong, and S. Hoi. "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation." In: *International Conference on Machine Learning (ICML)*. PMLR, 2022, pp. 12888–12900.

[6] D. Li, J. Li, H. Le, G. Wang, S. Savarese, and S. C. Hoi. "LAVIS: A One-stop Library for Language-Vision Intelligence." In: *arXiv preprint arXiv:2209.09019* (2023).

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. "Learning Transferable Visual Models from Natural Language Supervision." In: *International Conference on Machine Learning (ICML)*. PMLR, 2021, pp. 8748–8763.

[8] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. "Flamingo: A Visual Language Model for Few-Shot Learning." In: *Advances in Neural Information Processing Systems (NeurIPS)* 35 (2022), pp. 23716–23736.

[9] H. Liu, C. Li, Q. Wu, and Y. J. Lee. "Visual Instruction Tuning." In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 36. 2024.

[10] S. Zhang, Y. Xu, N. Usuyama, J. Bagber, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Teodoro, et al. "BiomedCLIP: A Multimodal Biomedical Foundation Model Pretrained from Fifteen Million Scientific Image-Text Pairs." In: *arXiv preprint arXiv:2303.00915* (2023).

[11] 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]`.

[12] G. Yang. "MedGemma: Medical Gemma." In: *Google AI Blog* (2024).

[13] Gemma Team, Google DeepMind. "Gemma: Open Models Based on Gemini Research and Technology." In: *arXiv preprint arXiv:2403.08295* (2024).

[14] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. "LLaMA: Open and Efficient Foundation Language Models." In: *arXiv preprint arXiv:2302.13971* (2023).

[15] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. "LoRA: Low-Rank Adaptation of Large Language Models." In: *arXiv preprint arXiv:2106.09685* (2022).

[16] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. "QLoRA: Efficient Finetuning of Quantized Large Language Models." In: *Advances in Neural Information Processing Systems (NeurIPS)* 36 (2023).

[17] B. Jing, P. Xie, and E. Xing. "On the Automatic Generation of Medical Imaging Reports." In: *arXiv preprint arXiv:1711.08195* (2018).

[18] Z. Chen, Y. Song, T.-H. Chang, and X. Wan. "Generating Radiology Reports via Memory-Driven Transformer." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020, pp. 1439–1449.

[19] Z. Chen, Y. Shen, Y. Song, and X. Wan. "Cross-modal Memory Networks for Radiology Report Generation." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*. 2022.

[20] A. Nicolson, J. Dowling, and B. Kober. "Improving Chest X-Ray Report Generation by Leveraging Warm Starting." In: *Artificial Intelligence in Medicine* 144 (2023), p. 102633.

[21] Z. Wang, L. Liu, L. Wang, and L. Zhou. "R2GenGPT: Radiology Report Generation with Frozen LLMs." In: *Meta-Radiology* (2023).

[22] J. Wasserthal, H.-C. Breit, M. T. Meyer, M. Pradella, D. Hinck, A. W. Sauter, T. Heye, D. T. Boll, J. Defined, et al. "TotalSegmentator: Robust segmentation of 104 anatomic structures in CT images." In: *Radiology: Artificial Intelligence* 5.5 (2023).

[23] R. Caruana. "Multitask Learning." In: *Machine Learning* 28.1 (1997), pp. 41–75. DOI: 10.1023/A:1007379606734.

[24] 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.

[25] 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.

[26] 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.

[27] 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].

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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).

[33] 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]`.

[34] 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).