Zürcher Hochschule für Angewandte Wissenschaften

Departement School of Engineering

Centre for Artificial Intelligence

BACHELORARBEIT

Kompakte Multimodale Sprachmodelle in der Pathologie-VQA: Leistungsvergleich und Erklärbarkeitsansätze

Autoren:
Arbnor Ziberi, Luka Sokcevic

Betreuerin:
Prof. Dr. Jasmina Bogojeska

Eingereicht am:
6. Juni 2025

Studiengang:
Data Science B.Sc.

Imprint

Projekt: Bachelorarbeit
Titel: Kompakte Multimodale Sprachmodelle in der Pathologie-VQA: Leistungsvergleich und Erklärbarkeitsansätze
Autoren: Arbnor Ziberi, Luka Sokcevic
Datum: 6. Juni 2025
Keywords: multimodal LLMs, pathology, visual question answering, fine-tuning, model interpretability

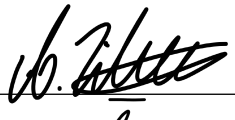
Studiengang:
Data Science B.Sc.
Zürcher Hochschule für Angewandte Wissenschaften

Betreuerin:
Prof. Dr. Jasmina Bogojeska
Zurich University of Applied Sciences
Email: jasmina.bogojeska@zhaw.ch
Web: Link

Erklärung der Eigenständigkeit

Wir, Arbnor Ziberi und Luka Sokcevic, erklären, dass diese Arbeit und die darin dargestellte Leistung von uns selbst stammen. Wir bestätigen, dass:

- Diese Arbeit ganz oder hauptsächlich während unserer Kandidatur für einen Forschungsgrad an der Zürcher Hochschule für Angewandte Wissenschaften entstanden ist.
- Falls Teile dieser Arbeit bereits für einen Abschluss oder eine andere Qualifikation an dieser Hochschule oder einer anderen Institution eingereicht wurden, wurde dies deutlich angegeben.
- Wenn wir auf veröffentlichte Arbeiten anderer zurückgegriffen haben, wurde dies stets eindeutig gekennzeichnet.
- Wenn wir aus Werken anderer zitiert haben, ist die Quelle immer angegeben. Abgesehen von solchen Zitaten ist diese Arbeit vollständig von uns verfasst.
- Wir haben alle wesentlichen Hilfsquellen angegeben.
- Wenn die Arbeit auf gemeinsam mit anderen durchgeführten Arbeiten basiert, haben wir genau angegeben, was von anderen geleistet wurde und welchen Beitrag wir selbst erbracht haben.
- Wir geben an, dass bestimmte Teile des Textes und des Codes mithilfe des Sprachmodells ChatGPT von OpenAI erstellt oder unterstützt wurden. Konkret wurde ChatGPT verwendet, um Formulierungen zu verfeinern, Umformulierungen vorzuschlagen und bei der Erstfassung einzelner Abschnitte sowie bei der Generierung bestimmter Code-Passagen, der Fehlersuche und der Strukturierung des Codes zu helfen.

Unterschrift: 

Unterschrift: 

Datum: 06.06.2025

Zusammenfassung

Die rasante Entwicklung multimodaler Large Language Models (MLLMs) eröffnet neue Möglichkeiten für automatisierte Assistenzsysteme im medizinischen Bereich. Diese Bachelorarbeit untersucht, inwiefern kompakte MLLMs wie Gemma 3 4B und Qwen 2.5VL 3B für Visual Question Answering (VQA) in der Pathologie eingesetzt werden können. Dabei werden die beiden modernen Modelle anhand des PathVQA-Datensatzes im Vergleich zu einem klassischen ResNet50-BioBERT-Ansatz bewertet. Der Fokus liegt sowohl auf der Leistungsfähigkeit in Zero-Shot- und Few-Shot-Szenarien als auch auf den Verbesserungen durch gezieltes Fine-Tuning. Zudem wird die Nachvollziehbarkeit der Modellentscheidungen mithilfe von Explainable-AI-Methoden wie Integrated Gradients und Self-Attention-Analyse untersucht. Die Resultate zeigen, dass kompakte MLLMs insbesondere nach dem Fine-Tuning die klassischen Ansätze sowohl in Bezug auf Genauigkeit als auch auf Flexibilität übertreffen. So erreichte Gemma 3 4B nach dem Fine-Tuning auf dem Testdatensatz einen ROUGE-L Wert von 0,5975. Erklärbare Methoden tragen dazu bei, die Entscheidungsprozesse der Modelle transparent und nachvollziehbar zu gestalten, bieten jedoch bislang nur punktuelle Einblicke und ermöglichen noch keine vollständige Transparenz. Damit leistet diese Arbeit einen wichtigen Beitrag zur Einschätzung des Potenzials und der Limitationen kompakter multimodaler Sprachmodelle im klinischen Einsatz der Pathologie.

Abstract

The rapid development of multimodal Large Language Models (MLLMs) is opening up new possibilities for automated assistance systems in the medical field. This bachelor's thesis examines to what extent compact MLLMs such as Gemma 3 4B and Qwen 2.5VL 3B can be used for Visual Question Answering (VQA) in pathology. The two modern models are evaluated using the PathVQA dataset and compared to a classical ResNet50-BioBERT approach. The focus is both on performance in zero-shot and few-shot scenarios and on improvements achieved through targeted fine-tuning. Furthermore, the interpretability of model decisions is investigated using Explainable AI methods such as Integrated Gradients and self-attention analysis. The results show that compact MLLMs, especially after fine-tuning, outperform classical approaches in terms of both accuracy and flexibility. For example, after fine-tuning, Gemma 3 4B achieved a ROUGE-L score of 0.5975 on the test dataset. Explainable methods help make the models' decision-making processes transparent and comprehensible, but so far they only provide selective insights and do not yet allow complete transparency. This work thus makes an important contribution to assessing the potential and limitations of compact multimodal language models for clinical use in pathology.

Danksagung

Wir danken Prof. Dr. Jasmina Bogojeska herzlich für ihre wertvolle Unterstützung während der gesamten Arbeit. Sie stand uns jederzeit mit fachkundiger Beratung und inspirierenden Ideen zur Seite. Sie hat sich in spezifische Themenbereiche unseres Forschungsvorhabens eingearbeitet, für uns wissenschaftliche Berichte recherchiert und uns immer geholfen, wenn wir Probleme hatten. Diese engagierte Hilfe hat entscheidend zum Fortgang und Erfolg dieser Bachelorarbeit beigetragen.

Inhaltsverzeichnis

Abkürzungsverzeichnis

MLLM	**M**ultimodal **L**arge **L**anguage **M**odels
LLM	**L**arge **L**anguage **M**odels
VQA	**V**isual **Q**uestion **A**nswering
CNN	**C**onvolutional **N**eural **N**etworks
RNN	**R**ecurrent **N**eural **N**etworks
ViT	**V**ision **T**ransformer
ResNets	**R**esidual **N**etworks
LoRA	**L**ow-**R**ank **A**daptation
ICL	**I**n-**C**ontext **L**earning
BLEU	**B**ilingual **E**valuation **U**nderstudy
ROUGE-L	**R**ecall**O**riented **U**nderstudy for **G**isting **E**valuation
BERTScore	**B**idirectional **E**ncoder **R**epresentations from **T**ransformers **S**core
XAI	**X**plainable **A**rtificial **I**ntelligence
IG	**I**ntegrated **G**radients
BioBERT	**B**iomedical **B**idirectional **E**ncoder **R**epresentations from **T**ransformers
BiLSTM	**B**idirectional **L**ong **S**hort-**T**erm **M**emory

Kapitel 1

Einleitung

Die zunehmende Verfügbarkeit medizinischer Bilddaten und die Fortschritte im Bereich der künstlichen Intelligenz eröffnen neue Möglichkeiten für automatisierte Diagnose- und Entscheidungsunterstützungssysteme. Gleichzeitig stellt sich die Frage, wie präzise und nachvollziehbar solche Systeme im klinischen Einsatz tatsächlich sind. Dieses Kapitel gibt einen Überblick über das Thema und beschreibt die Forschungsziele, die im Rahmen der Arbeit bearbeitet werden.

1.1 Kontext und Motivation

Die Verarbeitung medizinischer Bilddaten, inbesondere in der Pathologie, erfordert nicht nur umfangreiche Fachkenntnisse, sondern ist auch zeitaufwändig und anfällig für subjektive Interpretation. Gleichzeitig wächst durch den demografischen Wandel und individualisierte Therapiekonzepte der Bedarf an schnellen und verlässlichen Diagnoseverfahren. Die steigende Datenmengen steht jedoch im Widerspruch zur stagnierenden Anzahl spezialisierter Pathologen. Diese Entwicklung unterstricht die Relevanz automatisierte Assistenzsysteme, inbesondere multimodale Visual Question Answering (VQA) Systeme.

Multimodale Large Language Models (MLLMs), welche visuelle und sprachliche Informationen miteinander verbinden, haben jüngst erhebliche Fortschritte gemacht und zeigen auf allgemeinen Benchmarks beeindruckende Leistungen[1]. Dennoch stellt sich die Frage, ob solche Modelle auch im medizinischen Kontext, wo Präzision und Erklärbarkeit essentiell sind, zuverlässig eingesetzt werden können. Diese Unsicherheit bildet die zentrale Motivation dieser Arbeit. Wir untersuchen, ob und wie kompakte MLLMs wie Gemma 3 4B oder Qwen 2.5VL 3B im Bereich der Pathologie-VQA, basierend auf dem PathVQA-Datensatz[2], eingesetzt werden können und ob ihre Entscheidungsfindungen durch Erklärbarkeitsansätze nachvollziehbarer werden.

1.1.1 Forschungsziele

Die vorliegende Arbeit verfolgt folgende Forschungsziele:

- Entwicklung und Evaluierung zweier multimodaler VQA-Systeme für den medizinischen Bereich unter Einsatz neuester Deep-Learning-Modelle, mit besonderem Fokus auf kompakten Sprachmodellen und der Adressierung der domänenspezifischen Herausforderungen des PathVQA-Datensatzes.

- Vergleich der Leistungsfähigkeit klassischer VQA-Ansätze mit kompakten multimodalen Sprachmodellen auf dem PathVQA-Datensatz.
- Untersuchung möglicher Leistungssteigerungen durch gezieltes Fine-tuning der Modelle im Vergleich zu Zero-Shot- und Few-Shot-Szenarien.
- Analyse der Nachvollziehbarkeit der Modellentscheidungen des Gemma 3 4B Modells mithilfe von Feature-Importance-Methoden wie Integrated Gradients zur Erhöhung der Erklärbarkeit.

1.2 Aufbau der Arbeit

Die Arbeit gliedert sich in folgenden Kapitel: Zunächst erfolgt in Kapitel 2 eine umfassende Literaturübersicht, welche relevante Ansätze im Bereich des multimodalen Lernens und der medizinischen Bildinterpretation erläutert. In Kapitel 3 werden die theoretischen Grundlagen der Transformer-Architektur sowie relevante Konzepte aus dem Bereich des multimodalen Lernens und der Erklärbarkeit eingeführt. Kapitel 4 stellt den methodischen Forschungsrahmen und die verwendeten Evaluationsmetriken vor. Kapitel 5 beschreibt das experimentelle Setup einschliesslich Datenaufteilung und Evaluationstratiegie. In Kapitel 6 werden die Ergebnisse detatilliert analysiert und verglichen. Anschliessend erfolgt in Kapitel 7 eine tiefere Betrachtung der Erklärbarkeit mittels Integrated Gradients und Attention-Werten. Kapitel 8 bietet eine kritische Diskussion der Resultate im Vergleich zu bestehenden State-of-the-Art Modellen. Abschliessend fasst Kapitel 9 die Erkenntnisse der Arbeit zusammen und gibt einen Ausblick auf zukünftige Forschungsfelder und Anwendungsmöglichkeiten.

Kapitel 2

Literaturübersicht

2.1 Related Work

VQA bezeichnet den interdisziplinären Forschungsbereich, in dem Computermodelle natürliche Sprachfragen zu einer gegebenen Bilddarstellung beantworten. Erste Arbeiten wie Antol et al. [3] führten grundlegende Benchmark-Datensätze und Evaluationsprotokolle ein, um die Leistungsfähigkeit solcher Systeme vergleichbar zu bewerten. PathVQA ist eine spezialisierte Ausprägung des VQA-Tasks, die sich auf pathologische Bilder und medizinisch relevante Fragestellungen konzentriert. Im Gegensatz zu allgemeinen VQA Datensätzen, bei denen Alltagsobjekte im Vordergrund stehen, zeichnen sich PathVQA-Datensätze durch Fragen aus, die fachliche Domänenkenntnisse und komplexe Bildmerkmale erfordern [2]. Ziel dieses Related-Work-Abschnitts ist es, klassische CNN-RNN-Pipelines, Transformer-Architekturen und multimodale LLM-Ansätze systematisch vorzustellen und ihre jeweiligen Stärken und Limitationen herauszuarbeiten. Zunächst betrachten wir klassische CNN-RNN-Architekturen, die den Grundstein für alle späteren Entwicklungen legten.

Klassische VQA-Pipelines kombinieren CNNs zur Extraktion visueller Merkmale mit RNNs zur Kodierung der Fragestellung. Antol et al. [3] stellten ein erstes VQA-Basismodell vor, das Bild-Features aus einem vortrainierten CNN extrahiert und diese gemeinsam mit der Frage per LSTM decodiert. Zwar erreichte dieses Modell eine solide Benchmark-Performance, doch es zeigte sich, dass das starre CNN-Feature-Repräsentationsschema kaum auf unterschiedliche Fragekontexte adaptierte und lange Abhängigkeiten nur unzureichend abbildete. Pathologische Bilddaten stellen besondere Anforderungen an VQA-Modelle, da sie tiefgehende Domänenkenntnisse und feingranulare Bildanalyse erfordern. Sharma et al. [4] entwickelten mit MedFuseNet ein multimodales Modell, das neben CNN-Features auch einen Attention-basierten Text-Encoder zur Integration von Bildinformationen nutzt. Zwar verbessert diese Architektur die Genauigkeit auf dem PathVQA-Datensatz, jedoch führt ihre Komplexität zu einem hohen Rechenaufwand und erschwert die Generalisierung auf neue Datensätze. Diese Limitationen motivierten den Übergang zu Attention-basierten Architekturen, die Bild- und Textrepräsentationen verknüpfen.

Um starre Repräsentationsschemata zu überwinden, setzen neuere Ansätze auf Transformer Modelle mit Self- und Cross-Attention. Moon et al. [5] entwickelten einen multimodalen Vision-Language-Transformer für medizinische Bilder, der Masked Language Modeling und Image-Text Matching als Pre-Training Objectives einsetzt und damit eine verbesserte Ausrichtung zwischen Bild- und Textmodalitäten für

PathVQA demonstriert. Allerdings beschränkt sich die Evaluation auf einen einzigen Datensatz, wodurch die Generalisierbarkeit auf andere medizinische Bilddomänen unklar bleibt. Obwohl Vision-Language-Transformer eine enge Verzahnung zwischen Bild- und Textelementen ermöglichen, fehlt ihnen häufig das externe Wissen, das grosse Sprachmodelle (LLMs) bereitstellen können. Im folgenden Abschnitt werden daher VQA-Methoden untersucht, die auf LLMs basieren.

LLMs bieten als universelle Sprachmodelle eine attraktive Basis für VQA, da sie umfangreiches Weltwissen und flexible Prompt-Strategien integrieren. Li et al. [6] führten in „Tell-and-Answer“ einen zweistufigen Workflow ein, bei dem zuerst Attribute und Bildbeschriebe generiert und anschliessend zur Fragebeantwortung genutzt werden. Diese Modularität verbessert die Nachvollziehbarkeit, führt jedoch zu Fehler Akkumulation zwischen den Modulen. Yang et al. [7] untersuchten mit PICa den Einsatz von GPT-3 als implizites Wissens und Antwortmodul für Knowledge-based VQA. Sie wandelten Bilder über ein externes Captioning-Netzwerk in Bildbeschriebe um und nutzten wenige Beispiel-Prompts, um Fragen zu beantworten. Obwohl PICa auf OK-VQA und VQAv2 beachtliche Few-Shot-Leistungen zeigt, skaliert es kaum mit komplexen visuellen Fragen. Sterner et al. [8] verglichen einen prompt-basierten Ansatz mit einem Verfahren, das vorgelagerte Bild-Text-Features nutzt. Beide Methoden kommen ohne Fine-tuning aus, doch die reine Prompt-Variante erweist sich als weniger robust und benötigt häufig externe Feature-Encoder, um akzeptable Genauigkeit zu erreichen. Da reine Few-Shot-Prompting-Ansätze ohne Fine-tuning oft an ihre Grenzen stossen, wenden wir uns im nächsten Abschnitt einen Blick auf End-to-End multimodale LLMs, die Bild- und Textverarbeitung in einem Modell vereinen.

Statt Bild- und Sprachverarbeitung getrennt zu behandeln, setzen End-to-End MLLMs auf einen einheitlichen Ansatz, bei dem visuelle und textuelle Informationen gleichzeitig in einem Modell erlernt und genutzt werden. Naseem et al. [9] entwickelten ein auf Pathologie spezialisiertes Vision-Language-Modell, das CNN-basierte Bildmerkmale über mehrstufige Cross-Attention mit textuellen Repräsentationen eines Transformers kombiniert. Neben der leistungsstarken Bild-Text-Fusion integriert das Modell auch erklärbare Komponenten, etwa visuelle Attention-Maps, um die Nachvollziehbarkeit der Vorhersagen zu unterstützen. Jiang et al. [10] stellen mit Med-MoE ein leichtgewichtiges Mixture-of-Experts-Framework vor, das mehrere LLMs als domänenspezifische Experten und einen globalen Meta-Experten kombiniert. Nach einem dreistufigen Training verbessert Med-MoE die geschlossene Genauigkeit auf PathVQA merklich und übertrifft dabei LLaVA-Med-7B, obwohl es nur einen Bruchteil der Modellparameter hat. Die selektive Experten-Aktivierung senkt Speicherbedarf und Inferenzzeit, erfordert jedoch zusätzliches Router-Training und schneidet in offenen Fragen weiterhin schwächer ab als bei geschlossenen Fragen. Einen frühen, auf pathologisches VQA spezialisierten Transformer-Ansatz präsentierten Bazi et al. [11] mit dem Vision–Language Model for VQA in Medical Imagery. Das Modell nutzt ViT-Bildtokens, einen Text-Encoder und einen gemeinsamen Decoder, um Antworten autoregressiv zu erzeugen. Ohne umfangreiches Sprachvorwissen erreicht es auf PathVQA eine solide Leistung, liegt jedoch sowohl bei geschlossenen als auch bei offenen Fragen insgesamt unter Med-MoE. Gleichwohl zeigt diese Arbeit, dass bereits kompakte Encoder-Decoder-Architekturen durch konsequentes Cross-Modal-Training solide PathVQA-Ergebnisse erzielen können.

Kapitel 3

Grundlagen

Dieses Kapitel vermittelt die theoretischen Grundlagen, die für das Verständnis und die Umsetzung multimodaler VQA-Systeme im medizinischen Bereich erforderlich sind. Es beginnt mit der Einführung in die Transformer-Architektur, deren Varianten für Text- und Bildverarbeitung sowie deren Erweiterung auf multimodale Eingaben. Anschliessend werden ResNet-Modelle vorgestellt. Darauf folgt eine Übersicht über multimodales Lernen und dessen zentrale Fusionsstrategien, ergänzt durch effiziente Anpassungsverfahren wie LoRA und In-Context Learning. Zur Bewertung der Modellleistung werden gängige Evaluationsmetriken wie BLEU, ROUGE-L, BERTScore und G-EVAL erläutert. Abschliessend werden zentrale Ansätze der erklärbaren KI (XAI) behandelt sowie das Medical VQA-Paradigma und seine Abgrenzung zu allgemeinen VQA-Systemen diskutiert.

3.1 Transformer-Architekturen

Transformer-Architekturen bilden das Fundament moderner KI-Systeme. Sie beruhen auf Self-Attention-Mechanismen, die es ermöglichen, globale Zusammenhänge in Sequenzen gleichzeitig und effizient zu erfassen, und ersetzen damit rekurrente oder konvolutionale Strukturen.

3.1.1 Transformers

Die Transformer-Architektur, erstmals von Vaswani et al. [12] vorgestellt, markiert einen fundamentalen Fortschritt in der Modellierung von Sequenzen. Im Gegensatz zu RNNs oder CNNs, die inhärente Einschränkungen bei der parallelen Verarbeitung und der Modellierung langfristiger Abhängigkeiten aufweisen, basiert der Transformer vollständig auf Self-Attention-Mechanismen. Diese ermöglichen es, alle Positionen einer Sequenz direkt miteinander zu verknüpfen, wodurch die parallele Verarbeitung gefördert und globale Zusammenhänge effizient abgebildet werden.

Wie in Abbildung 3.1 dargestellt, besteht die Architektur aus einem Encoder- und einem Decoder-Stack. Jede der $N = 6$ identischen Schichten des Encoders enthält zwei Hauptkomponenten. Eine Multi-Head-Self-Attention-Schicht und ein Position-wise-Feed-Forward-Netzwerk. Durch Residual-Verbindungen und Layer Normalization wird eine stabile und effiziente Modellierung sichergestellt. Der Decoder erweitert dieses Konzept um eine Encoder-Decoder-Attention, die es ermöglicht, Informationen aus den Encoder-Repräsentationen direkt zu integrieren.

Das Herzstück des Transformers ist die Scaled Dot-Product Attention, die für jede Abfrage (Query) Q die Ähnlichkeit zu allen Schlüsseln (Keys) K berechnet und diese Ähnlichkeiten auf die Werte (Values) V anwendet:

$$Attention(Q, K, V) = softmax\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

Diese Attention-Mechanismen werden durch die Multi-Head Attention erweitert, bei der mehrere solcher Attention-Operationen parallel ausgeführt werden. Jeder Kopf kann dabei unterschiedliche Aspekte der Eingabe fokussieren. Die Ergebnisse werden anschliessend konkateniert und erneut projiziert, um die finale Ausgabe zu erzeugen. Da der Transformer selbst keine Informationen über die Reihenfolge der Tokens enthält, werden Positionsinformationen explizit über sinus- und cosinus-basierte Positional Encodings hinzugefügt. Diese ermöglichen es dem Modell, die Reihenfolge der Eingabesequenz korrekt zu berücksichtigen. Insgesamt verkürzt der Transformer die maximale Pfadlänge zwischen beliebigen Positionen auf eins, wodurch er in der Lage ist, auch langfristige Abhängigkeiten effizient zu modellieren.

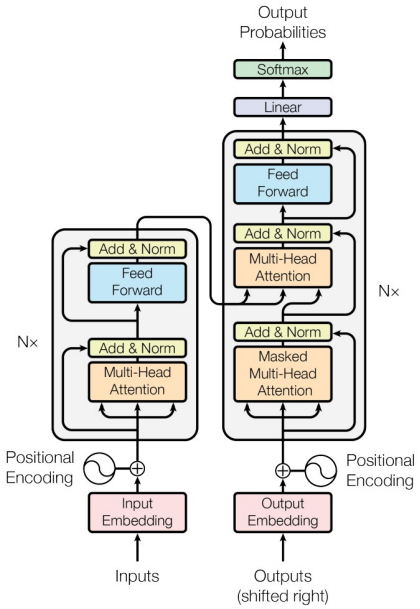


ABBILDUNG 3.1: Transformer mit Encoder- und Decoder-Stack

3.1.2 Vision Transformers

Der ViT überträgt das ursprünglich für die Sprachverarbeitung entwickelte Transformer-Konzept 3.1.1 auf Bilddaten. Dabei wird nach Dosovitskiy et al. [13] ein Bild in nicht überlappende Patches unterteilt, die jeweils linear eingebettet und mit Positionsinformationen ergänzt werden. Diese Sequenz von Patch-Embeddings dient als Eingabe für den Transformer-Encoder. Die Verarbeitungsblöcke des ViT bestehen aus alternierenden Schichten von Multi-Head Self-Attention und Feedforward-Netzwerken. Vor jedem Block wird eine Layer-Normalisierung angewendet, während Residual-Verbindungen zur Stabilisierung des Trainingsprozesses beitragen. Zusätzlich verwendet das Modell ein spezielles Klassifikationstoken, dessen Repräsentation am Ende zur Bildklassifikation dient. Die Positionsinformationen werden dabei ausschliesslich über die hinzugefügten Embeddings vermittelt. Im Gegensatz zu Convolutional Networks, die auf lokale Muster in Bilddaten ausgelegt sind, verlässt sich der ViT weitgehend auf globale Beziehungen zwischen den Patches.

3.1.3 Multimodale Transformer

Multimodale Transformer erweitern die klassische Transformer-Architektur, um verschiedene Modalitäten wie Text, Bilder, Audio oder Video gemeinsam zu verarbeiten. Dabei nutzen sie die Stärken des Self-Attention-Mechanismus, der es ermöglicht sowohl globale als auch insbesondere intermodale Zusammenhänge effizient zu erfassen [14].

Die Verarbeitung beginnt mit der spezifischen Tokenisierung und Einbettung jeder Modalität, beispielsweise durch Patch-basierte Ansätze für Bilder (siehe ViT 3.1.2) oder segmentierte Tokens für Texte, in einen gemeinsamen Repräsentationsraum. Für andere Datenquellen können dies z.B. Objekt-Regionen mit CNN-Features oder segmentierte Videosequenzen mit 3D-CNNs sein. Diese einheitliche Repräsentation ist entscheidend für die gemeinsame Verarbeitung.

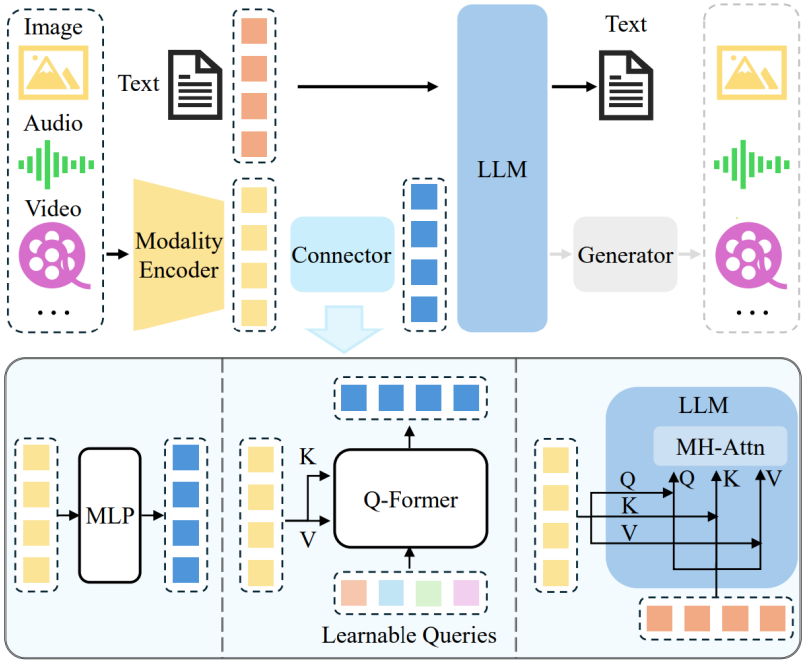


ABBILDUNG 3.2: Darstellung eines multimodalen LLMs.

Wie in Abbildung 3.2 illustriert, extrahieren in typischen multimodalen Systemen zunächst Modalitäts-Encoder die Merkmale der einzelnen Eingaben. Diese werden oft über ein Connector-Modul in eine für das zentrale Modell verständliche Repräsentation überführt. Der multimodale Transformer-Kern nutzt dann Self-Attention-Varianten, um sowohl intra- als auch intermodale Zusammenhänge zu modellieren.

Die Integration der multimodalen Informationen erfolgt durch spezifische Attention-Varianten. Ansätze reichen von Fusionstechniken wie der gewichteten Summierung von Token-Embeddings oder der direkten Verkettung bis hin zu hierarchischen Strukturen, in denen Modalitäten zunächst separat verarbeitet und dann kombiniert werden. Ein besonders leistungsfähiger Mechanismus ist hierbei die Cross-Attention, bei der Abfrage-, Schlüssel- und Wert-Embeddings gezielt zwischen den Modalitäten ausgetauscht werden, um deren wechselseitige Beziehungen zu lernen. Die Wahl des konkreten Attention-Mechanismus sowie der Fusionsstrategie beeinflusst massgeblich die Netzwerkarchitektur. Diese kann als Single-Stream-Ansatz realisiert sein, bei dem alle Modalitäten gemeinsam in einem Transformer verarbeitet

werden, als Multi-Stream-Architektur mit getrennten Verarbeitungspfaden und anschliessender Fusion, oder als hybrider Ansatz, der beide Konzepte kombiniert [15]. Diese werden im Abschnitt 3.3 genauer erläutert.

3.2 Residual Network

ResNets, wie von He et al. [16] beschrieben, ermöglichen den Einsatz tiefer neuronaler Netzwerke, ohne dass der Trainingsfehler mit zunehmender Schichtanzahl steigt. In klassischen tiefen Netzen tritt bei wachsender Tiefe oft das Degradierungsproblem auf, bei dem zusätzliche Schichten nicht zu einer verbesserten, sondern zu einer verschlechterten Leistung führen. ResNets begegnen diesem Problem durch eine grundlegende Umstrukturierung. Anstatt dass jede Schicht eine vollständige Repräsentation der Zielausgabe lernt, modellieren die Residual Blocks lediglich die Abweichung zur Eingabe. So kann das Netzwerk lernen, nur die notwendigen Änderungen zur vorherigen Repräsentation vorzunehmen.

Diese Residual Blocks beinhalten Shortcut-Verbindungen, die die Eingabe direkt auf spätere Schichten übertragen. Diese Shortcut-Verbindungen fügen keine zusätzlichen Parameter hinzu und haben keinen Einfluss auf die Komplexität des Netzwerks. Ihre Hauptaufgabe ist es, den Gradientenfluss während des Trainings zu erleichtern, was die Konvergenz, insbesondere in tiefen Netzwerken verbessert. Dadurch wird eine stabile und effiziente Optimierung ermöglicht, auch bei Netzwerken mit mehreren hundert Schichten.

Die Architektur von ResNets folgt einem modularen Aufbau. Residual Blocks werden wiederholt gestapelt und jeweils mit Batch-Normalisierung und Aktivierungsfunktionen wie ReLU ergänzt. Für besonders tiefe Architekturen wurden sogenannte Bottleneck-Blöcke eingeführt, die mehrere Convolutional-Layer in kompakter Form kombinieren und so die Effizienz und Rechenleistung weiter steigern.

3.3 Multimodales Lernen

Multimodales Lernen zielt darauf ab, Informationen aus unterschiedlichen Datenmodalitäten wie Text, Bild oder Audio zu einer einheitlichen Repräsentation zu verknüpfen. Ziel ist es, die unterschiedlichen Perspektiven und Merkmale dieser Modalitäten zu einer kohärenten und umfassenden Repräsentation zusammenzuführen. Durch die Kombination multimodaler Informationen können Aufgaben gelöst werden, die mit unimodalen Daten oft nicht oder nur eingeschränkt adressierbar sind[17].

Eine zentrale Herausforderung stellt die Fusion der unterschiedlichen Modalitäten dar. Hierbei existieren verschiedene Ansätze:

- **Early Fusion:** Merkmale der verschiedenen Modalitäten werden unmittelbar nach ihrer Extraktion zusammengeführt, etwa durch Konkatenation oder gewichtete Summation.
- **Late Fusion:** Jede Modalität wird zunächst separat verarbeitet, die daraus resultierenden Vorhersagen werden anschliessend kombiniert, beispielsweise durch gewichtete Mittelung oder Abstimmungsverfahren.

- **Hybrid oder Coordinated Fusion:** Diese Ansätze kombinieren frühe und späte Fusionstechniken, um sowohl auf Ebene der Merkmalsextraktion als auch bei der Ergebnisintegration multimodale Synergien zu nutzen.

Darüber hinaus wird zwischen gemeinsamer und koordinierter Repräsentation unterschieden. Während bei der gemeinsamen Repräsentation alle Modalitäten in einen gemeinsamen Merkmalsraum eingebettet werden, werden bei der koordinierten Repräsentation die Merkmale zunächst modalitätsspezifisch abgebildet und anschliessend durch Korrelations- oder Ähnlichkeitsmassnahmen synchronisiert.

Für die praktische Umsetzung multimodaler Systeme stehen verschiedene Lernparadigmen zur Verfügung, die sich grundlegend in ihrem Trainingsaufwand, ihrer Parametereffizienz und Flexibilität unterscheiden. Diese reichen von vollständig überwachtem Lernen mit umfangreichem Parameter-Update über parametereffiziente Adaptionsmethoden bis hin zu trainingsfreien Ansätzen zur Laufzeit.

3.3.1 Überwachtes Multimodales Lernen

Überwachtes Multimodales Lernen erfordert das Vorhandensein von Ground-Truth-Labels, um ein prädiktives Modell M_θ, parametrisiert durch θ, zu trainieren. Das Ziel besteht darin, ein Modell zu finden, das die Vorhersagegenauigkeit anhand multimodaler Eingaben maximiert, indem es die Diskrepanz zwischen Vorhersagen und tatsächlichen Labels minimiert. Mathematisch wird dieses Optimierungsproblem wie folgt dargestellt:

$$\theta^* = \arg\min_\theta \sum_{(X_i^k, y_i^k) \in D_m} \mathcal{L}_{sl} \left(M_\theta(X_i^1, \ldots, X_i^k), y^i \right)$$

Hierbei bezeichnet $\mathcal{L}_{sl}$ die Verlustfunktion, X_i^k die multimodalen Eingaben und y^i das zugehörige Ground-Truth-Label[18].

Überwachtes Multimodales Lernen stellt den klassischen Ansatz dar, erfordert jedoch das Training vieler Modellparameter. Für grosse vortrainierte Modelle sind daher parametereffiziente Alternativen wie Low-Rank Adaption besonders wichtig.

3.3.2 Low-Rank Adaptation

LoRA ist ein Ansatz zur effizienten Anpassung von LLMs, ohne dass alle Modellparameter aktualisiert werden müssen. Stattdessen werden zusätzliche, trainierbare Matrizen mit niedrigen Rängen in die Gewichtsmatrizen der Transformer-Architektur eingefügt. Diese werden während der Adaption trainiert, während die ursprünglichen Modellgewichte eingefroren bleiben. Der zentrale Vorteil dieses Verfahrens liegt in der deutlichen Reduktion der trainierbaren Parameterzahl. Im Vergleich zu einer vollständigen Feinabstimmung kann LoRA die Anzahl der trainierbaren Parameter um mehrere Grössenordnungen verringern. Gleichzeitig wird der Speicherbedarf für das Training und die Inferenz erheblich reduziert [19].

Mathematisch wird LoRA wie folgt beschrieben: Für eine vortrainierte Gewichtsmatrix $W_0 \in \mathbb{R}^{d \times k}$ wird die Anpassung ΔW als Produkt zweier niedrigrangiger Matrizen $B \in \mathbb{R}^{d \times r}$ und $A \in \mathbb{R}^{r \times k}$ dargestellt. Der Forward-Pass wird entsprechend

angepasst:

$$h = W_0 x + \Delta W x = W_0 x + BAx$$

Dabei bleibt W_0 eingefroren, während nur A und B während des Trainings optimiert werden. Durch diese Architektur kann LoRA eine effiziente Adaption an neue Aufgaben ermöglichen, ohne die Inferenzlatenz zu erhöhen. Insbesondere im Vergleich zu klassischen Adapter-Layern bietet LoRA eine grössere Flexibilität und Effizienz bei gleichzeitig hoher Modellqualität.

LoRA ermöglicht somit eine effiziente Adaption vortrainierter multimodaler Modelle an spezifische Aufgaben wie Medical VQA, ohne die computationellen Vorteile der ursprünglichen Architektur zu verlieren. Während LoRA jedoch weiterhin einen Trainingsprozess erfordert, existieren auch vollständig trainingsfreie Ansätze.

3.3.3 In-Context Learning

ICL ist eine Methode, bei der LLMs neue Aufgaben während der Inferenzphase lösen, ohne ihre Modellparameter anzupassen. Stattdessen werden Beispiele und Aufgabenbeschreibungen direkt als Eingabe in natürlicher Sprache in den Prompt integriert. Dies ermöglicht es dem Modell, sich während der Laufzeit flexibel an neue Aufgaben anzupassen, ohne explizites Fine-tuning oder zusätzliche Trainingsschritte. ICL funktioniert auf Basis der umfangreichen Vortrainingsphase von LLMs, in der diese bereits eine Vielzahl von Sprachmustern und Konzepten gelernt haben. Somit werden beim ICL die Modellparameter nicht angepasst, neue Aufgaben werden allein durch das Einfügen passender Beispiele in den Prompt gelöst [20].

3.4 Evaluationsmetriken

Die Bewertung automatisch generierter Texte ist ein zentrales Thema in der natürlichen Sprachverarbeitung. Um die Qualität solcher Ausgaben objektiv zu erfassen, werden verschiedene Evaluationsmetriken eingesetzt. Diese unterscheiden sich in ihrer Methodik und legen den Fokus auf formale Übereinstimmungen, semantische Ähnlichkeit oder Korrektheit. Folgend werden die Metriken BLEU, ROUGE-L, BERTScore und G-EVAL vorgestellt.

3.4.1 BLEU

Der BLEU-Score ist ein automatisches Gütemass, das die Ähnlichkeit einer maschinell erzeugten Ausgabe mit einer Menge menschlicher Referenzen anhand von n-Gramm Übereinstimmungen und einer Längenstrafe quantifiziert [21].

Zunächst wird für jede n-Gramm-Länge n die modifizierte Präzision p_n berechnet. Dabei zählt man alle n-Gramme der Modellantwort, wobei Mehrfachvorkommen nur bis zur maximalen Häufigkeit in einer der Referenzantworten berücksichtigt werden. p_n ergibt sich als Verhältnis der so begrenzten Trefferzahl zur Gesamtanzahl der n-Gramme in der Kandidatenantwort. Um eine Bevorzugung zu kurzer Ausgaben zu vermeiden, wird zusätzlich eine Brevity Penalty (BP) eingeführt. Der finale BLEU-Score berechnet sich als geometrisches Mittel der logarithmierten Präzisionen

über verschiedene n-Gramm-Längen hinweg, gewichtet mit w_n, und multipliziert mit der Längenstrafe:

$$\text{BLEU} = \text{BP} \cdot \exp\Big(\sum_{n=1}^{N} w_n \log p_n\Big)$$

Ein Wert nahe 1 signalisiert, dass Wortwahl, Phrasenstruktur und Länge der maschinellen Ausgabe eng an den Referenzübersetzungen liegen, während kleinere Werte grössere Abweichungen anzeigen.

3.4.2 ROUGE-L

Der ROUGE-Score ist ein Mass zur Bewertung automatisch generierter Texte, die über rein zusammenfassende Inhalte hinausgehen, etwa offene generierte Antworten oder Freitextausgaben. Während BLEU primär die Precision misst, fokussiert ROUGE auf den Recall, also den Anteil der Informationen aus den Referenzzusammenfassungen, der in der automatischen Zusammenfassung enthalten ist [22].

ROUGE-L basiert auf der Länge der längsten gemeinsamen Teilfolge (Longest Common Subsequence, LCS) zwischen der automatisch generierten und der Referenzzusammenfassung. Dadurch wird sichergestellt, dass auch semantisch ähnliche, aber nicht identisch aufeinanderfolgende Wortfolgen erkannt werden.

Mathematisch werden auf Basis von LCS der Recall R_{LCS}, die Präzision P_{LCS} und das F-Mass F_{LCS} berechnet:

$$R_{\text{LCS}} = \frac{\text{LCS}(X,Y)}{X}, \quad P_{\text{LCS}} = \frac{\text{LCS}(X,Y)}{Y}, \quad F_{\text{LCS}} = \frac{(1+\beta^2) \cdot R_{\text{LCS}} \cdot P_{\text{LCS}}}{R_{\text{LCS}} + \beta^2 \cdot P_{\text{LCS}}}$$

Dabei ist X die Referenz, Y die generierte Zusammenfassung und β ein Parameter, der den Recall stärker gewichtet.

3.4.3 BERTScore

Der BERTScore ist ein semantisch orientiertes Evaluierungsmass für generierte Texte, das im Gegensatz zu klassischen, n-Gram-basierten Metriken wie BLEU und ROUGE nicht auf exakten Wortübereinstimmungen basiert. Stattdessen nutzt es kontextuelle Einbettungen, die aus einem vortrainierten BERT-Modell gewonnen werden, um die Ähnlichkeit zwischen generierten Antworten und Referenztexten zu bewerten. Jeder Token der generierten Ausgabe wird mit dem semantisch ähnlichsten Token der Referenz verglichen, wodurch Synonyme und abweichende Ausdrucksweisen berücksichtigt werden[23].

Der finale BERTScore wird dabei als F1-Mass berechnet, das Präzision und Recall der semantischen Ähnlichkeit zwischen Referenz und Kandidat kombiniert:

$$P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} x_i^\top \hat{x}_j, \quad R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^\top \hat{x}_j, \quad F_{\text{BERT}} = \frac{2 \cdot P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}$$

Diese Eigenschaft macht den BERTScore besonders geeignet für die Bewertung offener oder freier generierter Antworten, bei denen exakte Wortübereinstimmungen nicht zwingend erforderlich sind.

3.4.4 LLM-basierte Evaluierung mit G-EVAL

Das Bewerten generierter Texte ist in der Sprachverarbeitung eine zentrale Herausforderung. Klassische Metriken wie BLEU und ROUGE korrelieren oft nur schwach mit menschlichen Bewertungen, da sie rein auf exakten n-Gramm-Übereinstimmungen basieren. Das Framework G-EVAL nutzt dagegen LLMs wie GPT-4, um Antworten auf Basis semantischer Kohärenz zu bewerten [24].

G-EVAL kombiniert dabei drei Elemente. Einen Aufgaben-Prompt mit Bewertungskriterien, automatisch generierte Chain-of-Thoughts (CoT) zur Begründung der Bewertung sowie eine Wahrscheinlichkeits-basierte Aggregation der Scores. Dadurch kann G-EVAL auch offene, längere Antworten zuverlässig evaluieren. Beispielsweise werden Begriffe wie *tumor* und *the tumour*, die semantisch identisch sind, von G-EVAL als gleichwertig erkannt und mit einem Score von 1 bewertet, während klassische Metriken hier eine Abwertung vornehmen würden. Damit eignet sich G-EVAL insbesondere für die Bewertung offener Antworten, bei denen inhaltliche Übereinstimmung entscheidender ist als exakte Wortfolgen.

3.5 Explainable AI

Explainable Artificial Intelligence (XAI) beschreibt Ansätze und Methoden, die darauf abzielen, die Entscheidungsprozesse von KI- und Machine-Learning-Modellen nachvollziehbarer und transparenter zu gestalten. Der Begriff fasst alle Bestrebungen zusammen, den sogenannten „Black-Box"-Charakter vieler moderner Modelle aufzubrechen und ihre Funktionsweise sowie die Gründe für ihre Vorhersagen oder Entscheidungen verständlich zu machen [25].

XAI-Methoden lassen sich grob in zwei Kategorien einteilen. Einerseits ante-hoc-Methoden, die bereits während des Trainings und der Modellentwicklung für mehr Transparenz sorgen, und post-hoc-Methoden, die nachträglich Erklärungen für bereits trainierte Modelle liefern. Zu den wichtigsten Verfahren gehören perturbationsbasierte Ansätze, Gradienten-basierte Methoden sowie neuere architekturbasierte Erklärverfahren, die auf Techniken wie Attention oder Transformer-Netzwerken basieren.

Im Folgenden werden exemplarisch ein Ante-hoc- und ein Post-hoc-Ansatz vorgestellt, die in dieser Arbeit zur Erklärbarkeit der eingesetzten Transformer-Modelle verwendet werden.

3.5.1 Ante-hoc: Self-Attention Werte als Erklärungsansatz

Self-Attention ist ein zentrales Werkzeug in modernen neuronalen Netzwerken, insbesondere in Transformer-basierten Modellen. Sie beschreibt, wie stark ein bestimmter Teil der Eingabedaten, wie beispielsweise ein Bildausschnitt oder ein Wort, sich auf andere Teile des Inputs bezieht. Diese Gewichtungen, sogenannte Attention Scores, geben damit Aufschluss darüber, welche Bereiche des Inputs für die Entscheidungsfindung des Modells besonders wichtig sind.

Im Kontext von MLLMs ermöglichen diese Self-Attention-Werte eine transparente Einsicht in die Verknüpfung von visuellen und textuellen Informationen. Durch die Analyse der Aufmerksamkeit auf die einzelnen Wörter einer Frage kann nachvollzogen werden, welche Begriffe für das Modell besonders relevant waren und wie Informationen über verschiedene Schichten hinweg verarbeitet werden. Damit tragen Self-Attention-Werte wesentlich dazu bei, die oft als „Black Box“ wahrgenommenen neuronalen Modelle besser verständlich und erklärbarer zu machen[26].

3.5.2 Post-hoc: Integrated Gradients

IG, eingeführt von Sundararajan et al. [27], stellen ein post-hoc-Attributionsverfahren dar, das zur Erklärung tiefer neuronaler Netze eingesetzt wird. IG quantifizieren den Beitrag einzelner Eingabefeatures zur Vorhersage des Modells, indem sie die Gradienten der Vorhersagefunktion entlang eines Pfades von einer neutralen Baseline-Eingabe bis zur tatsächlichen Eingabe integrieren. Dadurch adressieren IG zwei wichtige Anforderungen. Sensitivität und Invarianz gegenüber der Modellimplementierung.

Für eine Vorhersagefunktion $F : \mathbb{R}^n \to [0,1]$ und eine Eingabe $x \in \mathbb{R}^n$ mit einer Baseline $x' \in \mathbb{R}^n$ berechnen sich die **IG** entlang der i-ten Dimension wie folgt:

$$\text{IntegratedGrad}_i(x) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F\big(x' + \alpha(x - x')\big)}{\partial x_i}\, d\alpha$$

Diese Attributionsmethode hat den Vorteil, dass sie sowohl Sensitivität als auch Implementierungsinvarianz erfüllt. Im Vergleich zu Methoden, die ausschliesslich lokale Gradienten betrachten, erfasst IG die Relevanz von Eingabefeatures auch über nichtlineare Transformationen hinweg.

3.6 Medical Visual Question Answering

MedVQA ist ein spezialisiertes KI-Verfahren, bei dem medizinische Bilddaten gemeinsam mit Textfragen verarbeitet werden, um automatisch präzise Antworten zu generieren. In diesem Bereich werden maschinelle Lernmodelle entwickelt, um medizinische Bilder wie Röntgenaufnahmen, Magnetresonanztomographien, Computertomographien und pathologische Präparate zu interpretieren und zu analysieren. Das Hauptziel von MedVQA ist die Unterstützung der klinischen Entscheidungsfindung durch die Bereitstellung präziser und kontextuell relevanter Antworten auf Fragen von Klinikern. Dies trägt zur Entlastung des medizinischen Personals, zur Reduzierung von Diagnosefehlern und letztendlich zur Verbesserung der Patientenversorgung bei. Darüber hinaus haben MedVQA-Systeme das Potenzial, die Patientenversorgung durch die Integration in Online-Beratungsplattformen zu verbessern, insbesondere in Situationen, in denen kein direkter Zugang zu medizinischem Fachpersonal besteht [1].

Ein typisches MedVQA-System besteht aus vier zentralen Komponenten. Zunächst extrahiert ein visueller Merkmalsextraktor relevante Bildinformationen aus den medizinischen Aufnahmen. Parallel dazu wird die gestellte Frage durch einen textuellen Merkmalsextraktor in eine geeignete Repräsentation überführt. Anschliessend

werden diese beiden Informationsquellen in einem Merkmalsfusionsmodul zusammengeführt, um eine multimodale Repräsentation zu erzeugen. Auf deren Basis generiert die Antwortkomponente schliesslich eine präzise und kontextbezogene Antwort.

3.6.1 Abgrenzung zu allgemeinen VQA-Systemen

Obwohl MedVQA auf den Grundlagen allgemeiner VQA-Systeme aufbaut, weist es mehrere zentrale Unterschiede auf, die eine gesonderte Betrachtung rechtfertigen[1]:

- **Domänenspezifität:** MedVQA fokussiert sich ausschliesslich auf den medizinischen Bereich. Dies erfordert ein tiefes Verständnis medizinischer Terminologie, anatomischer Strukturen, Krankheiten und diagnostischer Verfahren.
- **Art der Bilder:** Genutzt werden spezifische medizinische Bildmodalitäten, wie Röntgen- oder MRT-Aufnahmen, die sich grundlegend von den Alltagsbildern in allgemeinen VQA-Systemen unterscheiden.
- **Art der Fragen:** Klinisch Fragen zielen auf Diagnosen, Therapiebewertungen oder die Interpretation komplexer medizinischer Befunde ab, während allgemeine VQA-Systeme meist einfache Szenen oder Objektbeschreibungen generieren.
- **Anforderungen an die Genauigkeit:** Da Fehler gravierende Folgen für die Patientenversorgung haben können, sind in MedVQA-Systemen besonders hohe Präzision und Zuverlässigkeit erforderlich.
- **Interpretierbarkeit:** Neben der Antwort selbst ist eine transparente Begründung wichtig, damit Mediziner das Ergebnis nachvollziehen und in den klinischen Arbeitsablauf integrieren können.

Kapitel 4

Methodik

Dieses Kapitel beschreibt das methodische Vorgehen dieser Arbeit und erläutert die eingesetzten Techniken und Modelle zur Beantwortung medizinischer Bild-Text-Fragen. Es beginnt mit dem gewählten Forschungsansatz, der einen Vergleich zwischen modernen multimodalen LLMs und einem klassischen Vision-Language-Ansatz vorsieht, gefolgt von einer Beschreibung der verwendeten Daten. Im Anschluss werden die Architekturen der drei untersuchten Modelle im Detail dargestellt, einschliesslich ihrer Ansätze zur Integration visueller und textueller Informationen. Daraufhin folgt die Erläuterung der multimodalen Prompt-Konstruktion. Für die Evaluierung wurden sowohl Zero-Shot- als auch Few-Shot-Szenarien umgesetzt, inklusive der Auswahl der Few-Shot-Beispiele. Abschliessend wird das verwendete Greedy Decoding beschrieben, das durch deterministische Vorhersagen konsistente und medizinisch präzise Antworten liefert.

4.1 Forschungsansatz

Ziel dieser Arbeit ist es, zu untersuchen, ob moderne multimodale LLMs wie Qwen 2.5VL 3B und Gemma 3 4B in der medizinischen Bild-Text-Verständnisaufgabe PathVQA leistungsfähiger sind als klassische Vision-Language-Ansätze. Neben dem Leistungsvergleich der Modelle wird auch erforscht, wie gut sich die Entscheidungsfindung der multimodalen LLMs durch Erklärbarkeitsmethoden transparenter gestalten lässt, um Vertrauen in medizinische Anwendungen zu schaffen. Für diesen Vergleich wurde ein experimenteller Ansatz gewählt. Die beiden multimodalen LLMs wurden zunächst im Zero-Shot- und Few-Shot-Szenario evaluiert, um ihre generelle Leistungsfähigkeit zu überprüfen. Anschliessend wurden alle drei Modelle, einschliesslich der Baseline, auf den PathVQA-Datensatz trainiert und anhand verschiedener Evaluationsmetriken beurteilt. Abschliessend wurde das leistungsfähigste Modell eingehend analysiert und für den Einsatz von Erklärbarkeitsansätzen vorbereitet.

4.2 PathVQA-Datensatz

Der PathVQA-Datensatz kombiniert pathologische Bilder mit Frage–Antwort-Paaren und deckt damit exakt den Informationsbedarf unserer Untersuchung ab. Der PathVQA Datensatz wurde 2020 von Xuehai et al. vorgestellt [2]. In dieser Arbeit nutzen wir die überarbeitete Version vom 15. Februar 2023, die zusätzliche Annotationskategorien sowie umfassende Fehlerkorrekturen enthält. Doppelt vorhandene Bild–Frage–Antwort-Triplets wurden in dieser überarbeiteten Version entfernt und

kleinere Inkonsistenzen bereinigt. Die Daten sind direkt über die Hugging-Face-Datasets-API verfügbar und bilden die Grundlage aller Experimente.

4.2.1 Qualitative Bewertung

Der PathVQA-Korpus vereint 4 289 histopathologische Bilder aus zwei Standardwerken, nämlich dem Textbook of Pathology [28] und Robbins Basic Pathology [29] sowie aus der open-access PEIR-Bibliothek [30]. Die ausschliessliche Nutzung von Lehrmaterial beseitigt datenschutzrechtliche Risiken, da keine patientenidentifizierbaren Informationen enthalten sind und führt zugleich zu einer stilistischen Homogenität der Abbildungen. Seltene Entitäten und Artefakte der Routinediagnostik sind damit unterrepräsentiert. Die Frage-Antwort-Paare wurden halbautomatisch aus den Bildlegenden extrahiert und anschliessend von Fachpathologinnen und -pathologen einzeln verifiziert.

4.2.2 Statistische Verteilung

Dere Datensatz umfasst 32 632 Bild–Frage–Antwort-Paare, im Folgenden als Beispiele bezeichnet. Diese Beispiele verteilen sich auf 19 654 Trainingseinträge, 6259 Validierungseinträge und 6719 Testeinträge. Insgesamt enthält der Datensatz 4289 unterschiedliche Bilder. Davon entfallen 2599 auf den Trainingssplit, 832 auf den Validierungssplit und 858 auf den Testsplit. Die Textlängen weisen eine kompakte Struktur auf. Fragen bestehen durchschnittlich aus $\mu_q = 6{,}28 \pm 4{,}58$ Wörtern, wobei der Median 5 Wörter beträgt und die Spannweite von 2 bis 38 Wörtern reicht. Antworten sind noch kürzer, ihr Mittelwert beträgt $\mu_a = 1{,}80 \pm 2{,}27$ Wörter, der Median liegt bei einem Wort und die Spannweite reicht von einem bis 34 Wörtern. Diese kurzen Antworten deuten darauf hin, dass der Datensatz vorwiegend geschlossene oder stark fokussierte Fragestellungen enthält. Die Verteilung der Fragetypen zeigt ein deutliches Übergewicht binärer Yes/No-Fragen. Knapp die Hälfte aller Einträge fällt in diese Kategorie, während die restlichen Typen von „What"-Fragen angeführt werden, die 41 Prozent ausmachen. Die genaue Aufschlüsselung ist in Tabelle 4.1 zusammengefasst. In Validierungs- und Testsplit liegen die Anteile jeweils höchstens 1 Prozent neben den Trainingswerten, sodass keine zusätzlichen Verzerrungen entstehen.

Fragetyp	Anzahl	Prozentsatz
Yes/No	16194	49.6 %
What	13337	40.9 %
Where	2155	6.6 %
How	695	2.1 %
Why	114	0.3 %
Other	86	0.3 %
When	49	0.2 %

TABELLE 4.1: Verteilung der Fragetypen

4.3 Modellarchitekturen

Dieser Abschnitt beschreibt die drei Modellarchitekturen, die in dieser Arbeit für die Beantwortung medizinischer Bild-Text-Fragen untersucht wurden. die selbst implementierte Baseline-Architektur auf Basis von ResNet50 und BioBERT, sowie die beiden multimodalen Large Language Models Qwen 2.5VL 3B und Gemma 3 4B. Für jedes Modell werden die technischen Grundlagen und die Art der multimodalen Fusion erläutert.

4.3.1 Baseline Resnet50 & BioBert

Für die Entwicklung der Baseline wurde die Methodik aus Naseem et al. [9] übernommen und für das eigene Projekt nachgebaut. Dabei wurde ein Vision-Language-Modell implementiert, das Bild- und Sprachinformationen kombiniert und für die Beantwortung medizinischer Fragestellungen genutzt wird.

Die Bild-Features wurden mit einem vortrainierten ResNet50 extrahiert, bei dem die finalen Klassifikationsschichten entfernt wurden, um ausschliesslich die tiefen Feature-Maps des Modells zu nutzen. Diese Features wurden anschliessend durch eine 2D-Convolution-Schicht und einen linearen Layer transformiert, um eine einheitliche Bildfeature-Dimension von 512 zu erzeugen.

Für die Sprach-Features kam BioBERT als Encoder zum Einsatz, um kontextualisierte Repräsentationen der Fragen zu generieren. Um die bidirektionale Kontextinformation weiter zu verstärken, wurden diese Ausgaben in ein BiLSTM-Modul eingespeist und danach auf 512 Dimensionen projiziert. Zusätzlich wurde eine Positional-Encoding-Schicht verwendet, um die Sequenzstruktur der Frage abzubilden.

Die fusionierten Bild- und Sprachfeatures wurden in einem zweistufigen Multi-Head-Attention-Mechanismus zusammengeführt. Diese Fusion ermöglichte es sowohl globale Bildkonzepte als auch lokale Textdetails in einer gemeinsamen Repräsentation zu verarbeiten. Im Anschluss daran wurde ein Transformer-Decoder mit zwei Schichten eingesetzt, der auf Basis der kombinierten Features die Antwortsequenz schrittweise generierte, wobei jedes Token abhängig von den zuvor erzeugten Tokens vorhergesagt wurde.

Der Aufbau orientert sich eng an der Originalarchitektur, wurde jedoch im Rahmen dieser Arbeit vollständig eigenständig implementiert.

4.3.2 Qwen 2.5VL

Für die zweite Modellvariante wurde das multimodale Sprachmodell Qwen 2.5VL 3B ausgewählt. Dieses Modell kombiniert ein autoregressives Sprachmodell mit einer visuellen Verarbeitungspipeline und wurde speziell für visuelle Frage-Antwort-Aufgaben entwickelt. Die Auswahl fiel auf Qwen 2.5VL 3B, da es zum Zeitpunkt März zu den leistungsstärksten öffentlich verfügbaren Modellen im Open VLM Benchmark [31] zählte und daher eine geeignete Grundlage für die Untersuchung von medizinischen Bild-Text-Verständnisaufgaben bot.

Das Architekturkonzept von Qwen2.5-VL basiert auf einem Encoder-Decoder ähnlichen Ansatz, bei dem Bild- und Textinformationen gemeinsam in einer einheitlichen Sequenz verarbeitet werden [32]. Der visuelle Encoder ist ein 32-schichtiger ViT, der

das Eingabebild in 14×14-Pixel-Patches zerlegt. Diese Bild-Patches werden als visuelle Tokens behandelt und in mehreren Schichten mit Self-Attention verarbeitet, um eine differenzierte visuelle Repräsentation zu erzeugen.

Für die Textverarbeitung wird ein Transformer-basiertes Sprachmodell eingesetzt. Die Fragen werden in Token zerlegt, mit Positionskodierungen versehen und in einem Sprachkontext modelliert.

Die multimodale Fusion erfolgt in einer Cross-Attention-Schicht, in der die visuellen Embeddings als Query-Vektoren und die textuellen Embeddings als Key- und Value-Vektoren eingebunden sind. Diese Integration ermöglicht es, visuelle und sprachliche Informationen in einer gemeinsamen Repräsentation zusammenzuführen. Der Decoder des Modells basiert ebenfalls auf einem Transformer-Stack und generiert autoregressiv die Antwortsequenz. Dabei wird jeder neue Token schrittweise in Abhängigkeit der bisherigen Tokens und der kombinierten multimodalen Embeddings erzeugt.

4.3.3 Gemma 3

Die dritte Modellvariante, die in dieser Arbeit untersucht wird, ist das multimodale Sprachmodell Gemma 3 4B. Dieses Modell vereint ein autoregressives Sprachmodell mit einem visuellen Verarbeitungssystem und wurde eigens für Aufgaben des visuellen Frage-Antwortens entwickelt. Zum Zeitpunkt März stellte Gemma 3 das aktuellste und leistungsfähigste Modell seiner Serie dar und bildet somit eine geeignete Ausgangsbasis für die Experimente zum medizinischen Bild-Text-Verständnis [33].

Im Gegensatz zu einem klassischen Encoder-Decoder-Ansatz nutzt Gemma 3 einen reinen Decoder-Transformer, in dem visuelle und sprachliche Eingaben gemeinsam in einer durchgehenden Sequenz verarbeitet werden. Für die visuelle Verarbeitung wird ein SigLIP Vision Transformer eingesetzt, der die Eingabebilder mit einer Auflösung von 896 × 896 Pixeln analysiert und dabei 256-dimensionale visuelle Tokens generiert. Diese visuellen Tokens werden direkt in die Sprachverarbeitung integriert, ohne dass eine zusätzliche Vorverarbeitungsschicht erforderlich ist.

Die Textverarbeitung erfolgt durch einen Transformer-Decoder mit Grouped-Query Attention, der verschiedene Self-Attention-Ebenen kombiniert. Dadurch werden sowohl lokale als auch globale semantische Beziehungen der Texttokens modelliert.

Die Fusion von Bild- und Texteingaben findet in einem einheitlichen Sequenzformat statt. Hierbei werden die visuellen und sprachlichen Embeddings in einem gemeinsamen Transformermodul zusammengeführt, was eine enge Kopplung beider Modalitäten ermöglicht. Der Decoder generiert schliesslich autoregressiv die Antworttokens, wobei die Integration der multimodalen Informationen in jedem Schritt berücksichtigt wird.

4.4 Evaluierungsmethoden

Dieser Abschnitt beschreibt die methodischen Schritte zur Bewertung der Modelle. Dazu zählen die Gestaltung der Prompts, die Durchführung von Zero- und Few-Shot-Evaluierungen sowie die vorbereitende Datenverarbeitung und Antwortgenerierung.

4.4.1 Multimodale Prompt-Konstruktion

Für beide Modelle, Qwen 2.5VL 3B und Gemma 3 4B, wurde für diese Arbeit ein systematischer Ansatz zur multimodalen Prompt-Konstruktion umgesetzt, um eine präzise Steuerung des Modellverhaltens zu gewährleisten. Der Kern dieses Ansatzes ist die Verwendung eines System-Prompts, der als übergeordnete Anweisung fungiert und das Modell gezielt auf die Bearbeitung medizinischer visueller Frage-Antwort-Aufgaben ausrichtet. Diese System-Prompts wurden so formuliert, dass sie die Rolle des Modells als medizinischer Pathologie-Experte klar definieren und gleichzeitig sicherstellen, dass sich die Antworten ausschliesslich auf die im Bild erkennbaren visuellen Informationen stützen.

Für die Eingabedaten wurde eine strukturierte Nachrichtenhierarchie implementiert, die aus drei aufeinanderfolgenden Nachrichten besteht. Einer System-Nachricht zur Kontextualisierung des Modells, einer Benutzer-Nachricht mit dem zu analysierenden Bild und der zugehörigen Frage, sowie einer Antwort-Nachricht des Modells. Das Bild und die Frage werden dabei gemeinsam als multimodaler Eingabekontext übergeben. Das System-Prompt diente dabei als Meta-Instruktion, um das Verhalten des Modells präzise zu steuern und eine konsistente Beantwortung der Fragen sicherzustellen. Die verwendete System-Nachricht lautete dabei:

> *You are a medical pathology expert. Your task is to answer medical questions based solely on the visual information in the provided pathology image. Focus only on what is visible in the image — do not rely on prior medical knowledge, assumptions, or external information. Your responses should be short, factual, and medically precise, using appropriate terminology. Do not include any explanations, reasoning, or additional text. Use a consistent format, without punctuation, and avoid capitalisation unless medically required. Only return the exact answer.*

Diese Konstruktion der multimodalen Prompts orientiert sich am Konzept der System Message Generalization, das in der Arbeit von Lee et al. [34] beschrieben wird. Durch die explizite Vorgabe des gewünschten Verhaltens in der System-Nachricht wird das Modell in die Lage versetzt, auch auf neue, bislang unbekannte System-Prompts robust und kontextgerecht zu reagieren.

4.4.2 Zero-Shot Evaluierung

Für beide Modelle, Qwen 2.5VL 3B und Gemma 3 4B, wurde eine Zero-Shot-Evaluierung durchgeführt. Dabei wurden dem Modell ausschliesslich das jeweilige Bild und die dazugehörige Frage als Eingabe zur Verfügung gestellt. Zusätzlich wurde ein präziser System Prompt definiert, wie in 4.4.1 beschrieben. Dieser System Prompt instruiert das Modell, sich ausschliesslich auf die im Bild erkennbaren visuellen Informationen zu konzentrieren und keine externen Annahmen oder medizinisches Vorwissen einzubeziehen.

Während der Inferenzphase wurden alle Validierungsbeispiele sequenziell verarbeitet. Diese Vorgehensweise stellt sicher, dass die Zero-Shot-Evaluierung unter identischen Bedingungen für beide Modelle durchgeführt wurde und ermöglicht so eine direkte Vergleichbarkeit der Ergebnisse.

4.4.3 Few-Shot Learning

In dieser Arbeit wird das Konzept des ICL, wie in Abschnitt 3.3.3 beschrieben, für visuelle medizinische Frage-Antwort-Aufgaben umgesetzt. Dazu wurden 8 visuell gestützte Frage-Antwort-Beispiele aus dem Trainingsdatensatz ausgewählt und in die Prompts der Modelle Qwen 2.5VL 3B und Gemma 3 4B integriert. Die Auswahl dieser Beispiele erfolgte mithilfe eines SelfDis-basierten Verfahrens wie im Abschnitt 4.4.4 beschriben, um eine möglichst grosse inhaltliche Diversität sicherzustellen.

Während der Inferenzphase wurden diese Beispiele in die Prompt-Struktur eingebettet, bestehend aus einer System-Nachricht, den Few-Shot-Beispielen und der eigentlichen Benutzeranfrage. Dieser Ansatz ermöglichte es den Modellen, die Beispiele als semantischen Kontext zu interpretieren und dadurch kontextgerechtere Antworten zu generieren, ohne dass eine Anpassung der Modellparameter erforderlich war. So konnte das Potenzial von ICL gezielt genutzt werden, um komplexe Aufgaben auf Basis weniger Beispiele flexibel zu bearbeiten.

4.4.4 Few-Shot Beispielauswahl

Für die Auswahl der Few-Shot-Beispiele wurde ein auf Ähnlichkeit basierender Ansatz umgesetzt, der sich am sogenannten *SelfDis*-Verfahren orientiert [35]. Hierbei wurde für jeden Datensatz ein kombinierter Feature Vektor erstellt, der sowohl textuelle Informationen mittels BERT-Embeddings als auch visuelle Bildinformationen mittels Clip-Embeddings enthält. Diese multimodalen Repräsentationen wurden durch Konkatenation zusammengeführt und als Grundlage für die Ähnlichkeitberechnung genutzt.

Das eigentliche Auswahlverfahren folgte einem sequentiellen Selbst-Dissimilaritätsprinzip. In jeder Iteration wurde das Beispiel ausgewählt, das im Durchschnitt am unähnlichsten zu den bereits selektierten Beispielen war. Die Dissimilarität für ein Beispiel x_i gegenüber den schon ausgewählten Beispielen $x_j \in D_{\text{sel}}$ wurde dabei durch folgende Formel quantifiziert:

$$\text{SelfDis}(x_i) = - \sum_{x_j \in D_{\text{sel}}} \cos(x_i, x_j)$$

wobei $\cos(x_i, x_j)$ die Cosinus-Similarität der Embeddings von x_i und x_j bezeichnet. Ein hoher Wert der negativen Summen-Similarität diente somit als Kriterium für die Auswahl des jeweils am stärksten dissimilaren Beispiels. Dieses Vorgehen stellt sicher, dass die ausgewählten Beispiele eine möglichst hohe Diversität und damit eine breite Abdeckung der vorhandenen Datenvielfalt aufweisen.

4.4.5 Vorverarbeitung für das Training multimodaler LLMs

Für das Fine-tuning der multimodalen LLMs wurde eine angepasste Funktion implementiert, die eine konsistente und verlustfreie Datenvorbereitung gewährleistet.

Diese Funktion übernimmt mehrere Aufgaben. Zunächst werden die textuellen Eingaben im passenden Prompt-Template formatiert und die zugehörigen Bilddaten in das vom Modell geforderte Eingabeformat überführt. Anschliessend erfolgt die Tokenisierung beider Modalitäten, sodass ein einheitlicher Tensor-Batch entsteht.

Besondere Bedeutung kommt der Generierung der Loss-Labels zu. Hierbei werden irrelevante Tokens wie Padding- und Bild-Tokens gezielt maskiert, damit sie nicht in die Loss-Berechnung einfliessen. Dieser präzise Maskierungsprozess sorgt dafür, dass die Modelle sich ausschliesslich auf die semantisch relevanten Informationen konzentrieren.

Die Implementierung dieser spezialisierten Funktion stellt sicher, dass die multimodalen Transformer-Modelle stabil und effizient trainiert werden können. Durch diese strukturierte Vorverarbeitung wird eine saubere Datenbasis geschaffen, die für das erfolgreiche Fine-tuning und die anschliessende Inferenz essenziell ist.

4.4.6 Greedy Decoding

Die Antwortgenerierung in dieser Arbeit erfolgt mithilfe eines deterministischen Greedy-Decoding-Verfahrens. Im Gegensatz zu stochastischen Sampling-Methoden wie Top-k oder Beam Search wird bei Greedy Decoding in jedem Generierungsschritt stets das wahrscheinlichste nächste Token ausgewählt. Dadurch werden konsistente und wiederholbare Ergebnisse erzielt, was insbesondere für die medizinische Domäne mit ihrer hohen Anforderung an Genauigkeit und Interpretierbarkeit von Vorteil ist.

Konkret wird die Antwort $A = \{a_1, a_2, \ldots, a_n\}$ iterativ generiert, wobei jedes Token a_i durch folgende Formel bestimmt wird:

$$a_i = \arg\max_{w \in V} P(w \mid a_1, a_2, \ldots, a_{i-1}, Q, I),$$

wobei V der Vokabularraum des Modells ist, Q die Frage und I das Eingabebild. Hierbei wird stets das Token w mit der höchsten bedingten Wahrscheinlichkeit gewählt, ohne Berücksichtigung alternativer Fortsetzungen oder Sampling-Strategien.

Dieser Ansatz vermeidet unnötige Diversität und sorgt für präzise, deterministische Vorhersagen, was sich laut den Arbeiten von Song et al.[36] und Wang et al.[37] in einer robusteren und zuverlässigeren Leistung niederschlägt. Somit stellt Greedy Decoding in unserem Szenario die bevorzugte Methode für die Generierung medizinisch präziser Antworten dar.

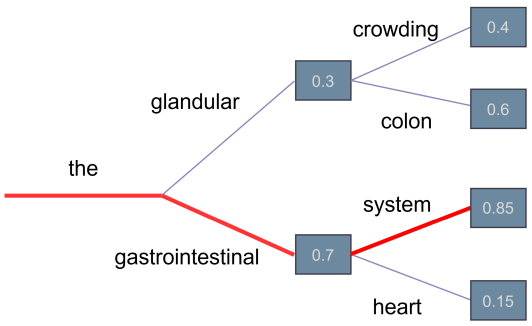


ABBILDUNG 4.1: Ablauf des Greedy Decoding

Kapitel 5

Experimentelles Setup

Dieses Kapitel beschreibt den experimentellen Aufbau der Arbeit. Es erläutert die Datenaufteilung sowie die Konfigurationen und Feinabstimmungen der Modelle, die für das Fine-tuning verwendet wurden. Ausserdem werden die eingesetzten Evaluationsmetriken und deren Anwendung für die Bewertung der Modellleistungen dargestellt, gefolgt von einem Überblick über die genutzte Hardware.

5.1 Datenaufteilung

Für die Experimente wurde der offizielle Trainings-, Validierungs- und Testsplit des Path-VQA-Datensatzes verwendet, wie in Abschnitt 4.2 beschrieben. Diese Aufteilung wurde übernommen, um eine Vergleichbarkeit mit anderen Arbeiten sicherzustellen, die denselben Datensatz und Split nutzen. Insgesamt enthält der Trainingssplit 19 654 Einträge, der Validierungssplit 6 259 und der Testsplitt 6 719 Einträge, wie in Tabelle 5.1 ersichtlich. Jeder Split umfasst sowohl offene als auch geschlossene Fragen, wodurch eine differenzierte Evaluierung der Modelle ermöglicht wird. Darüber hinaus wurde durch die offizielle Aufteilung gewährleistet, dass es keine Überschneidung von Bildinstanzen zwischen den Splits gibt.

Datenaufteilung	Datenpunkte	Bilder	Offene Fragen	Geschlossene Fragen
Train	19654	2599	9905	9749
Validation	6259	832	3134	3125
Test	6719	858	3358	3361

TABELLE 5.1: Train/Test/Validation Split

5.2 Vorgehen

Im Folgenden werden die konkreten Konfigurationen und Trainingsdetails der Modelle erläutert, um eine präzise Reproduzierbarkeit der Experimente zu gewährleisten.

5.2.1 Few-Shot Beispiele

Abbildung 5.1 zeigt exemplarisch drei der insgesamt acht ausgewählten Few-Shot-Beispiele. Jedes Beispiel besteht aus einem histologischen Bild sowie einer zugehörigen Frage-Antwort-Paarung. Diese wurden dem Modell als feste Kontextbeispiele bereitgestellt, um die Inferenz bei neuen Eingaben zu unterstützen.

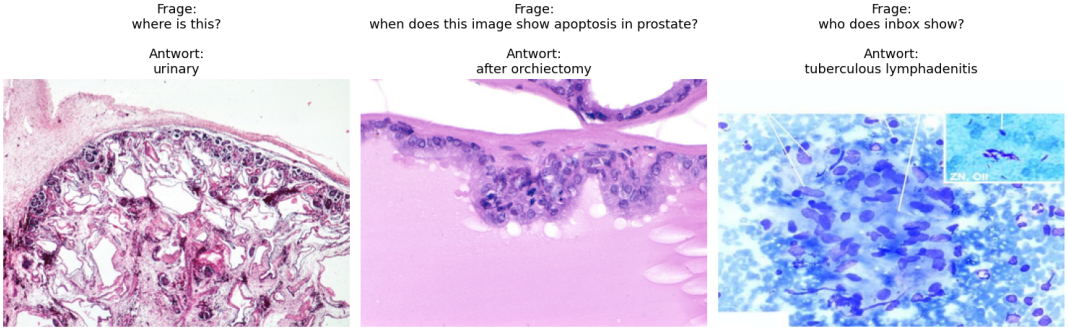


ABBILDUNG 5.1: Drei der acht ausgewählten Few-Shot-Beispiele aus dem Trainingsdatensatz

5.2.2 Baseline ResNet50 & BioBERT

Die Baseline-Implementierung kombiniert ein ResNet50-Modell zur Extraktion von Bildmerkmalen mit BioBERT für die Verarbeitung der textuellen Fragen. Die Trainings- und Modellarchitektur orientiert sich eng an den in Abschnitt 4.3.1 beschriebenen Aufbau. Die Hyperparameter wurden dabei nicht manuell festgelegt, sondern mittels eines systematischen Suchprozesses mit Optuna [38] optimiert.

Die optimalen ermittelten Hyperparameter umfassen eine Batch-Grösse von 16, eine Lernrate von 4.7×10^{-5} sowie je zwei Encoder- und Decoder-Layer. Die Multi-Head-Attention-Module verwenden 16 Köpfe und eine Feedforward-Dimension von 4096. Für die Regularisierung wurde ein Dropout-Wert von etwa 0.24 eingesetzt. Als Optimierer diente `Adam` in Kombination mit einem Cross-Entropy-Verlust. Dabei wurde Teacher Forcing verwendet, um eine stabile Generierung von Antwortsequenzen zu gewährleisten.

Das Training wurde für 20 Epochen durchgeführt. Diese Epochenzahl wurde dabei in Anlehnung an die ursprüngliche Arbeit von Naseem et al. [9] gewählt, um eine bestmögliche Vergleichbarkeit sicherzustellen.

5.2.3 Fine-tuning der multimodalen Sprachmodelle

Für das Fine-tuning der beiden multimodalen Modelle Qwen2.5 VL 3B und Gemma 3 4B wurden vortrainierte Modellgewichte verwendet, um von den bereits erlernten Repräsentationen zu profitieren. In beiden Fällen kam ein LoRA-Adapter mit einer Rank-Grösse von 8, einem LoRA-Alpha-Wert von 16 und einem Dropout-Wert von 0.05 zum Einsatz. Dieser Adapter modifiziert gezielt die Query- und Value-Projektionen der Attention-Schichten. Während die Lernrate für Qwen2.5 VL 3B bei 1×10^{-5} lag und sich an den Empfehlungen des Qwen-Entwicklungsteams orientierte, wurde für Gemma 3 4B eine Lernrate von 2×10^{-4} verwendet, basierend auf den offiziellen Veröffentlichungen zu Gemma 3 [39]. Beide Modelle wurden über zwei Epochen trainiert. Dabei kam der `paged_adamw_32bit`-Optimierer zum Einsatz, um den Speicherverbrauch während des Trainings zu reduzieren. Zusätzlich wurde in beiden Fällen eine 4-Bit-Quantisierung mithilfe von BitsAndBytes implementiert, um den Speicherbedarf weiter zu verringern. Die maximale Sequenzlänge wurde auf 128 Tokens festgelegt, basierend auf der maximalen Antwortlänge im Datensatz, wie im Abschnitt 4.2.2 beschrieben. Zudem wurde ein Warmup-Ratio von 0.1 und eine maximale Gradienten-Norm von 0.3 verwendet. Als Verlustfunktion

wurde in beiden Fällen der Cross-Entropy Loss verwendet. Für beide Modelle wurden unterschiedliche Batch-Grössen getestet, um deren Einfluss auf die Modellleistung zu evaluieren und eine Vergleichbarkeit der Ergebnisse zu gewährleisten.

5.3 Evaluation

Für die Evaluation der Modelle wurde ein differenzierter Ansatz verfolgt, um die Leistungsfähigkeit der Systeme unter verschiedenen Antwortformaten systematisch zu überprüfen. Die zur Bewertung eingesetzten Metriken werden im Folgenden detailliert vorgestellt.

5.3.1 Einsatz der Metriken in der Evaluation

Nachdem die verschiedenen Metriken zur Bewertung generierter Antworten vorgestellt wurden in Abschnitt 3.4, wird im Folgenden beschrieben, wie diese Metriken im Rahmen dieser Arbeit gezielt zur Evaluation der Modelle eingesetzt wurden. Dabei wird zwischen den drei Antworttypen unterschieden, um die Anwendung der Metriken klar zu strukturieren.

- **Close-Ended**: Dieser Antworttyp umfasst ausschliesslich Yes-/No-Fragen. Hierbei wurde der Fokus auf binäre Klassifikationen gelegt, um die Präzision des Modells bei klar definierten Antwortmöglichkeiten zu evaluieren. Entsprechende Metriken wie *Accuracy* und *F1-Scores* für die Klassen Yes und No dienten als Kennzahlen für die Bewertung.

- **Open-Ended**: Bei offenen Fragen waren längere Fliesstextantworten oder auch präzise Ein-Wort-Antworten gefordert. Die Evaluation konzentrierte sich auf die inhaltliche und sprachliche Qualität der generierten Antworten. Dafür wurden Metriken wie *BLEU-1*, *ROUGE-L*, *Token-F1* und *BERTScore-F1* verwendet. Zusätzlich kam hier *G-EVAL* zum Einsatz, um die medizinische Korrektheit und klinische Präzision der Antworten zu überprüfen. Dabei verglich G-EVAL die Modellantworten mit den korrekten Referenzantworten und lieferte sowohl eine quantitative Bewertung in Form eines Scores als auch eine qualitative Einschätzung bestehend aus einem Urteil und einer erklärenden Begründung. Um Varianzen in den Bewertungen zu minimieren, wurde der Evaluierungsprozess pro Modell dreimal durchgeführt, wobei jeweils die GPT-4o-mini-Version als Bewertungsinstanz verwendet wurde.

- **Combined**: In diesem Evaluationsmodus wurden beide Antworttypen Close Ended und Open Ended kombiniert betrachtet. Dies ermöglichte eine ganzheitliche Analyse der Modellfähigkeiten, da hier sowohl die binäre Klassifikation als auch die Fähigkeit zur Generierung offener Antworten in einem gemeinsamen Evaluationsrahmen zusammengeführt wurden. Hier wurden ebenfalls unsere Standardmetriken gebraucht, konkret wurden *BLEU 1*, *ROUGE L*, *Token F1* und *BERTScore F1* benutzt.

Zur Vergleichbarkeit mit bestehenden Arbeiten wurde eine umfassende Evaluationsstrategie umgesetzt, die sowohl Yes/No-Fragen als auch offene Antwortformate berücksichtigt. Dadurch lässt sich die Leistungsfähigkeit der Modelle differenziert beurteilen. Von der binären Entscheidungssicherheit bis hin zur inhaltlichen Qualität freier Textantworten.

5.4 Hardware

Für die Experimente in dieser Arbeit kamen unterschiedliche Hardware-Setups zum Einsatz. Das Fine-tuning der Modelle wurde auf der Plattform Lightning AI durchgeführt, die eine NVIDIA L40S GPU sowie 48 GB VRAM bereitstellte. Die Inferenz- und Evaluierungsschritte wurden hingegen auf einem Server der ZHAW durchgeführt, ausgestattet mit einer NVIDIA T4 GPU und 16 GB VRAM.

Kapitel 6

Ergebnisse

In diesem Kapitel werden die Ergebnisse der durchgeführten Experimente systematisch dargestellt. Ziel ist es, die Leistungsfähigkeit der untersuchten Modellarchitekturen im Kontext der Aufgabe des VQA auf dem PathVQA-Datensatz zu bewerten. Die Auswertung erfolgt anhand definierter Metriken für unterschiedliche Fragetypen sowie unter Berücksichtigung der verschiedenen Evaluationsmethoden wie Zero-Shot, Few-Shot und Fine-tuning. Darüber hinaus wird ein Vergleich mit dem klassischen Baseline-Modell durchgeführt. Abschliessend wird die Generalisierungsfähigkeit des leistungsstärksten Modells auf einem zuvor nicht gesehenen Testdatensatz überprüft.

6.1 Resultate

Die folgenden Abschnitte präsentieren die quantitativen Resultate der drei untersuchten Modellansätze. Zunächst werden die Ergebnisse des klassischen Baseline-Modells mit ResNet50 und BioBERT vorgestellt. Im Anschluss folgen die Resultate des Modells Qwen 2.5VL 3B, das in den drei Evaluationsmethoden Zero-Shot, Few-Shot und Fine-tuning getestet wurde. Danach wird die Auswertung des Modells Gemma 3 4B beschrieben, ebenfalls differenziert nach denselben drei Varianten. Zur besseren Nachvollziehbarkeit sind die Trainingsverläufe mit Validierungsverlust und der Entwicklung des ROUGE-L-Scores zusätzlich grafisch dargestellt.

6.1.1 Baseline Resnet50 & BioBERT

Die Baseline, bestehend aus ResNet50 für die Bild- und BioBERT für die Textverarbeitung, wurde zunächst auf dem PathVQA-Datensatz trainiert und evaluiert. Die finale Performance wird in Tabelle 6.1 dargestellt.

Modell	BLEU-1	ROUGE-L	Token-F1	BERTScore-F1
ResNet50 & BioBERT	0.5701	0.5759	0.5754	0.7816

TABELLE 6.1: Ergebnisse der Baseline auf dem vollständigen Validierungsset

Das Baseline-Modell erreichte einen BLEU-1-Wert von 0.5701, einen ROUGE-L-Score von 0.5759, einen Token-F1-Wert von 0.5754 sowie einen BERTScore-F1 von 0.7816. Diese Resultate bilden einen Referenzpunkt für die späteren Vergleiche mit den multimodalen Sprachmodellen Qwen2.5-VL-3B und Gemma 3 4B.

6.1.2 Qwen 2.5VL 3B

Tabelle 6.2 fasst die Ergebnisse der drei getesteten Varianten zusammen. Besonders hervorzuheben ist die Konfiguration mit Batch Size 2, welche im Vergleich zu den übrigen Einstellungen (siehe Anhang A.4) in allen Metriken die höchsten Werte erzielt. Der BLEU-1 Score liegt bei 0.4448, ROUGE-L erreicht 0.4505, Token-F1 beträgt 0.4501 und der BERTScore-F1 beläuft sich auf 0.7338. Die Few-Shot- und Zero-Shot-Ausführungen bleiben in sämtlichen Kennzahlen deutlich hinter diesen Werten zurück.

Modell	BLEU-1	ROUGE-L	Token-F1	BERTScore-F1
Fine-tuned	0.4448	0.4505	0.4501	0.7338
Few-Shot	0.3310	0.3375	0.3362	0.6903
Zero-Shot	0.3204	0.3264	0.3250	0.6985

TABELLE 6.2: Ergebnisse für Qwen 2.5 VL 3B auf dem vollständigen Validierungsset

Der Trainingsprozess für die Variante mit Batch Size 2 ist in Abbildung 6.1 visualisiert. Zu Beginn startet der Trainingsverlust bei einem Wert von über 1.2 und sinkt bereits innerhalb der ersten Epoche rapide auf etwa 0.2. Anschliessend bleibt er konstant niedrig. Gleichzeitig reduziert sich der Validierungsverlust gleichmässig von anfangs 0.25 auf rund 0.21. Der ROUGE-L Score verbessert sich dabei kontinuierlich und steigt von einem Anfangswert von etwa 0.32 auf einen stabilen Wert um 0.45 an. Besonders auffällig ist der gleichmässige Verlauf ohne starke Ausreisser, der sich gegen Ende des Trainings stabilisiert.

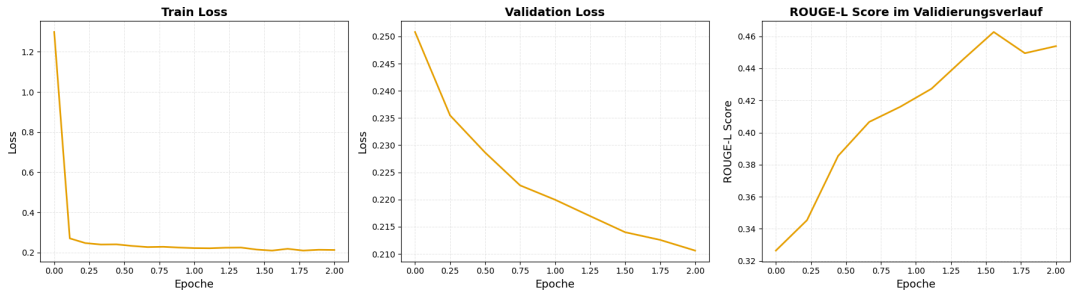


ABBILDUNG 6.1: Trainings- und Validierungsverlust sowie Entwicklung des ROUGE-L Scores für Qwen 2.5 VL 3B auf dem Validierungsset

Insgesamt zeigen diese Ergebnisse, dass das Modell Qwen 2.5-VL-3B mit der gewählten Konfiguration und dem gezielten Fine-tuning eine stabile und konsistente Trainingsdynamik erreicht.

6.1.3 Gemma 3 4B

Die kombinierten Metriken der drei getesteten Varianten sind in Tabelle 6.3 dargestellt. Dabei zeigt sich, dass die Variante mit Batch Size 2 (siehe Anhang A.3) in allen Metriken die besten Resultate erzielt. Sie erreicht einen BLEU-1 Score von 0.5946, einen ROUGE-L Score von 0.6028, einen Token-F1 Score von 0.6016 sowie einen BERTScore-F1 von 0.7964. Im Vergleich dazu schneiden die Few-Shot- und Zero-Shot-Varianten in sämtlichen Metriken niedriger ab, was auf die Vorteile eines gezielten Fine-tunings hinweist.

Modell	BLEU-1	ROUGE-L	Token-F1	BERTScore-F1
Fine-tuned	0.5946	0.6028	0.6016	0.7964
Few-Shot	0.3065	0.3124	0.3107	0.6834
Zero-Shot	0.2814	0.2878	0.2861	0.6311

TABELLE 6.3: Ergebnisse für Gemma 3 4B auf dem vollständigen Validierungsset

Der Trainingsverlauf der Fine-tuned-Variante mit Batch Size 2 ist in Abbildung 6.2 dargestellt. Zu Beginn des Trainings liegt der Trainingsverlust bei etwa 0.45 und sinkt innerhalb von zwei Epochen auf circa 0.18, was auf eine schnelle Konvergenz des Modells hinweist. Parallel dazu verringert sich der Validierungsverlust von rund 0.31 auf ungefähr 0.27. Auffällig ist dabei die kontinuierliche Steigerung des ROUGE-L Scores, der von einem Anfangswert von 0.48 auf fast 0.60 ansteigt. Eine kurzzeitige Schwankung in der Mitte des Trainingsverlaufs wird bis zum Ende der zweiten Epoche vollständig ausgeglichen. Diese Ergebnisse deuten darauf hin, dass die gewählte Konfiguration mit Batch Size 2 nicht nur stabile Trainingseigenschaften zeigt, sondern auch eine stetige Verbesserung in der Fähigkeit zur Erfassung von visuellen und textuellen Kontextinformationen bietet.

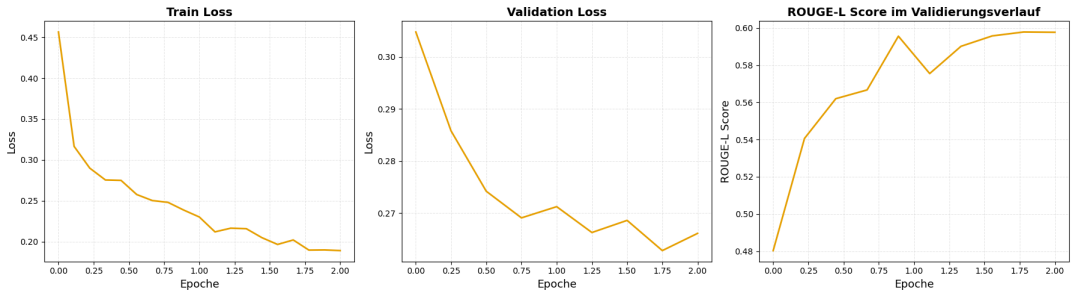


ABBILDUNG 6.2: Trainings- und Validierungsverlust sowie Entwicklung des ROUGE-L Scores für Gemma 4B auf dem Validierungsset

Insgesamt bestätigen diese Ergebnisse die Effektivität des Fine-tuned Gemma 4B Modells mit Batch Size 2 bei der Verarbeitung von multimodalen Daten. Die konstant hohen Werte bei den kombinierten Metriken und der stabile Trainingsverlauf unterstreichen die Eignung dieser Konfiguration für die VQA-Aufgabe und verdeutlichen gleichzeitig den Einfluss gezielten Fine-tuning auf die Leistungsfähigkeit von multimodalen Sprachmodellen.

6.1.4 Resultate offene Fragen

In Tabelle 6.4 sind die Resultate aller drei Modellansätze bei den offenen Fragen aufgeführt. Auf den ersten Blick zeigt sich, dass das Gemma 3 4B Modell in allen Metriken die stärkste Performance liefert. Es erreicht einen BLEU-1-Wert von 0.3094 und einen ROUGE-L-Wert von 0.3258, was beide deutlich höhere Werte sind als beim Baseline-Modell oder bei Qwen 2.5 VL 3B. Bezogen auf die Token-F1-Metrik erzielt Gemma 3 4B Fine-tuned mit 0.5405 den höchsten Wert, während das Baseline-Modell eine Token-F1 von 0.5143 aufweist. Die BERTScore-F1-Messung unterstreicht den Vorsprung von Gemma 3 4B (0.7435) im Vergleich zur Baseline (0.7259) und zu Qwen 2.5 VL 3B (0.6804). Auch im G-EVAL-Score liegt Gemma 3 4B Fine-tuned mit 0.404 ± 0.01 klar vor dem Baseline-Modell (0.333 ± 0.0008) und Qwen 2.5 VL 3B Fine-tuned (0.232 ± 0.001). Die sehr kleinen Standardabweichungen weisen darauf

hin, dass die Bewertungen über die Stichprobe hinweg äusserst konsistent sind. Die Few-Shot- und Zero-Shot-Varianten schneiden über alle Modelle hinweg deutlich schlechter ab. So erzielt Qwen 2.5 VL 3B Zero-Shot nur einen BLEU-1 von 0.0311 und einen ROUGE-L von 0.0431, und Gemma 3 4B Zero-Shot erreicht mit BLEU-1 0.0352 und ROUGE-L 0.0476 ebenfalls sehr geringe Werte. Der Fine-tuning Ansatz beider multimodaler Modelle weisen hingegen deutlich höheren Nutzen auf, wobei Gemma 3 4B mit Fine-tuning die beste Balance zwischen allen vier Metriken erreicht.

Modell	BLEU-1	ROUGE-L	Token-F1	BERT	G-Eval
ResNet50 & BioBERT	0.2272	0.2857	0.5143	0.7259	0.333 ± 0.0008
Qwen 2.5VL 3B Fine-tuned	0.1469	0.1582	0.4278	0.6804	0.232 ± 0.001
Qwen 2.5VL 3B Few-Shot	0.0445	0.0576	0.3582	0.6416	0.158 ± 0.0005
Qwen 2.5VL 3B Zero-Shot	0.0311	0.0431	0.3482	0.6409	0.199 ± 0.002
Gemma 3 4B Fine-tuned	**0.3094**	**0.3258**	**0.5405**	**0.7435**	$\mathbf{0.404 \pm 0.01}$
Gemma 3 4B Few-Shot	0.0375	0.0491	0.3521	0.6412	0.173 ± 0.015
Gemma 3 4B Zero-Shot	0.0352	0.0476	0.3509	0.6239	0.235 ± 0.0006

TABELLE 6.4: Evaluationsergebnisse offener Fragen auf dem Validierungsset

Gemma 3 4B Fine-tuned erzielt in sämtlichen Metriken die besten Ergebnisse, gefolgt von der Baseline. Dahinter liegt Qwen 2.5 VL 3B Fine-tuned, während alle Few-Shot- und Zero-Shot-Varianten deutlich schwächer abschneiden

6.1.5 Resultate geschlossene Fragen

Die Resultate für die geschlossenen Fragen sind in Tabelle 6.5 zusammengefasst. Gemma 3 4B erweist sich erneut als bestes Modell. Es erreicht eine Accuracy von 0.88 sowie jeweils hohe F1-Scores von 0.89 für die Antwort Yes und 0.87 für die Antwort No. Die Baseline erzielt eine Accuracy von 0.87 sowie F1-Scores von 0.88 für Yes und 0.85 für No, was einen sehr soliden Ausgangspunkt darstellt.

Modell	Accuracy	F1 Score Yes	F1 Score No
ResNet50 & BioBERT	0.87	0.88	0.85
Qwen 2.5VL 3B Fine-tuned	0.75	0.80	0.64
Qwen 2.5VL 3B Few-Shot	0.62	0.70	0.48
Qwen 2.5VL 3B Zero-Shot	0.68	0.75	0.56
Gemma 3 4B Fine-tuned	**0.88**	**0.89**	**0.87**
Gemma 3 4B Few-Shot	0.58	0.70	0.31
Gemma 3 4B Zero-Shot	0.63	0.65	0.62

TABELLE 6.5: Evaluationsergebnisse geschlossener Fragen auf dem Validierungsset

Das Qwen 2.5VL 3B Modell erreicht eine Accuracy von 0.75, einen F1-Score für Yes von 0.80 und für No von 0.64. Damit bleibt es unter den Leistungen von Gemma 3 4B und der Baseline. Besonders deutlich zeigen sich die Unterschiede in den Few-Shot- und Zero-Shot-Varianten beider multimodaler Modelle. Qwen 2.5VL 3B Few-Shot erzielt nur eine Accuracy von 0.62 und einen F1-Score für No von 0.48, was den niedrigsten Wert darstellt, während Zero-Shot mit einer Accuracy von 0.68 und einem F1-Score für No von 0.56 ebenfalls deutlich schlechter abschneidet als die Fine-tuned-Variante. Analog schneidet Gemma 3 4B Few-Shot mit einer Accuracy von 0.58 und einem F1-Score für No von 0.31 am schlechtesten ab, und Gemma 3 4B

Zero-Shot erreicht lediglich eine Accuracy von 0.63 und einen F1-Score für Yes von 0.65.

Gemma 3 4B Fine-tuned liegt auch hier in allen Metriken klar vorn, gefolgt von der Baseline. Qwen 2.5VL 3B Fine-tuned liegt dahinter, während alle Few-Shot- und Zero-Shot-Varianten deutlich abfallen.

6.2 Vergleich mit Baseline ResNet50 & BioBERT

Um die Fortschritte der multimodalen Modelle Qwen 2.5VL 3B und Gemma 3 4B im Vergleich zur klassischen VQA-Baseline einzuordnen, wurde deren Leistung systematisch für offene und geschlossene Fragen gegenübergestellt.

Offene Fragen: Gemma 3 4B Fine-tuned liefert durchweg die besten Resultate und übertrifft die Baseline in allen Metriken deutlich, insbesondere bei BLEU-1, ROUGE-L, Token-F1 sowie BERTScore-F1 (vgl. Tabelle 6.4). Im Gegensatz dazu bleibt Qwen 2.5VL 3B selbst mit Fine-tuning klar hinter der Baseline. Noch deutlicher fällt der Abstand bei den Few-Shot- und Zero-Shot-Varianten beider Modelle aus, die lediglich einen Bruchteil der Baseline-Leistung erreichen. Damit zeigt sich. Erst durch gezieltes Fine-tuning können multimodale Sprachmodelle ihre Vorteile bei komplexen offenen Fragen ausspielen.

Geschlossene Fragen: In der Yes/No-Kategorie fällt der Abstand geringer aus. Die Baseline erzielt nahezu identische Werte wie Gemma 3 4B Fine-tuned, mit nur marginalen Unterschieden bei Accuracy und F1-Scores. Qwen 2.5VL 3B bleibt hingegen hinter beiden zurück, wobei seine Few-Shot- und Zero-Shot-Konfigurationen nochmals deutlich schwächer abschneiden (vgl. Tabelle 6.5).

6.3 Evaluation auf dem Testdatensatz

Für die abschliessende Evaluation wurde das Modell Gemma 3 4B auf dem nicht zuvor gesehenen PathVQA-Testdatensatz überprüft. Die Ergebnisse dieser Evaluierung sind in Tabelle 6.6 zusammengefasst. Dabei erzielt Gemma 3 4B einen BLEU-1 Score von 0.5874, einen ROUGE-L Score von 0.5957, einen Token-F1 Score von 0.5944 sowie einen BERTScore-F1 von 0.7913. Diese Werte liegen auf einem vergleichbaren Niveau wie die Resultate auf dem Validierungsdatensatz, was auf eine gute Generalisierungsfähigkeit von Gemma 3 4B hinweist.

Modell	BLEU-1	ROUGE-L	Token-F1	BERTScore-F1
Gemma 3 4B	0.5874	0.5957	0.5944	0.7913

TABELLE 6.6: Ergebnisse für Gemma 3 4B auf dem Testdatensatz

Diese Auswertung belegt, dass Gemma 3 4B auch auf einem bislang unbekannten Testset konsistente Leistungen erzielt. Die durchweg stabilen Werte in allen vier Metriken deuten auf eine robuste Modellanpassung und eine hohe Zuverlässigkeit bei der Generierung präziser Antworten hin.

Im Anschluss an die oben beschriebenen Auswertungen folgen zusätzliche Analysen zur Robustheit von Gemma 3 4B. Die detaillierten, erweiterten Experimente hierzu befinden sich in Abschnitt A.1.

Kapitel 7

Erklärbarkeit

Kapitel 6 hat gezeigt, dass Gemma 3 4B das leistungsstärkste der untersuchten Modelle ist. Dennoch bleiben die internen Entscheidungsprozesse dieses Modells weitgehend intransparent. Um erste Einsichten zu gewinnen, wenden wir zwei komplementäre Erklärverfahren an. Zum einen nutzen wir Integrated Gradients, zum anderen analysieren wir mithilfe von Self-Attention-Werten die Verteilung der Aufmerksamkeitsgewichte. Dabei zielen beide Ansätze nicht darauf ab, das Modell vollständig offenzulegen, sondern punktuell sichtbar zu machen, welche Bild- und Textelemente die Vorhersagen am stärksten beeinflussen. Im Folgenden erläutern wir in Abschnitt 7.1 zunächst die Methodik und Anwendung von Integrated Gradients für geschlossene und offene Fragen und widmen uns anschliessend in Abschnitt 7.2 der Analyse mithilfe von Self-Attention-Werten.

7.1 Integrated Gradients

Wie in Abschnitt 3.5.2 erläutert, wird IG eingesetzt, um Gemma 3 4B gezielt transparenter zu machen, dadurch werden die relevanten Bild- und Textbereiche erkennbar. Die resultierenden Attributionswerte werden als visuelle Heatmaps für Bilder und als farbcodierte Relevanz für Texttokens dargestellt.

7.1.1 Implementation und Methodik

Die Analyse der Attributionen wurde vollständig mit Captum, einem Framework zur Erklärung neuronaler Netzwerke, umgesetzt [40]. Für die Analyse wird ein strukturiertes Vorgehen angewendet, das mehrere aufeinander abgestimmte Schritte umfasst.

Als Baseline dient eine stark verschwommene Version des Originalbildes mittels Gauss-Blur mit Radius 5 Pixel, die alle feinen Details entfernt, jedoch grobe Formen und Strukturen bewahrt. Dieser Blur-Baseline stellt einen neutralen Referenzpunkt entlang des Interpolationspfads dar. Damit vermeiden wir numerische Artefakte und sichern eine höhere Stabilität und Konsistenz der resultierenden Attributionen.

Für die eigentliche Attribution werden entlang eines linearen Pfads zwischen Baseline und Originaleingabe insgesamt n=300 Interpolationsschritte berechnet. In jedem dieser Schritte erstellt IG eine Zwischenversion der Bild- und Texteingaben. Dabei werden Bildpixel kontinuierlich angepasst und die Textembeddings entlang desselben Pfads interpoliert. Für jede dieser Zwischenstufen werden die Gradienten in Bezug auf das Modelloutput ermittelt, was eine präzise Quantifizierung der Relevanz

jedes Eingabeelements ermöglicht. Die Integration der Gradienten über alle Stufen hinweg liefert einen skalaren Attributionswert pro Pixel sowie pro Token.

Zur Untersuchung der Entscheidungsgrundlage bei geschlossenen Fragen wird das Logit des ersten generierten Tokens betrachtet, da dieses unmittelbar die Auswahl zwischen ‚Yes' und ‚No' bestimmt und somit als Zielwert für die Attribution dient. Bei offenen Fragen hingegen werden die Log-Wahrscheinlichkeiten aller generierten Tokens über den gesamten Antwortverlauf hinweg summiert, um auch hier den Einfluss sämtlicher Eingabeelemente erfassen zu können.

Zur Durchführung dieser komplexen Analyse wurde ein eigener Forward-Callback implementiert, der sicherstellt, dass Bild- und Texteingaben korrekt miteinander fusioniert und konsistent für alle Interpolationsstufen ausgewertet werden. Der Callback extrahiert zunächst Bild-Features mit einem visuellen Encoder, der ein Patch-Grid erzeugt, das in seiner Struktur auf die Länge der Textsequenz abgestimmt ist. Diese Features werden dann in das Chat-Template eingebettet und ersetzen dort den Bildplatzhalter.

Zusätzlich repliziert der Callback die Attention-Maps für alle IG-Stufen, sodass das Modell in jeder Phase vollständigen Kontext hat. Bei geschlossenen Fragen fokussiert sich der Forward-Callback ausschliesslich auf das Logit des ersten Tokens und liefert so Attributionen, die den unmittelbaren Einfluss der Bild- und Textelemente auf die Yes/No-Entscheidung zeigen. Für offene Fragen wird stattdessen eine Log-Softmax-Transformation auf alle Logits der generierten Antwort durchgeführt, deren aufsummierte Log-Wahrscheinlichkeiten die Zielgrösse für die Attribution bilden. Diese Summen ermöglichen es IG, den Einfluss über die gesamte Antwort hinweg zu erfassen, anstatt nur punktuelle Relevanzwerte zu berechnen.

7.1.2 Interpretation und Visualisierung der Ergebnisse

Die resultierenden Attributionen werden in einer dualen Visualisierung dargestellt, die sowohl Bild- als auch Textaspekte betont und deren Zusammenspiel nachvollziehbarer macht.

Für die Bildkomponente werden die drei Farbkanäle des Attributionstensors zu einer einzigen Heatmap pro Pixel zusammengefasst und anschliessend auf die Originalauflösung skaliert. Dieses Heatmap-Overlay wird als halbtransparentes Graustufenbild neben das Originalbild gelegt. Dunkle Bereiche deuten auf eine hohe Relevanz hin, da sie einen starken Einfluss auf die Modellentscheidung haben, während helle Bereiche eine geringe Relevanz anzeigen.

Die Textattributionen werden auf Tokenebene aggregiert, indem alle Embedding-Dimensionen summiert werden. So erhält jedes Token einen einzelnen Skalarwert, der seine Bedeutung für das Modelloutput widerspiegelt. Um Unterschiede zwischen verschiedenen Eingaben auszugleichen, werden die Token-Attributionswerte anhand ihrer euklidischen Norm angepasst. Dadurch lassen sich auch unterschiedliche Antworten oder Prompts direkt miteinander vergleichen.

Anschliessend erfolgt eine Normierung der Attributionen für die gewählte Baseline, damit ihr individueller Einfluss vergleichbar wird. Technische Sonder- und Rollentokens aus dem Chat-Template werden entfernt, sodass bei geschlossenen Fragen nur die relevanten Frage-Tokens und das finale Antwort-Token verbleiben. In der

finalen Visualisierung wird die Attribution der Bildbereiche als leicht transparentes Overlay dargestellt, das die relevanten Bildbereiche deutlich sichtbar macht.

Im Textbereich werden die Tokens nach ihrem Attributionswert eingefärbt. Grün markiert positive Beiträge, die entlang des IG-Pfads die Log-Wahrscheinlichkeit der vorhergesagten Klasse erhöhen. Rot kennzeichnet negative Beiträge, welche diese Wahrscheinlichkeit senken. Negative Werte können auch bei einer korrekten Vorhersage auftreten, sie zeigen jene Eingabeelemente, die isoliert betrachtet gegen die Entscheidung sprechen, deren Einfluss jedoch von stärker positiven Signalen anderer Elemente überkompensiert wird. Die gekoppelte Visualisierung von Bild- und Textattributionen macht solche Wechselwirkungen sichtbar und erlaubt eine differenzierte Analyse des Entscheidungsprozesses.

7.1.3 Anwendungsbeispiele

Um die zuvor beschriebene Methodik zu veranschaulichen, werden zwei konkrete Beispiele präsentiert. Zuerst eine geschlossene Yes/No-Frage, anschliessend eine offene Frage.

Geschlossene Frage

Abbildung 7.1 illustriert das Resultat für die Frage *Is acute lymphocytic leukemia present?* Oberhalb des Bildpaares sind die Wort-Relevanzen dargestellt.

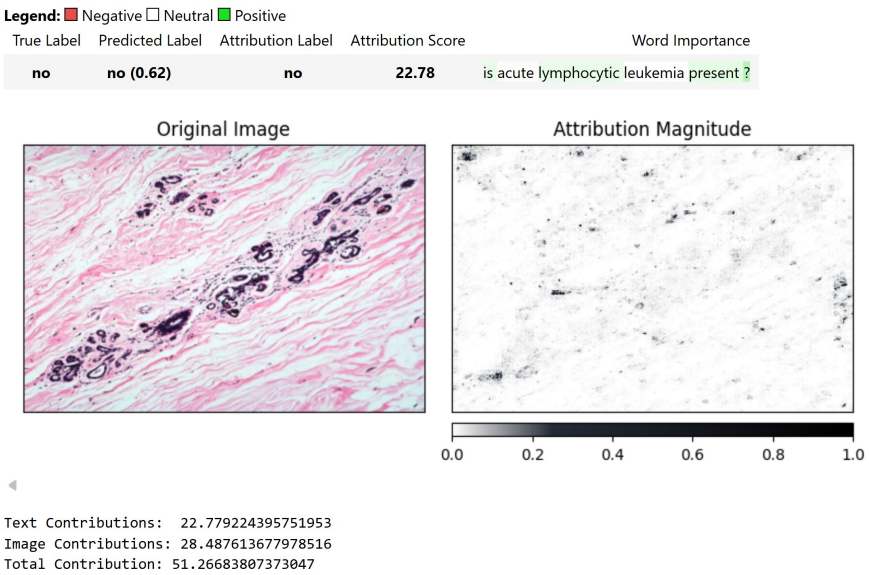


ABBILDUNG 7.1: IG für geschlossene Frage

Mit einem Gesamtattributionswert von 51.27 liefert die Eingabe einen deutlichen Hinweis auf die Antwort No. Davon entfallen 28.49 auf das Bild, während 22.78 dem Text zuzuschreiben sind. Besonders hohe positive Beiträge liefern die Tokens *is*, *lymphocytic*, *present* sowie das abschliessende Fragezeichen, da es den Fragetyp klar kennzeichnet. In der Heatmap heben sich insbesondere Zellaggregate im Bindegewebe durch intensive Färbung hervor. Negative Token-Attributionen fallen hingegen gering aus und werden durch die kombinierte Bild- und Text-Evidenz deutlich überlagert.

Offene Frage

Abbildung 7.2 zeigt das Ergebnis für die offene Frage *Where is this part in the figure?* Oberhalb des Bildpaares sind die Wort-Relevanzen dargestellt.

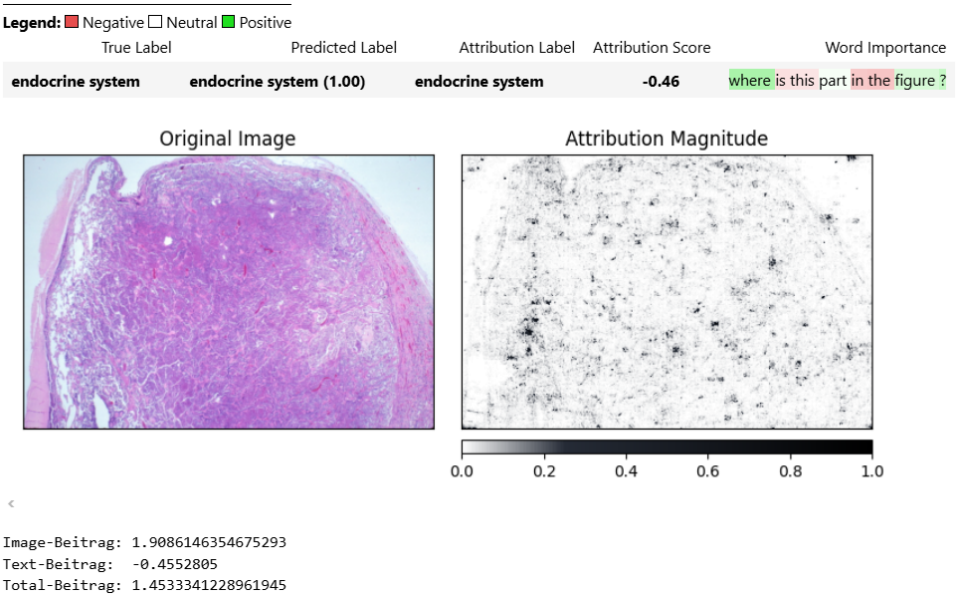


ABBILDUNG 7.2: IG für offene Frage

Hier fällt der Gesamtbeitrag mit 1.45 deutlich kleiner aus als im ersten Beispiel. Der Bildanteil 1.91 schiebt die Log-Wahrscheinlichkeit in Richtung *endocrine system*, während der Textbeitrag leicht negativ ausfällt mit −0.46. Die Wörter *in the* und *is this* sind in Rot markiert, da sie eine negative Attribution aufweisen und dadurch die Log-Wahrscheinlichkeit für *endocrine system* leicht senken. Sie liefern kaum bildbezogene Evidenz und dienen eher als strukturelle Füllwörter. Im Gegensatz dazu sind die Tokens *where*, *part*, *figure* und *?* farblich neutral bis leicht positiv hervorgehoben. Dies deutet darauf hin, dass sich das Modell zumindest teilweise auf diese inhaltlich bedeutungsvolleren Bestandteile stützt, um den Fragetyp zu erfassen. Dennoch stammt die entscheidende Evidenz für die Klassifikation fast ausschliesslich aus den visuellen Bildmerkmalen.

7.2 Self-Attention-Analyse

Für die Analyse der Self-Attention-Werte wurde für alle Fragen des Validierungsdatensatzes die Attention des letzten Transformer-Layers von Gemma 3 4B direkt aus dem Modelloutput extrahiert. Dabei wurden die Attention-Gewichte für jedes Token ermittelt, indem alle Attention-Heads zu einem Mittelwert zusammengefasst wurden. Dieser Schritt liefert eine zweidimensionale Attention-Matrix, die angibt, wie stark jedes Token von allen anderen beachtet wird.

Anschliessend wurde diese Matrix entlang der Zielrichtung gemittelt, sodass für jedes Eingabe-Token ein einzelner aggregierter Attention-Wert berechnet wurde. Diese Scores wurden mit den vom Tokenizer zurückkonvertierten Tokens verknüpft,

um den relativen Einfluss jedes Tokens auf die Modellverarbeitung sichtbar zu machen. Zur Analyse der Bedeutung verschiedener Wortarten wurden die Tokens anschliessend mit der spaCy-Bibliothek automatisch in Wortarten kategorisiert. Für jede Wortart wurde der durchschnittliche Attention-Anteil berechnet, indem die Scores aller Tokens dieser Kategorie aufsummiert und auf die Gesamtsumme der Attention normiert wurden.

7.2.1 Ergebnisse

Die Ergebnisse wurden über alle Fragen des Validierungsdatensatzes gemittelt und in Abbildung 7.3 dargestellt. Dabei zeigt sich, dass Gemma 3 4B den grössten Teil seiner Aufmerksamkeit auf Nomen richtet, die mit rund 21% den höchsten Anteil ausmachen. Danach folgen die Verben mit etwa 20% und die Adjektive mit ungefähr 19%. Diese Wortarten sind besonders relevant für die Bild-Text-Verknüpfung und tragen massgeblich zur inhaltlichen Interpretation der Fragen bei. Satzzeichen und Fragewörter nehmen moderate Anteile ein, da sie wichtige Satzstrukturen und Fragekontexte markieren. Determinatoren, Adverbien und Pronomen hingegen spielen eine eher untergeordnete Rolle.

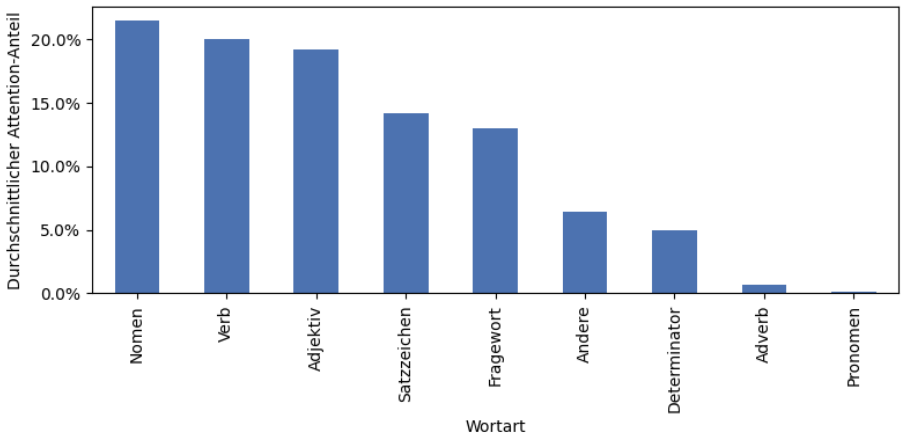


ABBILDUNG 7.3: Durchschnittlicher Attention-Anteil pro Wortart über den gesamten Validierungsdatensatz.

Der Fokus auf die Nomen des Gemma 3 4B Modells erklärt sich dadurch, dass medizinische Fragen meist auf konkrete Objekte, Gewebe oder anatomische Strukturen abzielen. Mit der starken Gewichtung von Nomen gelingt dem Modell eine präzise Verknüpfung von Bild- und Textelementen, wodurch es eine konsistente und fachlich fundierte Interpretation ermöglicht. Insgesamt geben diese Ergebnisse wertvolle Einblicke in die Arbeitsweise von Gemma 3 4B und bestätigen seine Eignung für die zuverlässige Verarbeitung medizinisch relevanter Inhalte.

Kapitel 8

Diskussion

8.1 Diskussion der Ergebnisse

Die vorliegende Arbeit liefert einen umfassenden Vergleich zwischen einem klassischen, spezialisierten Baseline-Ansatz mit einem ResNet50 & BioBERT und kompakten, modernen multimodalen LLMs wie Qwen 2.5VL und Gemma 3 4B auf dem PathVQA Datensatz. Im Zentrum stand dabei die Frage, inwiefern generative, multimodale Modelle nicht nur bestehende Methoden im medizinischen VQA leistungsmässig übertreffen, sondern sich auch in ihrer Entscheidungsfindung durch Erklärbarkeitsansätze transparenter gestalten lassen.

Die Ergebnisse zeigen, dass das Baseline-Modell insbesondere bei geschlossenen Fragen eine starke und robuste Performance erzielt. In diesem Aufgabenbereich erreicht es mit einer Accuracy von 87 Prozent nahezu die gleiche Leistung wie das deutlich grössere und komplexere Modell Gemma 3 4B, das eine Accuracy von 88 Prozent erzielt. Dies unterstreicht, dass klassische, spezialisierte Architekturen mit gezieltem Training auf domänenspezifische Daten für strukturierte Aufgaben nach wie vor konkurrenzfähig sind. Gleichzeitig zeigt sich jedoch, dass die Baseline insbesondere bei offenen Fragen, die komplexere Antworten erfordern, an ihre Grenzen stösst. So erreichte sie lediglich einen ROUGE-L Score von 0,28. Gemma 3 4B konnte hingegen insbesondere nach gezieltem Fine-tuning in allen Metriken deutlich höhere Werte erzielen und erreichte im ROUGE-L einen Score von 0,33. Dies legt nahe, dass multimodale Sprachmodelle ihre Stärken insbesondere bei komplexeren, unstrukturierten Fragestellungen ausspielen können.

Ein möglicher Grund dafür liegt in der grossen Bandbreite und Menge an Trainingsdaten, die bei der Vortrainierung solcher Modelle zum Einsatz kommen. Multimodale LLMs wie Gemma 3 4B wurden mit erheblich grösseren, diverseren Bild-Text-Datensätzen trainiert, wodurch sie ein tieferes Verständnis für visuelle und sprachliche Konzepte entwickeln können. Dieses umfassende Vorwissen versetzt multimodale LLMs in die Lage, auch anspruchsvolle medizinische VQA-Aufgaben besser zu bewältigen als klassische, separat trainierte Modelle.

Ein weiteres zentrales Ergebnis der Experimente ist die erhebliche Leistungssteigerung durch gezieltes Fine-tuning der multimodalen LLMs auf die VQA-Aufgabe. Sowhl bei Qwen 2.5VL 3B als auch bei Gemma 3 4B liegen die Few-Shot und Zero-Shot Ergebnisse klar unter den Resultaten der Fine-tuned Modelle. Die Unterschiede sind dabei teilweise sehr deutlich. Beispielsweise erreicht Gemma 3 4B im Zero-Shot bei offenen Fragen einen Rouge-L Score von nur 0.0476, während im Fine-tuning

0.326 erzielt werden. Dies verdeutlicht, dass das blosse Anwenden multimodaler Sprachmodelle auf medizinische Datensätzen in der Regel nicht aufreicht, um die Perfomance spezialisierter Baselines zu übertreffen. Erst durch gezieltes, domänenspezifisches Training kann das Potenzial moderner Modell ausgeschöpft werden.

Im direkten Vergleich der beiden multimodalen LLMs zeigt sich, dass Gemma 3 4B das Modell Qwen 2.5VL 3B in nahezu allen Bewertungsmetriken übertrifft, sowohl bei offenen als auch bei geschlossenen Fragen. Die Ursachen für diese Unterschiede könnten in der jeweiligen Modellarchitektur, der grösseren Anzahl an Parametern bei Gemma 3 4B sowie in den verwendeten Trainingsdaten und der Vortrainingsstrategie liegen. Gemma 3 4B verfügt im Vergleich etwa über eine Milliarde mehr Parameter, was ebenfalls zur besseren Leistung beitragen dürfte. Während Qwen 2.5VL 3B bereits im Fine-tuning Schwierigkeiten zeigt, mit der klassischen Baseline mitzuhalten, erzielt Gemma 3 4B regelmässig signifikantere Ergebnisse, insbesondere nach dem Fine-tuning. Dies spricht für die Überlegenheit des Gemma 3 4B Modells im Kontext der PathVQA Aufgabe.

Auffällig ist, dass in den durchgeführten Experimenten die besten Ergebnisse stets mit einer Batch Size von 2 erzielt wurden. Dies steht im Kontrast zu gängigen Empfehlungen für grosse Modelle, bei denen oft grössere Batchgrössen bevorzugt werden. Die Ursache hierfür könnte in der spezifischen Beschaffenheit des PathVQA-Datensatzes liegen. Die enthaltenen medizinischen Bilder und Fragestellungen sind hochspezialisiert und weisen grosse semantische Heterogenität auf. Eine kleine Batchsize ermöglicht dem Modell möglicherweise, individueller und feiner auf die einzelnen Beispiele zu reagieren, was insbesondere bei stark domänenspezifischen Aufgaben von Vorteil sein kann.

Ein weiterer Blick auf die eingesetzten Erklärbarkeitsmethoden zeigt, dass das IG-Verfahren mit Captum nachvollziehbar macht, welche Bildregionen und Texttokens vom Modell bei offenen und geschlossenen Fragen besonders stark gewichtet werden. So lassen sich beispielsweise fokussierte Zellstrukturen oder klinisch relevante Fachbegriffe gezielt hervorheben. Die Attention-Analyse verdeutlicht gleichzeitig, dass Gemma 3 4B semantisch wichtige Wortklassen wie Nomen, Verben und Adjektive priorisiert, was im medizinischen Kontext essenziell für das korrekte Verständnis der Fragestellungen ist. Beide Verfahren machen sichtbar, wie Bild- und Sprachelemente zusammenwirken, um etwa pathologische Muster im Gewebe zu erkennen. Allerdings liefern weder IG noch Attention Maps kausale Erklärungen oder Einblicke in tieferliegende Modellentscheidungen, sondern zeigen lediglich punktuell gewichtete Einflüsse.

Darüber hinaus zeigte sich im Rahmen erweiterter Experimente, dass das Modell trotz der statistisch nachgewiesenen Korrelation des Begriffs *image* mit Yes-Antworten keinen generellen Bias in Richtung Yes entwickelt hat. Zwar belegte ein Chi-Quadrat-Test die Verzerrung im Datensatz, doch Gemma 3 4B konnte die geschlossenen Fragen auch ohne das Schlüsselwort korrekt beantworten, was auf eine robuste semantisch-visuelle Verarbeitung hinweist. Bei offenen Fragen blieb die Leistung ebenfalls weitgehend stabil, was zeigt, dass das Modell nicht auf oberflächliche sprachliche Muster angewiesen ist. Weitere Tests wie Bildrauschen, Textverfälschung und Paraphrasierung bestätigten die Robustheit bei geschlossenen Fragen, während bei offenen Fragen eine höhere Empfindlichkeit gegenüber Störungen sichtbar wurde.

8.2 Vergleich mit State-of-the-Art Modellen

Der Vergleich unserer Ergebnisse mit anderen Arbeiten gestaltet sich schwierig, da in vielen Veröffentlichungen der verwendete Datensplit nicht angegeben ist oder die Version des PathVQA-Datensatzes nicht dokumentiert ist. Dennoch konnten wir eine Arbeit finden, die ebenfalls MLLM-basierte Methoden einsetzt und einen Vergleich erlaubt. In der Arbeit von Jiang et al.[10] wurden verschiedene multimodale Sprachmodelle evaluiert, darunter Med-MoE (Phi-2), LLaVA-Med, das auf LLaMA basiert und LLaVA 7B, jeweils mit und ohne Fine-tuning auf dem PathVQA-Datensatz.

Für geschlossene Fragen erreichte das leistungsstärkste Modell, Med-MoE mit Phi-2, eine Accuracy von 91,98 Prozent. Damit übertraf es unser bestes Modell Gemma 3 4B mit 88 Prozent um etwa vier Prozentpunkte. Dennoch zeigt sich, dass unser Modell bei den geschlossenen Fragen besser abschneidet als mehrere andere getestete Modelle in derselben Arbeit. So erreichte das Standardmodell LLaVA 7B, das auf den PathVQA-Datensatz trainiert wurde, eine Accuracy von lediglich 63,20 Prozent. Das ist ein deutlich schwächeres Ergebnis im Vergleich zu unserem Gemma 3 4B Modell.

Noch klarer zeigt sich der Unterschied bei offenen Fragen. Während Med-MoE (Phi-2) und LLaVA-Med (LLaMA 7B) BLEU-1 Scores von 32,85 Prozent bzw. 32,89 Prozent erreichten, lag unser Fine-tuned Gemma 3 4B bei deutlich höheren 59,46 Prozent. Trotz geringerer Modellgrösse konnte unser Ansatz somit eine erheblich präzisere sprachliche Generierung erreichen.

Insgesamt verdeutlicht dieser Vergleich, dass Gemma 3 4B in Bezug auf offene Fragen selbst moderne State-of-the-Art-Modelle übertreffen kann. Bei geschlossenen Fragen reicht es zwar nicht ganz an die besten Ergebnisse heran, übertrifft jedoch mehrere leistungsstarke Modelle wie das ursprüngliche LLaVA 7B oder Varianten mit bis zu 2,7 Milliarden Parametern. Dies zeigt das Potenzial kompakter und optimierter MLLMs im medizinischen VQA-Kontext.

Kapitel 9

Fazit und Ausblick

Dieses Kapitel fasst die zentralen Ergebnisse der Arbeit zusammen, reflektiert bestehende Limitationen und gibt einen Ausblick auf mögliche zukünftige Forschungsrichtungen.

9.1 Zentrale Erkenntnisse

- **Multimodale Sprachmodelle sind bei offenen Fragen überlegen**
  Gemma 3 4B übertraf die Baseline aus ResNet50 und BioBERT in allen Metriken nach dem Fine-tuning deutlich. Dies zeigt die Stärke multimodaler Sprachmodelle bei komplexen Fragestellungen im medizinischen Bereich.

- **Fine-tuning ist entscheidend für gute Ergebnisse**
  Ohne gezieltes Fine-tuning blieben Zero-Shot und Few-Shot Varianten deutlich schwächer. Erst durch Anpassung an den PathVQA Datensatz konnten die Modelle spezialisierte Methoden übertreffen.

- **Few-Shot Learning zeigt geringe Effektivität**
  Trotz durchdachter Beispielauswahl mit SelfDis konnte kein signifikanter Leistungsgewinn erzielt werden. Dies unterstreicht die Notwendigkeit echter Modellanpassung über Fine-tuning im medizinischen Bereich.

- **Klassische Modelle bleiben bei geschlossenen Fragen stark**
  Die Baseline erzielte bei Yes/No Fragen ähnlich gute Resultate wie Gemma 3 4B. Qwen 2.5VL 3B blieb selbst mit Fine-tuning hinter der Baseline. Klassische Architekturen sind hier nach wie vor konkurrenzfähig.

- **Kleine Batchgrössen verbessern die Modellleistung**
  Eine Batchgrösse von zwei führte durchgängig zu den besten Resultaten. Das spricht für eine feinere Modellanpassung bei komplexen medizinischen Daten durch kleinere Updates.

- **Erklärbarkeitsmethoden bieten Einblicke, aber keine vollständige Transparenz**
  Ansätze wie Integrated Gradients und die Auswertung von Self Attention Werten liefern punktuelle Erklärungen. Eine umfassende, globale Nachvollziehbarkeit bleibt jedoch weiterhin offen.

9.2 Limitationen

Im Rahmen dieser Arbeit ergaben sich mehrere Limitationen. Erstens musste die Analyse auf vergleichsweise kleine multimodale LLMs beschränkt werden, da die verfügbare GPU-Rechenleistung begrenzt war. Dies verhinderte die Untersuchung leistungsstärkerer Modelle, die ein differenziertes Bild der tatsächlichen Potenziale multimodaler LLMs aufzeigen könnten. Zweitens war das Hyperparameter-Tuning aufgrund dieser Hardware-Einschränkungen auf wenige Experimente mit verschiedenen Batchgrössen beschränkt. Ein umfassendes Tuning aller relevanten Parameter war somit nicht möglich.

Eine weitere Limitationen liegt in der Schwierigkeit, medizinische VQA-Ergebnisse valide zu bewerten. Es existieren derzeit keine standardisierten Evaluationsmetriken, die medizinische Relevanz eindeutig und umfassen bewerten können. Hierfür wäre die Einbeziehungen von Expertenwissen, in unserem Fall von Pathologen notwendig.

9.3 Ausblick

Zukünftige Forschungsarbeiten könnten von der Anwendung grösserer multimodaler Modelle profitieren, um deren Leistungsfähigkeit im Bereich medizinischer VQA genauer zu untersuchen. Darüber hinaus stellt der Einsatz von Reinforcement Learning-basierten Fine-tuning-Strategien, wie der kürzliche vorgestellte Ansatz des Group Relative Policy Optimization (GRPO), eine besondere vielversprechende Erweiterung dar [41]. Diese Studie deutet darauf hin, dass solche Methoden sowohl Genauigkeit als auch klinische Relevanz deutlich steigern könnten.

Des Weiteren wäre es wichtig, standardisierte Evaluationsmetriken für medizinische VQA zu entwickeln, die spezifische Anforderungen des medizinischen Kontexts berücksichtigen und idealerweise in Zusammenarbeit mit klinischen Experten validiert werden. Ein solcher Ansatz würde helfen, die Vergleichbarkeit und klinische Nützlichkeit der Ergebnisse deutlich zu verbessern und somit die praktische Relevanz multimodaler LLMs im medizinischen Alltag weiter zu stärken.

Anhang A

Anhang

A.1 Erweiterte Experimente auf Gemma 3 4B

Zur Prüfung der Robustheit von Gemma 3 4B wurden vier gezielte Anpassungen getestet. Dabei wurden alle Experimente ausschliesslich auf dem bereits durch Finetuning angepassten Gemma 3 4B Modell durchgeführt. Die folgenden Schritte wurden umgesetzt

Entfernung des Wortes „image“: Durch eine Korrelationsanalyse wurde festgestellt, dass das Wort „image“ stark mit der Antwort „Yes“ korreliert, sodass das trainierte Modell anstelle des tatsächlichen Bildinhalts lediglich auf dieses Stichwort reagiert. Von insgesamt 6259 Fragen entfallen 3125 auf Yes/No-Fragen. Ein Chi-Quadrat-Test zeigt für das Wort „image“ einen besonders hohen Wert ($\chi^2 = 450{,}77,\ p = 4{,}91 \times 10^{-100}$). In den Fragen, die „image“ enthalten, gab es 573 Yes-Antworten gegenüber lediglich 8 No-Antworten. Durch das Entfernen dieses Wortes zwingen wir das Modell dazu, semantische Signale zu nutzen, statt sich ausschliesslich am Wort „image“ zu orientieren.

Bildaugmentation: Um zu prüfen, wie robust das Modell gegenüber visuellen Variationen ist, haben wir während der Inferenz verschiedene Augmentationen auf die Eingangsbilder angewendet. Verwendet wurde eine Kombination aus Rotation, Kontraständerungen, Gauss'schem Rauschen und Bewegungsunschärfe.

Einführung von Typos: Zusätzlich haben wir in den Fragen zufällig Zeichen vertauscht oder Zeichen eingefügt, sowie ganze Wortpaare vertauscht. Diese zufälligen Störungen simulieren Tippfehler und unstrukturierte Fragestellungen, wie sie in realen klinischen oder Nutzerkontexten vorkommen können.

Paraphrasierung mit GPT-4o-mini: Schliesslich wurden alle Fragen mit GPT-4o-mini nach definierten Richtlinien umformuliert. Die Vorgaben stellten sicher, dass semantische Bedeutung und Fachterminologie erhalten bleiben.

Diese vier Eingriffe zeigen in Kombination, wie stabil Gemma 3 4B gegen Schlüsselwort-Bias, Bildverzerrungen, Tippfehler und alternative Formulierungen bleibt und wo seine Grenzen liegen.

Offene Fragen: Die durch Fine-tuning angepasstes Gemma 3 4B Modell dient als Referenz für alle Experimente. Beim Entfernen des Wortes „image“ sinken alle Metriken nur leicht. Bei Bildaugmentation fallen die Werte deutlich stärker ab. Durch Textaugmentation reduziert sich die Performance weiter unter das Niveau der Bildstörung. Die Paraphrasierung mittels GPT-4o-mini liegt etwas über der reinen Textaugmentation, bleibt aber ebenfalls unterhalb des Referenzmodells. Insgesamt zeigen die offenen Fragen, dass Bild- und Textstörungen die Leistung deutlicher reduzieren als das Entfernen des Bias-Wortes.

Experiment	BLEU-1	ROUGE-L	Token-F1	BERT	G-Eval
Entfernen des Wortes „image“	0.3024	0.3185	0.5350	0.7402	0.390
Bildaugmentation	0.2341	0.2468	0.4877	0.7170	0.315
Textaugmentation	0.2072	0.2072	0.4765	0.7000	0.305
Paraphrasierung (GPT-4o-mini)	0.2307	0.2477	0.4871	0.7117	0.361
Gemma 3 4B Fine-tuned	**0.3094**	**0.3258**	**0.5405**	**0.7435**	**0.392**

TABELLE A.1: Evaluation der offenen Fragen: Vergleich der vier Experimente

Geschlossene Fragen: Bei den geschlossenen Fragen zeigen sich klare Leistungseinbussen unter verschiedenen Störbedingungen. Wird das Wort image entfernt, sinken sowohl die Accuracy als auch die F1-Werte für Yes und No. Auch Bildrauschen beeinträchtigt die Leistung, jedoch weniger stark als bei den offenen Fragen. Textuelle Fehler führen zu einem stärkeren Leistungsabfall als Bildrauschen, bleiben jedoch über dem Niveau der Image Entfernung. Paraphrasierungen erzielen ein ähnliches Leistungsniveau wie die Textfehler und liegen insgesamt zwischen Bildrauschen und dem Referenzmodell.

Experiment	Accuracy	F1 Score Yes	F1 Score No
Entfernen des Wortes „image“	0.75	0.74	0.76
Bildaugmentation	0.85	0.86	0.84
Textaugmentation	0.77	0.76	0.79
Paraphrasierung (GPT-4o-mini)	0.77	0.79	0.75
Gemma 3 4B Fine-tuned	**0.88**	**0.89**	**0.87**

TABELLE A.2: Evaluation der geschlossenen Fragen: Vergleich der vier Experimente

Bias bei Yes/No-Fragen: Abbildung A.1 vergleicht die Häufigkeiten der Yes/No-Labels im Validierungsdatensatz (blau) im Vergleich zu den Modellvorhersagen von Gemma 3 4B (orange). Die nahezu identischen Höhen der blauen und orangen Balken bei Yes/No zeigen deutlich, dass Gemma 3 4B keinen ausgeprägten Yes/No-Bias übernimmt. Trotz dieses fehlenden Bias resultiert beim Entfernen des Stichworts „image“ ein Performanceeinbruch. Da das Modell nicht einfach auf eine übermässige Anzahl Yes-Antworten zurückgreift, lässt sich der Rückgang vielmehr auf den Wegfall semantischer Hinweise zurückführen. Die Verschlechterung der Metriken entsteht also nicht durch einen trivialen Yes/No-Bias, sondern durch die Herausforderung, semantische Inhalte und Bildinformationen ohne das Schlüsselwort zu kombinieren.

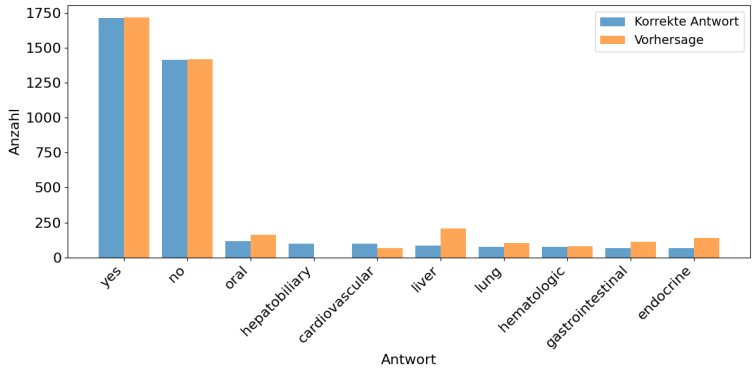


ABBILDUNG A.1: Histogramm der zehn häufigsten Modellvorhersagen im Vergleich zu den tatsächlichen Antworten.

A.2 Vergleich verschiedener Batchgrössen

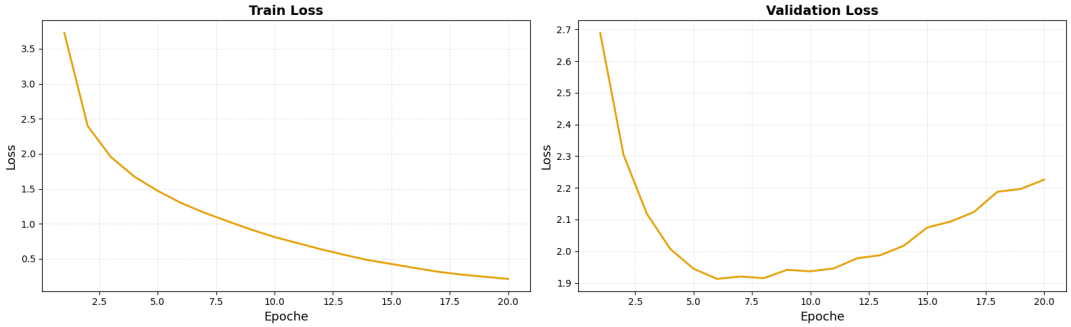


ABBILDUNG A.2: Trainings- und Validierungsverlauf für Baseline

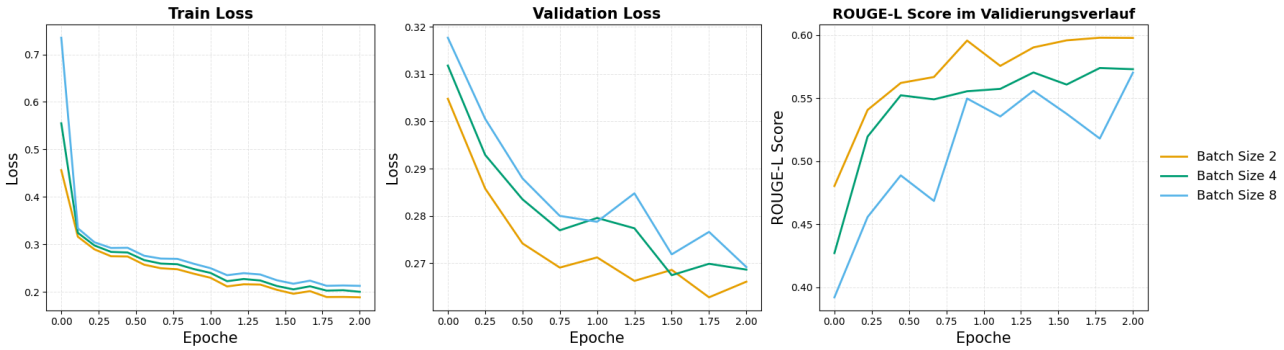


ABBILDUNG A.3: Traings- und Validierungsverlauf für Gemma 3 4B bei unterschiedlichen Batch-Grössen

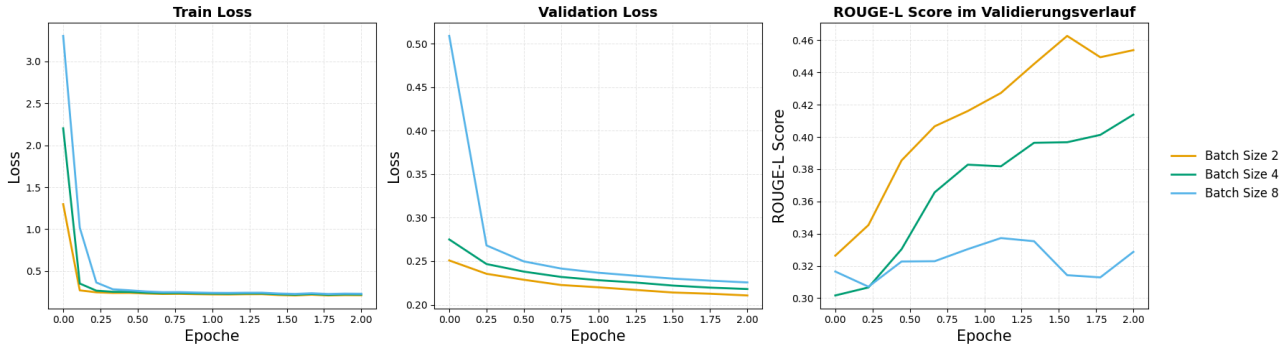


ABBILDUNG A.4: Traings- und Validierungsverlauf für Qwen 2.5VL 3B bei unterschiedlichen Batch-Grössen

A.3 Weitere Beispiele für Integrated Gradients

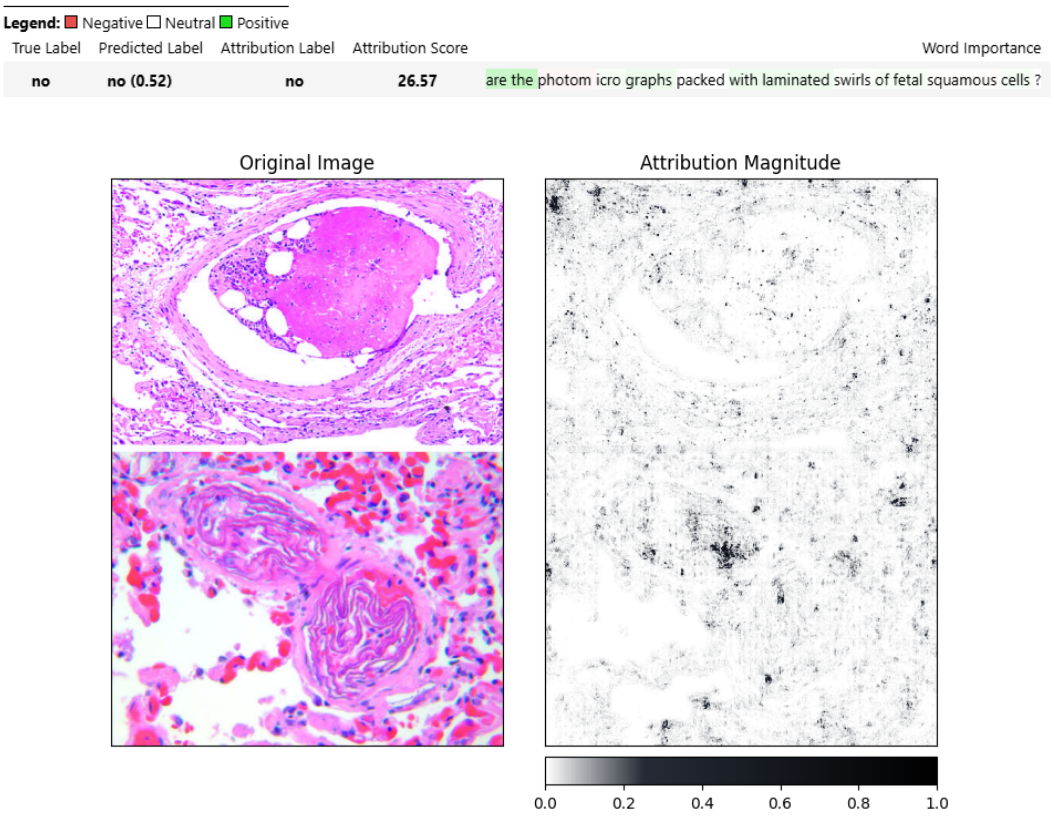


ABBILDUNG A.5: IG für die Frage „Are the photomicrographs packed with laminated swirls of fetal squamous cells?“

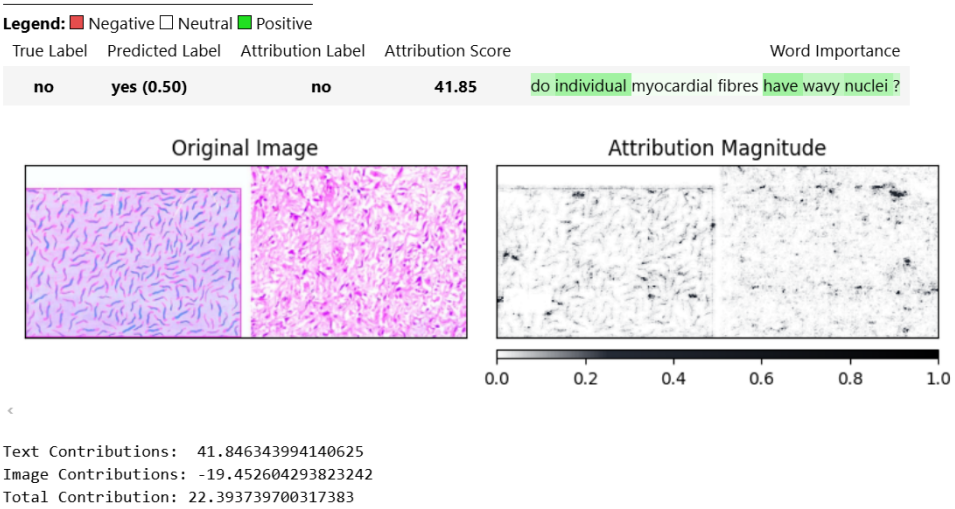


ABBILDUNG A.6: IG für die Frage „Do individual myocardial fibres have wavy nuclei?“

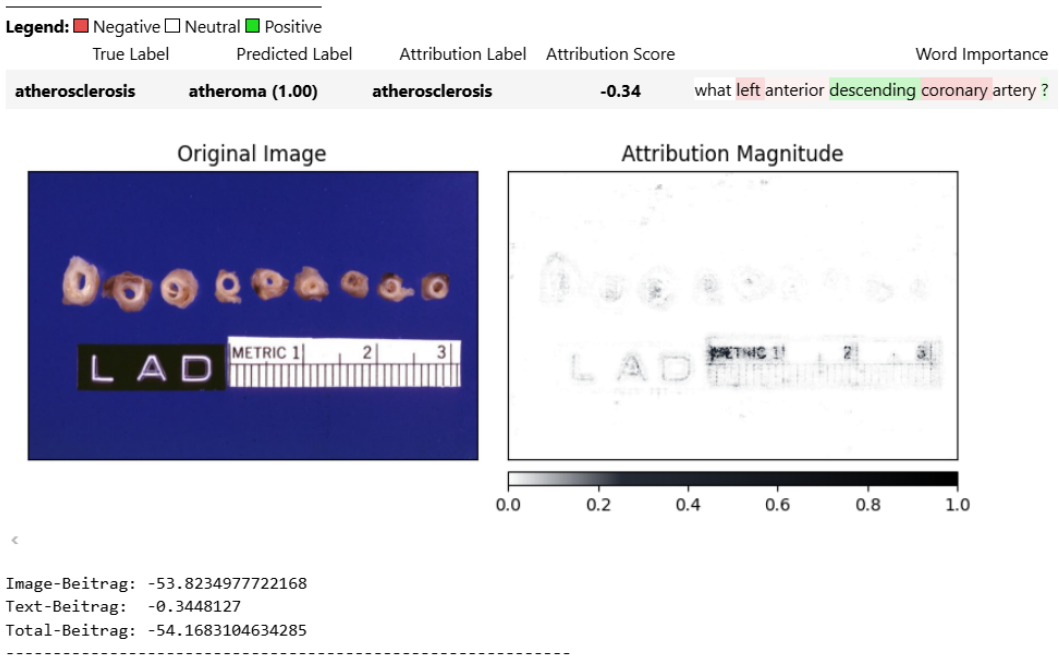


ABBILDUNG A.7: IG für die Frage „What left anterior descending coronary artery?“

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

[1] Wenjie Dong u. a. „Generative Models in Medical Visual Question Answering: A Survey“. In: *Appl. Sci.* 15 (2025), S. 2983. DOI: 10.3390/app15062983. URL: https://doi.org/10.3390/app15062983.

[2] Xuehai He u. a. *PathVQA: 30000+ Questions for Medical Visual Question Answering*. 2020. arXiv: 2003.10286 [cs.CL]. URL: https://arxiv.org/abs/2003.10286.

[3] Stanislaw Antol u. a. „VQA: Visual Question Answering“. In: *Proceedings of the IEEE international conference on computer vision (ICCV)*. 2015, S. 2425–2433.

[4] Dhruv Sharma, Sanjay Purushotham und Chandan K. Reddy. „MedFuseNet: An attention-based multimodal deep learning model for visual question answering in the medical domain“. In: *Scientific Reports* 11 (2021), S. 19826. DOI: 10.1038/s41598-021-98390-1. URL: https://www.nature.com/articles/s41598-021-98390-1.

[5] Jong Hak Moon u. a. „Multi-Modal Understanding and Generation for Medical Images and Text via Vision-Language Pre-Training“. In: *IEEE Journal of Biomedical and Health Informatics* 26.12 (Dez. 2022), S. 6070–6080. ISSN: 2168-2208. DOI: 10.1109/jbhi.2022.3207502. URL: http://dx.doi.org/10.1109/JBHI.2022.3207502.

[6] Qing Li u. a. *Tell-and-Answer: Towards Explainable Visual Question Answering using Attributes and Captions*. 2018. arXiv: 1801.09041 [cs.CV]. URL: https://arxiv.org/abs/1801.09041.

[7] Zhengyuan Yang u. a. „An Empirical Study of GPT-3 for Few-Shot Knowledge-Based VQA“. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.3 (Juni 2022), S. 3081–3089. DOI: 10.1609/aaai.v36i3.20215. URL: https://ojs.aaai.org/index.php/AAAI/article/view/20215.

[8] Igor Sterner u. a. *Few-Shot VQA with Frozen LLMs: A Tale of Two Approaches*. 2024. arXiv: 2403.11317 [cs.CL]. URL: https://arxiv.org/abs/2403.11317.

[9] Usman Naseem, Matloob Khushi und Jinman Kim. „Vision-Language Transformer for Interpretable Pathology Visual Question Answering“. In: *IEEE Journal of Biomedical and Health Informatics* 27.4 (2023), S. 1681–1690. DOI: 10.1109/JBHI.2022.3163751. URL: https://pubmed.ncbi.nlm.nih.gov/35358054/.

[10] Songtao Jiang u. a. „Med-MoE: Mixture of Domain-Specific Experts for Lightweight Medical Vision-Language Models“. In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, Nov. 2024, S. 3843–3860.

[11] Yakoub Bazi u. a. „Vision–Language Model for Visual Question Answering in Medical Imagery“. In: *Bioengineering* 10.3 (2023). ISSN: 2306-5354. DOI: 10.3390/bioengineering10030380. URL: https://www.mdpi.com/2306-5354/10/3/380.

[12] Ashish Vaswani u. a. „Attention Is All You Need“. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Bd. 30. Curran Associates, Inc., 2017, S. 5998–6008. URL: https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[13] Alexey Dosovitskiy u. a. „An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale“. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2021. URL: https://arxiv.org/abs/2010.11929.

[14] Shukang Yin u. a. „A Survey on Multimodal Large Language Models“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023). URL: https://arxiv.org/abs/2306.13549.

[15] Peng Xu, Xiatian Zhu und David A. Clifton. „Multimodal Learning with Transformers: A Survey“. In: *arXiv preprint arXiv:2206.06488* (2023). URL: https://arxiv.org/abs/2206.06488.

[16] Kaiming He u. a. „Deep Residual Learning for Image Recognition“. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, S. 770–778. URL: https://arxiv.org/abs/1512.03385.

[17] Konstantinos Poulinakis. *Multimodal Deep Learning: Definition, Examples, Applications*. Online verfügbar: https://www.v7labs.com/blog/multimodal-deep-learning-guide. 2022. URL: https://www.v7labs.com/blog/multimodal-deep-learning-guide.

[18] Prashant Shrestha u. a. „Medical Vision Language Pretraining: A survey“. In: *arXiv preprint arXiv:2312.06224* (2023). URL: https://arxiv.org/abs/2312.06224.

[19] Edward J. Hu u. a. „LoRA: Low-Rank Adaptation of Large Language Models“. In: *arXiv preprint arXiv:2106.09685* (2021). URL: https://arxiv.org/abs/2106.09685.

[20] Lakera. *What is In-Context Learning, and how does it work: The Beginner's Guide*. https://www.lakera.ai/blog/what-is-in-context-learning. Zugegriffen am 26. Mai 2025. 2024.

[21] Kishore Papineni u. a. „Bleu: a Method for Automatic Evaluation of Machine Translation“. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Hrsg. von Pierre Isabelle, Eugene Charniak und Dekang Lin. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Juli 2002, S. 311–318. DOI: 10.3115/1073083.1073135. URL: https://aclanthology.org/P02-1040/.

[22] Chin-Yew Lin. „ROUGE: A Package for Automatic Evaluation of Summaries“. In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Juli 2004, S. 74–81. URL: https://aclanthology.org/W04-1013/.

[23] Tianyi Zhang u. a. „BERTScore: Evaluating Text Generation with BERT“. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2020. URL: https://arxiv.org/abs/1904.09675.

[24] Yang Liu u. a. „G-EVAL: NLG Evaluation using GPT-4 with Better Human Alignment“. In: *arXiv preprint arXiv:2303.16634* (2023). URL: https://arxiv.org/abs/2303.16634.

[25] Melkamu Mersha u. a. „Explainable Artificial Intelligence: A Survey of Needs, Techniques, Applications, and Future Direction“. In: *arXiv preprint arXiv:2409.00265v2* (2024). Online verfügbar: https://arxiv.org/abs/2409.00265v2. URL: https://arxiv.org/abs/2409.00265v2.

[26] Omri Kaduri, Shai Bagon und Tali Dekel. „What's in the Image? A Deep-Dive into the Vision of Vision Language Models“. In: *arXiv preprint arXiv:2411.17491* (2024). URL: https://arxiv.org/abs/2411.17491.

[27] Mukund Sundararajan, Ankur Taly und Qiqi Yan. „Axiomatic Attribution for Deep Networks“. In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*. PMLR, 2017, S. 3319–3328. URL: https://arxiv.org/abs/1703.01365.

[28] H. Mohan. *Textbook of Pathology*. Jaypee Brothers Medical Publishers Pvt. Limited, 2018. ISBN: 9789352705474. URL: https://books.google.ch/books?id=BPr6ugEACAAJ.

[29] Vinay Kumar, Abul K. Abbas und Jon C. Aster. *Robbins Basic Pathology*. 11. Aufl. Philadelphia: Elsevier, 2023. ISBN: 9780323930005.

[30] *Pathology Education Informational Resource (PEIR) Digital Library*. https://peir.path.uab.edu. Zugegriffen am 28. Mai 2025.

[31] OpenCompass Contributors. *OpenCompass: Open-VLM Leaderboard*. https://huggingface.co/spaces/opencompass/open_vlm_leaderboard. Accessed: 2025-05-31. 2024.

[32] Shuai Bai u. a. *Qwen2.5-VL Technical Report*. 2025. arXiv: 2502.13923 [cs.CV]. URL: https://arxiv.org/abs/2502.13923.

[33] Gemma Team u. a. *Gemma 3 Technical Report*. 2025. arXiv: 2503.19786 [cs.CL]. URL: https://arxiv.org/abs/2503.19786.

[34] Seongyun Lee u. a. „Aligning to Thousands of Preferences via System Message Generalization“. In: *arXiv preprint arXiv:2405.17977* (2024).

[35] Ernie Chang u. a. „Exploring example selection for few-shot text-to-sql semantic parsing“. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 2021, S. 8–13.

[36] Yifan Song u. a. „The Good, The Bad, and The Greedy: Evaluation of LLMs Should Not Ignore Non-Determinism“. In: *arXiv preprint arXiv:2407.10457* (2024). URL: https://arxiv.org/abs/2407.10457.

[37] Oscar Mañas, Benno Krojer und Aishwarya Agrawal. „Improving Automatic VQA Evaluation Using Large Language Models“. In: *arXiv preprint arXiv:2310.02567* (2023). URL: https://arxiv.org/abs/2310.02567.

[38] Takuya Akiba u. a. *Optuna: A Next-generation Hyperparameter Optimization Framework*. 2019. arXiv: 1907.10902 [cs.LG]. URL: https://arxiv.org/abs/1907.10902.

[39] Google DeepMind. *Gemma Vision QLoRA Fine-Tuning Guide*. https://ai.google.dev/gemma/docs/core/huggingface_vision_finetune_qlora?hl=de. 2024.

[40] N. Kokhlikyan u. a. „Captum: A Unified and Generic Model Interpretability Library for PyTorch“. In: *arXiv preprint arXiv:2009.07896* (2020).

[41] Wenhui Zhu u. a. „Toward Effective Reinforcement Learning Fine-Tuning for Medical VQA in Vision-Language Models“. In: *arXiv preprint arXiv:2505.13973* (2025). URL: https://arxiv.org/abs/2505.13973.