



UNIVERSIDADE FEDERAL DO MARANHÃO

UNIVERSIDADE FEDERAL DO PIAUÍ

Doutorado em Ciência da Computação Associação UFMA/UFPI

Gilvan Veras Magalhães Junior

XRaySwinGen: Pré-Laudos Médicos Automáticos para Exames de Raio X de Tórax com Modelo Multimodal

Orientador: Prof. Dr. Pedro de Alcântara dos Santos Neto

Co-orientador: Prof. Dr. Anselmo Cardoso de Paiva

Teresina - PI
Agosto, 2025

Gilvan Veras Magalhães Junior

XRaySwinGen: Pré-Laudos Médicos Automáticos para Exames de Raio X de Tórax com Modelo Multimodal

DEFESA DE TESE DE DOUTORADO

Tese apresentada como requisito parcial para obtenção do título de Doutor em Ciência da Computação, ao Doutorado em Ciência da Computação, Associação UFMA/UFPI.

Orientador: Prof. Dr. Pedro de Alcântara dos Santos Neto
Coorientador: Prof. Dr. Anselmo Cardoso de Paiva

Teresina - PI
Agosto, 2025

FICHA CATALOGRÁFICA
Universidade Federal do Piauí
Sistema de Bibliotecas UFPI - SIBi/UFPI
Biblioteca Setorial do CCN

M189x Magalhães Júnior, Gilvan Veras.
XRaySwinGen: pré-laudos médicos automáticos para exames de Raio X de tórax com modelo multimodal / Gilvan Veras Magalhães Júnior. -- 2025.
68 f. : il. color.

Tese (Doutorado) - Universidade Federal do Piauí. Centro de Ciências da Natureza. Programa de Pós-Graduação em Ciência da Computação, Teresina, 2025.
"Orientador: Prof. Dr. Pedro de Alcântara dos Santos Neto. Coorientador: Prof. Dr. Anselmo Cardoso de Paiva."

1. Visão Computacional. 2. Processamento de Linguagem Neural. 3. Laudo Médico. 4. *Swin Transformer*. I. Santos Neto, Pedro Alcântara dos. II. Paiva, Anselmo Cardoso. III. Título.

CDD 004.77

Bibliotecária: Caryne Maria da Silva Gomes - CRB3/1461

Documento assinado digitalmente
GILVAN VERAS MAGALHAES JUNIOR
Data: 01/10/2025 12:56:28-0300
Verifique em https://validar.iti.gov.br

Gilvan Veras Magalhães Junior

XRaySwinGen: Pré-Laudos Médicos Automáticos para Exames de Raio X de Tórax com Modelo Multimodal

A presente Tese de Doutorado foi avaliada e aprovada por banca examinadora composta pelos seguintes membros:

Prof. Dr. Pedro de Alcântara dos Santos Neto
Orientador
Universidade Federal do Piauí

Prof. Dr. Anselmo Cardoso de Paiva
Co-orientador
Universidade Federal do Maranhão

Prof. Dr. António Manuel Trigueiros da Silva Cunha
Universidade de Trás-os-Montes e Alto Douro

Prof. Dr. Cláudio de Souza Baptista
Universidade Federal de Campina Grande

Prof. Dr. Rodrigo De Melo Souza Veras
Universidade Federal do Piauí

Prof. Dr. Kelson Romulo Teixeira Aires
Universidade Federal do Piauí

Certificamos que esta é a versão original e final da Tese de Doutorado que foi julgada adequada para obtenção do título de Doutora em Ciência da Computação.

Documento assinado digitalmente
PEDRO DE ALCANTARA DOS SANTOS NETO
Data: 02/10/2025 06:20:35-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Pedro de Alcântara dos Santos Neto
Orientador

Documento assinado digitalmente
ANSELMO CARDOSO DE PAIVA
Data: 02/10/2025 15:31:58-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Anselmo Cardoso de Paiva
Coordenador

Teresina - PI, 20 de Agosto de 2025

Aos meus pais Francisco Gilvan e Verbena Elane, à minha esposa Larissa Magalhães e ao meu filho Gilvan Magalhães por sempre terem me dado força para persistir.

Agradecimentos

Agradecimentos

Agradeço, primeiramente, a Deus, por iluminar meus caminhos e fortalecer minha jornada.

Às minhas maiores inspirações, meus pais, Francisco Gilvan e Verbena Elane, agradeço profundamente pela educação exemplar, pelo amor incondicional e por sempre acreditarem no meu potencial, oferecendo o melhor de si em cada etapa da minha vida.

À minha esposa, Larissa Magalhães, e ao meu filho, Gilvan Magalhães, minha gratidão eterna pelo amor, incentivo e compreensão. O apoio de vocês foi essencial para que eu pudesse seguir com determinação e alcançar meus objetivos pessoais e acadêmicos.

À minha querida família, deixo o mais sincero agradecimento pelo carinho, pela fé constante em meu caminho e por todo o suporte oferecido ao longo dessa trajetória.

Ao meu orientador, Prof. Dr. Pedro de Alcântara dos Santos Neto, sou imensamente grato pelos ensinamentos, orientações, paciência e apoio, tanto acadêmico quanto pessoal. Ao meu coorientador, Prof. Dr. Anselmo Cardoso de Paiva, agradeço pelas valiosas contribuições, pelas discussões instigantes e pelos questionamentos precisos, que muito contribuíram para meu amadurecimento acadêmico e científico. O trabalho conjunto de ambos foi fundamental para a obtenção de resultados significativos nesta pesquisa. Meu reconhecimento e gratidão pela dedicação e apoio constantes.

Aos colegas do programa de doutorado, agradeço pela amizade, parceria e convivência ao longo dessa jornada. Ao amigo Dr. Luis Henrique, sou grato pela generosidade ao compartilhar seus conhecimentos e experiências. Um agradecimento especial ao Prof. Dr. Roney Lira, pela amizade, pelas discussões enriquecedoras, pelas colaborações acadêmicas e pelos valiosos conselhos.

A todos os docentes do Programa de Pós-Graduação em Ciência da Computação Associação UFMA/UFPI, expresso meu sincero agradecimento.

“Cedo ou tarde,
você vai aprender,
assim como eu aprendi,
que existe uma diferença entre
CONHECER o caminho
e TRILHAR o caminho.”
(Morpheus)

Resumo

A radiologia tem papel crucial na medicina moderna ao fornecer diagnósticos precisos por meio de imagens não invasivas. Entretanto, a elaboração manual de laudos médicos é um processo demorado e sujeito a falhas humanas. Esta tese propõe um modelo multimodal para a geração automática de pré-laudos médicos a partir de radiografias de tórax, combinando técnicas de Visão Computacional e Processamento de Linguagem Natural com base na arquitetura Transformer. Inicialmente, foi desenvolvida uma abordagem com codificador visual baseado no Swin Transformer e decodificador textual integrando camadas de atenção cruzada e treinamento bilíngue com conjuntos de dados em Português PT-BR ou Inglês. Posteriormente, a arquitetura foi aprimorada com a introdução de um módulo de memória relacional, permitindo a retenção de informações contextuais de longo prazo durante a geração dos textos. O modelo final integra de forma coesa os componentes visuais e textuais por meio de normalização condicional orientada à memória. Os experimentos, realizados nas bases de imagens Proposta, IU Chest X-ray, NIH Chest X-ray e MIMIC-CXR-JPG, evidenciaram ganhos consistentes. Na avaliação com o conjunto de dados completo da MIMIC-CXR-JPG, o modelo com Swin Transformer e decodificador textual atingiu ROUGE-L de 0.304, METEOR de 0.233 e BLEU-4 de 0.054. A inclusão da memória relacional elevou essas métricas para 0.321, 0.281 e 0.114, respectivamente. Na versão do mesmo conjunto de dados sem o histórico clínico, o desempenho do modelo com memória relacional foi ainda maior, alcançando ROUGE-L de 0.416, METEOR de 0.384 e BLEU-4 de 0.187. A melhora consistente das métricas com a adição do módulo de memória relacional demonstra o impacto positivo da retenção de contexto de longo prazo na qualidade textual dos pré-laudos gerados. Esses resultados reforçam a relevância do modelo proposto e motivam sua adoção em cenários clínicos que demandam precisão, fluidez e confiabilidade na geração automática de relatórios médicos.

Palavras-chave: Visão Computacional, Pré-Laudo Médico, Multimodal, Processamento de Linguagem Natural, *Transformers*

Abstract

Radiology plays a crucial role in modern medicine by providing accurate diagnoses through non-invasive imaging. However, the manual creation of medical reports is a time-consuming process and prone to human error. This thesis proposes a multimodal model for the automatic generation of preliminary medical reports from chest radiographs, combining Computer Vision and Natural Language Processing techniques based on the Transformer architecture. Initially, an approach was developed with a visual encoder based on the Swin Transformer and a textual decoder integrating cross-attention layers and bilingual training with datasets in Brazilian Portuguese (PT-BR) or English. Subsequently, the architecture was enhanced with the introduction of a relational memory module, enabling the retention of long-term contextual information during text generation. The final model cohesively integrates the visual and textual components through memory-oriented conditional normalization. The experiments, conducted on the Proposta, IU Chest X-ray, NIH Chest X-ray, and MIMIC-CXR-JPG image datasets, showed consistent gains. In the evaluation with the complete MIMIC-CXR-JPG dataset, the model with the Swin Transformer and textual decoder achieved a ROUGE-L of 0.304, a METEOR of 0.233, and a BLEU-4 of 0.054. The inclusion of the relational memory module raised these metrics to 0.321, 0.281, and 0.114, respectively. In the version of the same dataset without clinical history, the performance of the model with relational memory was even higher, reaching a ROUGE-L of 0.416, a METEOR of 0.384, and a BLEU-4 of 0.187. The consistent improvement in metrics with the addition of the relational memory module demonstrates the positive impact of long-term context retention on the textual quality of the generated pre-reports. These results reinforce the relevance of the proposed model and encourage its adoption in clinical scenarios that demand accuracy, fluidity, and reliability in the automatic generation of medical reports.

Keywords: Computer Vision, Medical Report, Multimodal, Natural Language Processing, Transformers

Lista de ilustrações

Figura 1 – Arquitetura *Transformer* codificador-decodificador. 31
Figura 2 – (a) *Self-Attention*. (b) *Multi-head attention*. O MSA consiste em vários blocos SA (*heads*) concatenados juntos por canal como proposto em Vaswani et al. (2017) 33
Figura 3 – Arquitetura ViT e a estrutura do seu bloco codificador. *Patches* de imagens sequenciais são usados como entrada e processados por um codificador para produzir a saída de classificação final. 36
Figura 4 – Arquitetura do modelo *XRaySwinGen* para geração automática de pré-laudos médicos. 40
Figura 5 – Módulos da arquitetura do *Swin Transformer*. 42
Figura 6 – Arquitetura do modelo MMMR proposto. 44
Figura 7 – Comparativo visual do desempenho entre os modelos *XRaySwinGen* e MMMR nos cenários *MIMIC-CXR-JPG* com e sem histórico clínico, considerando as métricas BLEU-4, ROUGE-L, METEOR e SPICE. . 53
Figura 8 – *Webpage* criada para prova de conceito dos modelos *XRaySwinGen* e MMMR. 57

Lista de tabelas

Tabela 1 – Resumo dos trabalhos relacionados. 29
Tabela 2 – Mapeamento de classes *NIH Chest X-ray*. 47
Tabela 3 – Resumo das bases de dados utilizadas. 47
Tabela 4 – Resultados obtidos pelos modelos *XRaySwinGen* e MMMR em diferentes bases de dados. 50
Tabela 5 – Comparação dos resultados dos modelos MMMR (com variação de temperatura), *XRaySwinGen* e de outros autores em bases *MIMIC-CXR* e *MIMIC-CXR-JPG*. 51
Tabela 6 – Exemplos de laudos gerados pelo modelo *MMMR* em comparação com o laudo de referência (*ground truth*). 54

Lista de abreviaturas e siglas

AMC	Atenção Multi Cabeça
BERT	*Bidirectional Encoder Representations from Transformers*
BLEU	*Bilingual Evaluation Understudy*
BPE	*Byte-Pair Encodings*
CLM	*Causal Language Modeling*
CRM	Conselho Federal de Medicina
TC	*Tomografia Computadorizada*
GELU	*Gaussian Error Linear Unit*
GPT-2	*Generative Pretrained Transformer 2*
GPUs	Unidades de Processamento Gráfico
IA	Inteligência Artificial
LN	*LayerNorm*
METEOR	*Metric for Evaluation of Translation with Explicit ORdering*
MLM	*Masked Language Model*
MLP	*Multilayer Perceptron*
MRI	*Magnetic Resonance Imaging*
MSA	*Multi-head Self-Attention*
NSP	*Next Sentence Prediction*
PACS	*Picture Archiving and Communication System*
PLN	Processamento de Linguagem Natural
PMCID	*PubMed Central Identifier*
PP	Partição de Partes
RIS	*Radiology Information System*

RNC	Redes Neurais Convolucionais
RNR	Redes Neurais Recorrentes
ROUGE	*Recall-Oriented Understudy for Gisting Evaluation*
RX	Raio-X
SPICE	*Semantic Propositional Image Caption Evaluation*
SW-MSA	*Shifted Window MSA*
T5	*Text-to-Text Transfer Transformer*
TPUs	Unidades de Processamento de Tensor
VC	Visão Computacional
ViT	*Vision Transformer*
W-MSA	*Window MSA*

Sumário

1 INTRODUÇÃO 15
1.1 Motivação 16
1.2 Objetivo 18
1.3 Hipóteses 18
1.4 Contribuições 19
1.5 Organização 20

2 TRABALHOS RELACIONADOS 22
2.1 Arquiteturas baseadas em Rede Neural Convolucional e Rede Neural Recorrente 23
2.2 Arquiteturas baseadas em *Transformers* 25
2.3 Considerações Finais 28

3 FUNDAMENTAÇÃO TEÓRICA 30
3.1 *Transformers* 30
3.1.1 *Vision Transformer* 34
3.1.2 Módulo de Memória Relacional 35
3.2 Métricas de Avaliação 37
3.3 Considerações Finais 38

4 METODOLOGIA 39
4.1 *XRaySwinGen* 39
4.2 Codificador - *Swin Transformer* 40
4.3 Decodificador - GPT-2 *Transformer* 42
4.4 Modelo Multimodal de Memória Relacional: MMMR 43
4.5 Considerações Finais 45

5 EXPERIMENTOS E RESULTADOS 46
5.1 Bases de dados 46
5.2 Configurações de Treinamento 48
5.2.1 *XRaySwinGen*: Configurações Arquiteturais e de Treinamento 48
5.2.2 MMMR: Configuração Arquitetural e Treinamento 48
5.3 Resultados e Discussões 50
5.4 Considerações Finais 53

6 CONCLUSÃO E TRABALHOS FUTUROS 56
6.1 Contribuições 58

6.2 Trabalhos Futuros 59
6.3 Considerações Éticas e Limitações Clínicas 59
6.4 Produção Científica 60

REFERÊNCIAS 61

1 Introdução

A Inteligência Artificial (IA) tem desempenhado um papel central no avanço da medicina moderna, especialmente na radiologia, onde a análise de imagens é essencial para diagnósticos precisos e não invasivos. Contudo, a geração manual de laudos médicos ainda representa um gargalo significativo na prática clínica. O processo é inerentemente demorado, exige atenção constante e julgamento clínico apurado por parte dos profissionais, estando ainda sujeito a erros humanos e a uma variabilidade terminológica que frequentemente resulta em laudos não padronizados. Tal cenário representa o desafio central que motivou o desenvolvimento desta pesquisa.

A superação deste gargalo exige uma abordagem que transcenda a análise de imagem isolada, demandando uma solução capaz de realizar a complexa tarefa de tradução entre domínios distintos: o visual e o textual. Tal abordagem visa otimizar o fluxo de trabalho radiológico, oferecendo ao especialista um ponto de partida estruturado e coerente, com o potencial de acelerar a entrega de diagnósticos, reduzir a carga de trabalho e elevar a padronização e a qualidade dos relatórios clínicos.

Diante disso, esta tese investiga a automação da geração de pré-laudos médicos a partir de radiografias de tórax, explorando como os avanços em Visão Computacional e Processamento de Linguagem Natural poderiam ser integrados de forma sinérgica. A premissa fundamental era que uma arquitetura multimodal, projetada para compreender tanto as nuances das imagens radiológicas quanto a semântica da linguagem clínica, poderia otimizar este processo. Nesse contexto, a eficácia da arquitetura *Transformer* (VASWANI et al., 2017) em capturar relações contextuais complexas em dados sequenciais a estabeleceu como a base teórica mais promissora para o desenvolvimento de tal solução.

A primeira materialização desta hipótese foi a concepção e implementação da arquitetura **XRaySwinGen**. Este modelo foi metodicamente projetado para servir como uma robusta linha de base, combinando um codificador visual hierárquico, o *Swin Transformer* (LIU et al., 2021), escolhido por sua eficiência em capturar características visuais em múltiplas escalas, com um decodificador textual generativo, baseado em GPT (RADFORD et al., 2019). A integração entre as modalidades foi viabilizada por mecanismos de atenção cruzada, com o objetivo de criar um sistema de ponta a ponta capaz de traduzir as representações visuais de uma radiografia de tórax em um pré-laudo médico coerente e informativo.

Contudo, apesar de atingir o objetivo inicial de realizar a tradução intermodal, a avaliação sistemática do *XRaySwinGen* revelou uma limitação crítica. Embora o modelo

fosse capaz de gerar textos clinicamente relevantes, ele apresentava dificuldades em reter e propagar dependências semânticas ao longo de sequências extensas. Essa deficiência se manifestava em laudos que, por vezes, careciam de fluidez e consistência narrativa, especialmente em descrições que exigiam referência a múltiplos achados. A análise desses resultados tornou claro que um simples alinhamento entre imagem e texto era insuficiente; para alcançar um nível de qualidade superior, era imperativo dotar a arquitetura de uma memória contextual explícita.

A constatação dessa necessidade direcionou a pesquisa para uma nova e decisiva fase, focada no aprimoramento arquitetural para endereçar a questão da coerência textual. O resultado desta etapa foi o **Modelo Multimodal de Memória Relacional (MMMR)**, que representa a principal contribuição desta tese. Ao integrar um módulo de memória relacional explícito na estrutura do decodificador, o MMMR adquiriu a capacidade de armazenar, inter-relacionar e atualizar informações contextuais a cada passo da geração. Este aprimoramento, complementado por mecanismos de normalização condicional e portões que regulam dinamicamente o fluxo de informação, permitiu a produção de pré-laudos significativamente mais fluidos, consistentes e semanticamente ricos, superando a limitação fundamental da proposta anterior.

Esta linha de pesquisa, que parte de um desafio clínico bem definido, avança pela criação e avaliação crítica de uma proposta inicial e culmina no desenvolvimento de uma solução aprimorada e inovadora, forma a espinha dorsal desta tese. Nos capítulos seguintes, cada etapa deste processo é rigorosamente detalhada, apresentando os fundamentos teóricos que embasaram as decisões, as metodologias empregadas na construção dos modelos, os experimentos conduzidos para sua validação e os resultados quantitativos que demonstram a eficácia e a relevância da abordagem final.

1.1 Motivação

A adoção de sistemas automáticos para geração de pré-laudos radiológicos representa um avanço significativo rumo a uma prática médica mais eficaz, acessível e segura. A radiologia, por suas técnicas de imagem não invasivas, ocupa posição central na medicina moderna, permitindo o diagnóstico e monitoramento de diversas condições clínicas com alta precisão (ZHOU; RUECKERT; FICHTINGER, 2019). No entanto, a elaboração manual de laudos exige tempo, concentração e julgamento clínico cuidadoso por parte dos radiologistas, que frequentemente enfrentam sobrecarga de trabalho, longas jornadas e escassez de profissionais, especialmente em regiões remotas ou durante plantões com equipe reduzida. Essas condições ampliam o risco de fadiga, inconsistências terminológicas, erros humanos e atrasos no diagnóstico, comprometendo diretamente a segurança do paciente (ZHAO et al., 2023).

A aplicação de técnicas de IA à imagiologia médica tem impulsionado avanços

expressivos, especialmente em tarefas de classificação de achados radiológicos (KIM et al., 2022; ZAIDI et al., 2022; TIAN et al., 2022; NAWAZ et al., 2023; KANDASAMY, 2023; SWATHI; BAI, 2025). Esses sistemas apresentam elevado desempenho na identificação e categorização de patologias, contribuindo significativamente para triagens automatizadas e apoio à decisão clínica. Contudo, a prática radiológica vai além da detecção de padrões isolados: a construção de laudos clínicos exige articulação descritiva de observações imagiológicas, como localização, morfologia, dimensões e padrões, em linguagem textual padronizada, precisa e interpretativa (ÇALLI et al., 2021).

Essa lacuna entre a detecção automatizada e a expressão clínica interpretativa impulsionou o surgimento de um campo emergente: a geração automática de laudos radiológicos em linguagem natural (XIONG; DU; YAN, 2019; LIU et al., 2021; YOU et al., 2021; ALFARGHALY et al., 2021; MOHSAN et al., 2022; YU et al., 2023). A proposta desses modelos é sintetizar informações visuais em descrições clínicas relevantes, coerentes e contextualizadas. Além de reduzir a carga cognitiva dos radiologistas, tal abordagem pode otimizar o fluxo de trabalho, padronizar terminologias e democratizar o acesso a diagnósticos de qualidade, inclusive em cenários com restrição de recursos.

(CHEN et al., 2020; JING; XIE; XING, 2018; LI et al., 2018; JOHNSON et al., 2019; JING; WANG; XING, 2020),

Apesar dos avanços já alcançados, o desafio de produzir laudos automáticos que combinem acurácia diagnóstica, riqueza descritiva e confiabilidade clínica ainda permanece. Este desafio é abordado por linhas de pesquisa recentes que focam no aprimoramento do alinhamento multimodal (CHEDDI; HABBANI; NAIT-CHARIF, 2024a), na consistência entre múltiplas vistas de imagem (WANG et al., 2026) e na injeção de conhecimento externo para aumentar a precisão (HOU et al., 2025). A formulação de pré-laudos não estruturados demanda competências linguísticas complexas, além de capacidades visuais, um ponto em que técnicas modernas de geração automática de texto têm ganho destaque (FATIMA et al., 2022; BISWAS, 2023). Adicionalmente, avanços recentes na interpretação automatizada de imagens médicas (KAUR et al., 2021; YANG; TANAKA; ISHIDA, 2023; SHARMA et al., 2025) e na automação de relatórios clínicos (DICUONZO et al., 2023; SAHNI et al., 2023; FYTAS et al., 2024) reforçam o potencial dessas soluções como alternativas escaláveis e integráveis à prática médica contemporânea.

A motivação desta tese reside na construção de uma solução computacional que transite entre a análise visual e a expressão linguística de forma coerente, flexível e precisa. Ao adotar uma abordagem multimodal com suporte à retenção de memória contextual, almeja-se aproximar a geração automática de pré-laudos das necessidades reais dos sistemas de saúde, promovendo diagnósticos mais ágeis, padronizados e confiáveis.

1.2 Objetivo

O objetivo principal desta tese é propor, desenvolver e avaliar uma arquitetura multimodal baseada em *Transformers* (VASWANI et al., 2017) para a geração automática de pré-laudos médicos textuais a partir de radiografias de tórax. A proposta é concebida com o propósito de oferecer uma ferramenta computacional robusta que possa auxiliar radiologistas na tarefa de elaboração de laudos, especialmente em contextos com alta demanda, recursos limitados ou necessidade de triagem rápida, defendendo a tese de que a incorporação de mecanismos de memória explícita é fundamental para elevar a coerência e a fidelidade clínica desses sistemas.

Para alcançar esse objetivo geral, foram definidos os seguintes objetivos específicos:

- Aplicar técnicas de Processamento de Linguagem Natural aos laudos médicos, como normalização textual, remoção de entidades não relacionadas ao paciente e caracteres estranhos, e tokenização para que haja o entendimento do laudo em linguagem computacional;
- Projetar uma arquitetura multimodal que integre um codificador visual e um decodificador textual, com atenção cruzada entre modalidades;
- Incorporar um módulo de memória relacional à arquitetura, capaz de capturar dependências de longo prazo e regular a retenção e atualização de contexto textual;
- Aplicar validação cruzada com *k-fold* para garantir rigor estatístico na avaliação dos modelos em diferentes divisões de dados;
- Avaliar o desempenho do modelo utilizando métricas amplamente aceitas em geração de linguagem natural, como BLEU, ROUGE-L, METEOR e SPICE, em bases públicas como a *MIMIC-CXR-JPG*, *IU Chest X-ray* e *NIH Chest X-ray*.

1.3 Hipóteses

As hipóteses que conduzem os objetivos apresentados nesta tese são:

- É possível superar as limitações dos sistemas existentes de geração de laudos automáticos através do desenvolvimento de uma arquitetura de aprendizagem profunda multimodal. Tal arquitetura, ao aprender a mapear de forma coesa e contextual as características visuais hierárquicas de uma radiografia de tórax para a estrutura sequencial e complexa de um relatório médico, pode gerar pré-laudos que sejam não apenas linguisticamente fluentes, mas também clinicamente mais precisos e confiáveis.

- A coerência narrativa e a retenção de contexto em laudos longos podem ser significativamente aprimoradas pela incorporação de um módulo de memória relacional explícito na arquitetura, superando as limitações de modelos que dependem apenas de mecanismos de atenção para capturar dependências de longo prazo.

A partir destas hipóteses, a tese defendida e comprovada neste trabalho é que a incorporação de um módulo de memória relacional a uma arquitetura multimodal baseada em *Transformers* aumenta significativamente a coerência contextual e a fidelidade clínica na geração automática de pré-laudos médicos, superando abordagens que dependem exclusivamente de mecanismos de atenção para o alinhamento intermodal.

1.4 Contribuições

Esta tese apresenta uma contribuição original e significativa no campo da geração automática de pré-laudos médicos, propondo e avaliando arquiteturas multimodais baseadas em *Transformers* para sintetizar informações visuais oriundas de radiografias de tórax em linguagem médica descritiva. O trabalho foi conduzido com ênfase na eficiência computacional, na qualidade semântica dos textos gerados e na aplicabilidade clínica do modelo proposto.

Inicialmente, foi concebida uma modelagem do *Swin Transformer* como codificador visual, explorando sua estrutura hierárquica de autoatenção com janelas deslocadas (shifted windows) para representar imagens médicas com granularidade local e global. Essa configuração possibilitou o desenvolvimento de um extrator de características específicas para imagens de RX, com alto poder de representação sem aumento significativo de complexidade. Em conjunto, utilizou-se o GPT-2 como decodificador textual, incorporando camadas de atenção cruzada conectadas diretamente às representações visuais, o que permitiu uma integração eficaz entre os domínios visual e linguístico.

Durante o processo de investigação, identificaram-se limitações em termos de retenção de contexto e fluidez textual nas gerações. Em resposta a isso, a arquitetura foi estendida com a introdução de um módulo de memória relacional, um componente que armazena e atualiza informações sequenciais ao longo do tempo, contribuindo para a geração de textos mais consistentes, informativos e contextualizados. Essa evolução resultou na proposta do Modelo Multimodal de Memória Relacional, que se destaca por sua capacidade de capturar dependências de longo prazo por meio de mecanismos de atenção múltipla, portões de entrada/esquecimento e normalização condicional orientada à memória.

Outro diferencial relevante desta tese está na eficiência computacional da arquitetura: mesmo com 237 milhões de parâmetros, o modelo obteve desempenho

competitivo em relação a abordagens muito maiores, que frequentemente ultrapassam 1 bilhão de parâmetros, como LLMs aplicadas à medicina. Esse equilíbrio entre desempenho e custo computacional reforça a aplicabilidade do modelo em cenários reais, inclusive em ambientes com restrição de recursos computacionais.

Por fim, uma interface web foi desenvolvida como prova de conceito, permitindo testar o modelo com imagens reais e verificar sua capacidade de geração textual em tempo real.

As principais contribuições desta tese são:

- A criação de um extrator visual específico para RX;
- Arquitetura *XRaySwinGen* proposta com espinha dorsal *Swin Transformer base* e GPT-2 para geração automática de pré-laudos médicos;
- Arquitetura MMMR proposta com espinha dorsal *Swin Transformer tiny*, decodificador baseado na arquitetura GPT e módulo de memória relacional.
- A utilização de técnicas baseada no módulo *shift window* (deslocamento de janela) (LIU et al., 2021) para extrair atributos globais e locais das imagens, a qual não tem sido usada em trabalhos da área para gerar automaticamente pré-laudos textuais;
- A obtenção de um modelo menor (237M parâmetros) com desempenho comparável ou superior a modelos com bilhões de parâmetros;
- A disponibilização de uma ferramenta para a análise de pré-laudos gerados pelo modelo multimodal.

Além das contribuições técnicas, esta tese gerou produção científica relevante, com a publicação de dois artigos em canais de visibilidade internacional. O artigo *"XRaySwinGen: Automatic medical reporting for X-ray exams with multimodal model"*, publicado na revista *Heliyon* (*Elsevier*) e o artigo *"Relational Memory Multimodal Models for Automated X-Ray Medical Reporting"*, publicado na conferência internacional *International Conference on Health and Social Care Information Systems and Technologies* (HCist).

1.5 Organização

Esta tese está estruturada em seis capítulos, além das referências bibliográficas, apêndices e anexos. A seguir, descreve-se brevemente o conteúdo de cada capítulo:

No Capítulo 2, **Trabalhos Relacionados**, é realizada uma revisão da literatura, abordando os principais métodos e abordagens utilizados na geração automática de laudos médicos a partir de imagens radiológicas.

No Capítulo 3, **Fundamentação Teórica**, são descritas as principais técnicas e conceitos que embasam esta pesquisa, incluindo Visão Computacional, Processamento de Linguagem Natural e arquiteturas baseadas em *Transformers*.

No Capítulo 4, **Metodologia**, são detalhados os modelos computacionais propostos, os processos de codificação de imagens, geração textual e a evolução arquitetural que culminou na proposta final do modelo multimodal com memória relacional.

No Capítulo 5, **Experimentos e Resultados**, são apresentados os conjuntos de dados utilizados, os procedimentos de treinamento e validação, as métricas de avaliação e os resultados obtidos em diferentes cenários experimentais.

Por fim, no Capítulo 6, **Conclusão e Trabalhos Futuros**, são discutidas as conclusões da pesquisa, suas limitações, as lições aprendidas e as perspectivas para investigações futuras.

2 Trabalhos Relacionados

Neste capítulo é apresentada uma revisão da literatura com o objetivo de identificar, analisar e comparar as principais técnicas de aprendizagem profunda que vêm sendo aplicadas à tarefa de geração automática de pré-laudos médicos textuais a partir de exames de imagem, especialmente radiografias de tórax. A revisão visa mapear o estado da arte na área, com foco em modelos que exploram arquiteturas neurais para extrair informações visuais e convertê-las em linguagem natural clínica coerente, avaliando suas metodologias, limitações, desempenho e potencial de generalização.

A busca foi conduzida nas bases de dados *ScienceDirect*, *IEEE Xplore* e *ArXiv*, utilizando os termos "*medical report generation*"e “*chest x-ray medical report generation*”. As pesquisas foram refinadas para as áreas de *Computer Science* e *Engineering*, resultando em 1.377 artigos para o primeiro termo e 247 para o segundo. A seleção considerou o título, o resumo e a relevância dos estudos em relação aos objetivos deste trabalho. Além disso, a palavra-chave *"Computer Vision"* foi incluída como critério adicional nas buscas, ampliando o escopo para incluir abordagens visuais aplicadas à geração automática de laudos médicos.

A análise do material coletado evidenciou que a maioria das soluções propõe arquiteturas baseadas na integração entre Redes Neurais Convolucionais, responsáveis pela extração de atributos visuais, e Redes Neurais Recorrentes, utilizadas como decodificadores textuais. Mais recentemente, houve um movimento crescente em direção a modelos baseados em *Transformers*, que oferecem vantagens no tratamento de sequências longas e na paralelização do treinamento. Além disso, observou-se o surgimento de arquiteturas multimodais que combinam diferentes técnicas a fontes de entrada de imagem e texto em pipelines de aprendizado conjunto, dentre elas: atenção cruzada para alinhar os domínios visual e linguístico; a exploração de modelos de visão-linguagem pré-treinados (GOWDA et al., 2025); injeção de conhecimento externo (HOU et al., 2025) e introduzem mecanismos que ampliam a explicabilidade (SWATHI; BAI, 2025; KANDASAMY, 2023).

Esta revisão da literatura também identificou um número crescente de trabalhos que incorporam *prompt learning*, pré-treinamento em larga escala, aprendizagem auto-supervisionada e mecanismos de memória, ainda que poucos explorem explicitamente módulos de memória relacional dedicados ao controle do estado contextual durante a geração de texto, um diferencial central desta tese. As subseções a seguir apresentam os principais paradigmas investigados na literatura, com ênfase na sua evolução arquitetural, desempenho em *benchmarks* públicas e relação com a proposta desenvolvida neste

trabalho.

2.1 Arquiteturas baseadas em Rede Neural Convolucional e Rede Neural Recorrente

As Redes Neurais Convolucionais têm desempenhado um papel essencial na análise de imagens médicas ao longo dos últimos anos, sendo amplamente utilizadas para extrair características visuais de exames radiológicos em tarefas de classificação e, mais recentemente, na geração automática de laudos médicos (JING; XIE; XING, 2018; ALLAOUZI et al., 2018; LI et al., 2018; WANG et al., 2018; XIONG; DU; YAN, 2019; HUANG et al., 2019; JING; WANG; XING, 2019; GAJBHIYE; NANDEDKAR; FAYE, 2020; MONSHI; POON; CHUNG, 2020; LIU et al., 2021; YOU et al., 2021; ALFARGHALY et al., 2021; PANDEY et al., 2022; KUMAR et al., 2022a; KUMAR et al., 2022b; SWATHI; BAI, 2025). Esses modelos têm sido empregados como extratores de características visuais em sistemas de geração automática de texto, geralmente combinados com Redes Neurais Recorrentes, como as Memórias de Longo Prazo, que assumem a função de decodificadores textuais. Essa combinação arquitetural RNC-RNR deu origem aos primeiros sistemas de geração de laudos médicos a partir de imagens, marcando uma etapa inicial e significativa na evolução da área.

Embora a geração automática de descrições para imagens médicas já fosse objeto de estudo em anos anteriores, com iniciativas baseadas em predição de legendas genéricas adaptadas para o domínio radiológico (JING; XIE; XING, 2018), o trabalho de (ALLAOUZI et al., 2018) representa um marco inicial na aplicação da arquitetura codificador-decodificador especificamente voltada à geração de pré-laudos médicos. Nesse modelo, Redes Neurais Convolucionais foram utilizadas para a codificação das imagens, enquanto Redes Neurais Recorrentes desempenharam o papel de decodificadores textuais. Posteriormente, a introdução do mecanismo de atenção proposto por (VASWANI et al., 2017) ampliou o potencial dessas arquiteturas, sendo incorporada a experimentos voltados ao alinhamento entre regiões de interesse na imagem e partes do pré-laudo médico. Nesse contexto, (YUAN et al., 2019) propuseram um modelo com mecanismos de atenção hierárquica em nível de sentença, combinando características visuais multivisão com fusão tardia, com o objetivo de melhorar a coesão semântica das descrições clínicas.

Demonstrando a contínua relevância das abordagens modulares, o trabalho de Muharram et al. (2023) propôs um sistema que utiliza múltiplos modelos de classificação binária, onde cada um é responsável por detectar uma anormalidade específica. O resultado combinado dessas classificações gera um código que, por sua vez, seleciona sentenças pré-determinadas para construir o laudo final, uma abordagem que, embora

não seja ponta a ponta, busca aumentar a precisão em achados específicos. Em um enfoque complementar, não na geração de laudos, mas na sua classificação e extração de conceitos, o CCheXR-Attention (RANI et al., 2024) utiliza uma LSTM Bidirecional com atenção multi cabeça para classificar laudos radiológicos como normais ou anormais, destacando a importância da extração de características textuais para a tomada de decisão.

O trabalho de Xiong, Du e Yan (2019) representa um marco de transição entre abordagens convencionais e modelos baseados em *Transformers*, ao propor o *Hierarchical Reinforced Transformer*. A metodologia combinava a DenseNet-121 (HUANG et al., 2017) como codificador visual e um decodificador baseado em *Transformer*, treinado por reforço com o objetivo de gerar relatórios mais coerentes e informativos. Os experimentos foram conduzidos com a base *IU Chest X-ray* (DEMNER-FUSHMAN et al., 2016), avaliando os resultados com as métricas BLEU (PAPINENI et al., 2002) e CIDEr (VEDANTAM; ZITNICK; PARIKH, 2015).

Em Liu et al. (2021), foi apresentado um sistema composto por três módulos que busca simular a rotina de trabalho dos radiologistas. O modelo realiza a extração de regiões anormais, codifica informações textuais anteriores e destila o conhecimento adquirido para gerar relatórios mais precisos. As imagens foram codificadas com a ResNet-152 (HE et al., 2016), e o decodificador textual utilizava um modelo baseado em *Transformer*. Os experimentos utilizaram os conjuntos *IU Chest X-ray* (DEMNER-FUSHMAN et al., 2016) e MIMIC-CXR (JOHNSON et al., 2019), avaliados com BLEU, METEOR (BANERJEE; LAVIE, 2005), CIDEr e ROUGE (LIN, 2004).

Seguindo uma linha semelhante, You et al. (2021) introduziram uma arquitetura hierárquica que busca atenuar o viés de dados por meio do alinhamento entre regiões visuais de anormalidade e etiquetas de doenças. Utilizando ResNet-50 (HE et al., 2016) como codificador e um decodificador baseado em *Transformer*, o modelo foi avaliado nas bases MIMIC-CXR (JOHNSON et al., 2019) e IU Chest X-ray, apresentando ganhos notáveis na redução do viés e melhoria na cobertura semântica. As métricas utilizadas foram BLEU, METEOR e ROUGE.

O trabalho de Alfarghaly et al. (2021) apresentou um mecanismo de condicionamento semântico que visa melhorar a geração textual em termos de sobreposição de palavras e complexidade computacional. Os autores utilizaram DenseNet-121 (HUANG et al., 2017) para codificação visual e condicionaram um modelo GPT-2 (RADFORD et al., 2019) com atributos ponderados. O modelo foi avaliado com BLEU, METEOR, CIDEr e ROUGE, sendo um dos primeiros a empregar métricas de similaridade semântica para avaliação quantitativa da geração de laudos médicos.

Outro trabalho que se enquadra na categoria de abordagem modular com codificação visual seguida de geração textual é o de (CHEN et al., 2020). Nele, os autores

propõem um pipeline em duas etapas: a primeira realiza uma classificação multietiqueta baseada em CNN para extrair palavras-chave relevantes da imagem, enquanto a segunda utiliza um modelo de linguagem baseado em *Transformer* para gerar os laudos a partir dessas palavras-chave. A estratégia de adaptação dinâmica das palavras-chave e o uso de frequência para reforçar achados raros são diferenciais que buscam aumentar a interpretabilidade do processo, ainda que em detrimento de uma modelagem totalmente ponta a ponta.

As arquiteturas neurais dominantes nas primeiras abordagens, como Redes Neurais Convolucionais e Redes Neurais Recorrentes, são inerentemente locais e dependem de redes profundas para capturar informações distantes (WÖDLINGER, 2025). Essa limitação estrutural impacta tanto a extração de relações globais em imagens quanto a manutenção de coerência ao longo de sequências textuais extensas. Além disso, muitas dessas abordagens utilizam pipelines fragmentados ou acoplamento rígido entre os módulos visual e textual, dificultando o alinhamento semântico e comprometendo a fluência narrativa dos laudos gerados. Diante desse cenário, esta tese propõe uma arquitetura multimodal integrada, com um extrator visual hierárquico e um gerador textual mais expressivo e sensível ao contexto sequencial.

2.2 Arquiteturas baseadas em *Transformers*

Modelos baseados em *Transformers* revolucionaram a geração automática de laudos médicos ao superarem limitações presentes em arquiteturas sequenciais, como as RNN, especialmente no que se refere à captura de dependências de longo alcance e à paralelização do treinamento. O mecanismo de autoatenção proposto por (VASWANI et al., 2017) permite não apenas processar informações contextuais de maneira mais ampla, mas também integrar modalidades distintas, como imagem e texto, por meio da atenção cruzada. Embora as Redes Convolucionais (CNNs) continuem relevantes em tarefas puramente visuais, oferecendo vantagens em termos de eficiência computacional, precisão e simplicidade arquitetural, elas não foram originalmente projetadas para lidar com interações explícitas entre regiões espaciais distantes ou entre diferentes modalidades. Nesse contexto, modelos multimodais recentes passaram a empregar codificadores visuais e decodificadores textuais baseados inteiramente em *Transformers*, como ocorre na arquitetura proposta nesta tese, que integra esses componentes de forma unificada para capturar relações semânticas mais amplas entre imagens radiológicas e descrições clínicas.

Abordagens iniciais para geração de pré-laudos médicos com *Transformers* mantinham codificadores visuais baseados em RNC e introduziam *Transformers* apenas como decodificadores. Trabalhos como os de (XIONG; DU; YAN, 2019; LIU et al., 2021; YOU et al., 2021; ALFARGHALY et al., 2021) seguiram essa linha, combinando *DenseNet*

ou *ResNet* com decodificadores baseados em *Transformer*, alguns com mecanismos de atenção hierárquica ou condicionamento semântico, conforme discutido na seção anterior.

Com o surgimento do *Vision Transformer* (ViT) (DOSOVITSKIY et al., 2020a), novas arquiteturas passaram a empregar *Transformers* também no codificador de imagem, explorando sua capacidade de modelar longas dependências espaciais. Chen et al. (CHEN et al., 2020) propuseram um modelo baseado em *Transformer* impulsionado por memória, no qual uma memória relacional registra informações-chave do processo de geração e é incorporada ao decodificador por meio de uma camada de normalização condicional. Consolidando essa tendência, Agarwal e Verma (2024) apresentaram um sistema *dual-transformer* que utiliza um ViT como codificador de imagem e um decodificador GPT-3, demonstrando a eficácia da combinação de transformers tanto no domínio visual quanto no textual. Da mesma forma, Cheddi, Habbani e Nait-Charif (2024b) empregaram um ViT para extrair características visuais, que são subsequentemente fundidas com representações textuais do Word2Vec através de redes de atenção de canal e fusão *cross-modal*.

A superioridade dos ViTs em certas tarefas de imagem médica foi corroborada por estudos comparativos como o de Sharma et al. (2025) e Kandasamy (2023), que indicam que ViTs tendem a alcançar maiores pontuações de AUC, especialmente com grandes conjuntos de dados, enquanto as CNNs ainda podem dominar em cenários com dados limitados. Mohsan et al. (MOHSAN et al., 2022) investigaram o impacto da perda de viés espacial em codificadores visuais baseados em ViT pré-treinado, combinando-o com diferentes *Transformers* textuais no decodificador. O modelo foi avaliado no conjunto *IU Chest X-ray* com as métricas BLEU (PAPINENI et al., 2002), METEOR (BANERJEE; LAVIE, 2005), ROUGE (LIN, 2004) e CIDEr (VEDANTAM; ZITNICK; PARIKH, 2015), mostrando que a combinação adequada entre codificadores e decodificadores *Transformers* pode produzir relatórios mais descritivos e coesos.

Lee et al. Lee et al. (2022) em seu trabalho propuseram dois métodos para gerar melhores pré-laudos médicos a partir de imagens. O *Global-Local Visual Extractor* (GLVE) que captura recursos globais e locais da imagem e o *Cross Encoder-Decoder Transformer* (CEDT) para injetar características de vários níveis do codificador no processo de decodificação. O GLVE extrai não apenas características visuais globais que podem ser obtidas de uma imagem inteira, como tamanho de órgão ou estrutura óssea, mas também características visuais locais que podem ser geradas a partir de uma região local, como área de lesão. Dada uma imagem, o CEDT pode criar uma descrição detalhada dos recursos gerais injetando saídas do codificador de baixo e alto nível no decodificador. O modelo proposto foi avaliado no conjunto de dados *IU Chest X-ray* (DEMNER-FUSHMAN et al., 2016) com base nas métricas BLEU (PAPINENI et al., 2002), METEOR (BANERJEE; LAVIE, 2005) e ROUGE (LIN, 2004).

Trabalhos mais recentes também buscaram melhorar o alinhamento visual-textual com estratégias específicas. You et al. (2021) propuseram o *AlignTransformer*, que primeiro prevê categorias clínicas e depois alinha essas categorias com regiões visuais por meio de atenção hierárquica. Tanida et al. (2023) apresentaram um modelo baseado na geração de sentenças por região anatômica, usando uma *Faster R-CNN* para segmentação e um GPT-2 para redação textual, permitindo maior interpretabilidade e controle clínico. De forma similar, aprimorando o alinhamento e a qualidade do laudo, Cheddi, Habbani e Nait-Charif (2024a) propuseram o RG-MASR, um método que, após extrair características visuais e textuais, emprega um módulo de alinhamento multimodal e uma técnica de auto-refinamento para otimizar a coerência entre a imagem e o texto gerado.

Wang et al. (2025) introduziram a arquitetura LVMed-R2, baseada em modelo multimodal de larga escala com raciocínio iterativo e verificação automática. Após gerar o pré-laudo médico, o modelo aplica uma etapa de autoavaliação baseada em conhecimento médico, buscando reduzir inconsistências factuais. Essa preocupação com a precisão clínica também é central em outras abordagens recentes. Hou et al. (2025), por exemplo, propõem o framework RADAR, que otimiza a geração de laudos ao alavancar tanto o conhecimento interno de um LLM quanto informações externas recuperadas, evitando redundância e enriquecendo o laudo. Outra abordagem inovadora é a de Hou et al. (2024), que introduz um método de geração controlável (ECRG) baseado em modelos de energia, permitindo integrar conhecimento médico ou modelos especialistas externos durante a inferência, sem a necessidade de re-treinamento, para guiar a geração de texto. A consistência entre múltiplas fontes de imagem também foi explorada, como no C2M-DoT (WANG et al., 2026), que utiliza uma rede de transferência de domínio para garantir que o laudo gerado seja coerente mesmo quando apenas uma única vista da imagem está disponível.

Além das arquiteturas adaptadas diretamente de *Transformers* unimodais, nos últimos anos tem-se destacado o avanço de modelos multimodais pré-treinados em larga escala, capazes de integrar visão e linguagem de forma mais geral e flexível. Modelos como o BEiT-3 (WANG et al., 2023) estendem o aprendizado baseado em *tokens* visuais e textuais com objetivos de pré-treinamento conjuntos, enquanto o GIT (WANG et al., 2022) combina um codificador visual e um decodificador textual unificado por um cabeamento de atenção leve. Já o Flamingo (ALAYRAC et al., 2022), desenvolvido pela *DeepMind*, introduz uma arquitetura com atenção cruzada intercalada para gerar descrições e raciocínio visual com poucas imagens de entrada, destacando-se em tarefas *zero-shot*. Embora tais modelos tenham alcançado resultados expressivos em *benchmarks* genéricos de *Visual Question and Answering* (VQA) e *Image Captioning*, sua aplicação direta em domínios especializados como o médico ainda é limitada por questões de custo computacional, carência de dados anotados e necessidade de

explicabilidade.

Nesse sentido, trabalhos como o de Gowda et al. (2025) buscam superar essas limitações ao propor modelos de pré-treinamento de visão-linguagem específicos para o domínio médico, que são conscientes da incerteza e utilizam relatórios estruturados para enriquecer o contexto clínico. Adicionalmente, arquiteturas generativas alternativas, como o modelo baseado em GAN de Kumar, Karthik e Kamath (2024), exploram o treinamento adversário para forçar o gerador a produzir laudos mais indistinguíveis dos reais, melhorando a fidelidade clínica. Em um campo sinérgico, Fytas et al. (2024) introduzem o RadPrompt, uma metodologia que demonstra o potencial de sistemas baseados em regras para aprimorar as capacidades de LLMs na classificação de laudos radiológicos, mostrando que a combinação de abordagens pode superar o desempenho de cada modelo isoladamente.

A presente tese propõe uma arquitetura multimodal compacta, treinada de ponta a ponta, que combina o *Swin Transformer* como codificador visual e um decodificador textual baseado na arquitetura GPT, ambos integrados a um módulo de memória relacional. Essa configuração foi projetada para manter a coerência contextual e a fluidez textual na geração de laudos, ao mesmo tempo em que preserva a viabilidade computacional em ambientes com restrições de recursos. A estrutura proposta permite o alinhamento eficaz entre representações visuais e linguísticas, superando limitações observadas em modelos com acoplamento rígido ou sem mecanismos explícitos de memória.

2.3 Considerações Finais

Este capítulo apresentou uma revisão aprofundada da literatura sobre modelos neurais aplicados à geração automática de pré-laudos médicos textuais a partir de radiografias de tórax. As soluções investigadas demonstram a evolução do campo desde as primeiras arquiteturas CNN-RNN, que utilizavam codificadores convolucionais e decodificadores recorrentes, até os sistemas baseados em *Transformers*, que incorporam mecanismos de autoatenção e estratégias multimodais com alinhamento visual-textual. A Tabela 1 resume os principais trabalhos analisados, destacando suas contribuições, bases de dados e métricas de avaliação.

Embora codificadores convolucionais ainda sejam adotados em algumas soluções recentes (WANG et al., 2023), os *Transformers* se destacam por sua flexibilidade arquitetural, capacidade de modelar relações espaciais amplas e compatibilidade natural com mecanismos de atenção multimodal. Essas características favorecem seu uso em tarefas que demandam alinhamento visual-textual e geração de conteúdo coerente. Assim, a escolha por um codificador baseado em *Swin Transformer*, conforme adotado nesta tese, reflete não apenas o avanço técnico do estado da arte, mas também uma decisão

fundamentada nos requisitos específicos da tarefa de geração de laudos médicos.

Tabela 1 – Resumo dos trabalhos relacionados.

Método	**Contribuições**	**Bases de Dados**	**Métricas**
RTMIC(XIONG; DU; YAN, 2019)	• Metodologia *Hierarchical Reinforced Transformer*; • Uso inicial de *Transformers* e aprendizagem por reforço.	IU Chest X-ray	BLEU CIDEr
PPKED(LIU et al., 2021)	• Sistema modular que simula o fluxo de trabalho do radiologista; • Extração de regiões anormais e destilação de conhecimento.	IU Chest X-ray MIMIC-CXR	BLEU METEOR CIDEr ROUGE
Align Transformer(YOU et al., 2021)	• Mitigação de viés de dados com alinhamento hierárquico; • Alinhamento iterativo entre regiões visuais e etiquetas de doenças.	IU Chest X-ray MIMIC-CXR	BLEU METEOR ROUGE
CDGPT2(ALFARGHALY et al., 2021)	• Mecanismo de condicionamento semântico com GPT-2; • Uso de métricas de similaridade semântica para avaliação.	*IU Chest X-ray*	BLEU METEOR CIDEr ROUGE
MCLN(CHEN et al., 2020)	• *Transformer* impulsionado por memória relacional; • Normalização condicional para incorporar informações da memória.	IU Chest X-ray MIMIC-CXR	BLEU METEOR ROUGE F1
CEDT(LEE et al., 2022)	• Extrator visual Global-Local (GLVE) e Transformer Cross Encoder-Decoder (CEDT); • Injeção de *features* de múltiplos níveis do encoder no decoder.	*IU Chest X-ray*	BLEU METEOR ROUGE
REG-GPT2(TANIDA et al., 2023)	• Geração de sentenças por região anatômica (Faster R-CNN); • Decodificação com GPT-2 e interpretabilidade por visual grounding.	IU Chest X-ray	BLEU METEOR ROUGE
Multi-Model DL(MUHARRAM et al., 2023)	• Pipeline com múltiplos modelos de classificação binária para cada anormalidade; • Geração baseada em sentenças pré-determinadas a partir de um "código de resultado".	*NIH Chest X-ray*	ACC
Enhancing CXR Analysis(KANDASAMY, 2023)	• Estudo comparativo de CNNs e ViTs com XAI (GradCAM, SHAP); • ViTs alcançam maior AUC com menor tempo de treinamento.	NIH-CXR-LT MIMIC-CXR-LT	AUC
RG-MASR(CHEDDI; HABBANI; NAIT-CHARIF, 2024a)	• Alinhamento de características multimodais com estratégia de auto-refinamento; • Melhora a coerência e qualidade do laudo gerado.	IU X-ray *NIH Chest X-ray*	BLEU ROUGE
Multi-modal fusion(CHEDDI; HABBANI; NAIT-CHARIF, 2024b)	• Encoder ViT para visual e Word2Vec para texto; • Fusão com redes de atenção de canal e informação cross-modal.	IU X-ray *NIH Chest X-ray*	BLEU ROUGE
Dual-Transformer(AGARWAL; VERMA, 2024)	• Sistema com encoder Vision Transformer (ViT) e decoder GPT-3; • Arquitetura totalmente baseada em transformers para imagem e texto.	IU X-ray *NIH Chest X-ray*	BLEU ROUGE CIDEr METEOR
GAN-MLC(KUMAR; KARTHIK; KAMATH, 2024)	• Classificador multi-rótulo baseado em Redes Generativas Adversariais (GANs); • Treinamento adversário para gerar laudos mais realistas e precisos clinicamente.	NIH *IU Chest X-ray*	AUC BLEU
ECRG(HOU et al., 2024)	• Geração de laudos controlável usando Modelos Baseados em Energia (EBMs); • Integra conhecimento médico na inferência sem re-treinamento.	MIMIC-CXR Chest ImaGenome	BLEU METEOR ROUGE F1
CCheXR-Attention(RANI et al., 2024)	• Extração de conceitos clínicos e classificação de laudos (normal/anormal); • Usa Bi-LSTM com atenção multi cabeça para extração de *features* textuais.	MIMIC-CXR	ACC Precision Recall F1
RadPrompt(FYTAS et al., 2024)	• Metodologia de prompting que usa sistema baseado em regras para aprimorar LLMs; • Demonstra sinergia entre modelos de regras e LLMs para classificação de laudos.	MIMIC-CXR	F1
Explainable DL(SWATHI; BAI, 2025)	• Foco em classificação multi-rótulo de 14 doenças torácicas; • Usa CNNs pré-treinadas (ResNet, DenseNet) para identificar e localizar patologias.	*NIH Chest X-ray*	ACC
RADAR(HOU et al., 2025)	• Injeção de conhecimento suplementar para aprimorar LLMs; • Distingue e agrega conhecimento interno do modelo e conhecimento externo.	MIMIC-CXR IU X-ray	BLEU RG-F1 e RGER RadCliQ$_0$ CheXpert
Comparative Analysis(SHARMA et al., 2025)	• Análise comparativa entre ViTs e CNNs para classificação de imagens médicas; • ViTs se destacam em grandes datasets, CNNs em datasets menores.	N/A	ACC
Distribution-based VLM(GOWDA et al., 2025)	• Modelo de visão-linguagem pré-treinado consciente da incerteza; • Utiliza relatórios estruturados para enriquecer o contexto clínico.	MIMIC-CXR RSNA Pneumonia SIIM-ACR Pneumothorax CheXpert NIH COVIDx CXR	AUC F1 ACC
C2M-DoT(WANG et al., 2026)	• Geração consistente para imagens de múltiplas vistas (multi-view); • Rede de transferência de domínio para inferência com vista única.	IU Chest X-ray MIMIC-CXR	BLEU METEOR ROUGE

3 Fundamentação Teórica

Este capítulo descreve as principais técnicas usadas na concepção do modelo proposto e as estruturas arquiteturais que subsidiaram sua implementação. O objetivo do modelo é gerar automaticamente pré-laudos médicos a partir de exames de imagem de RX de tórax.

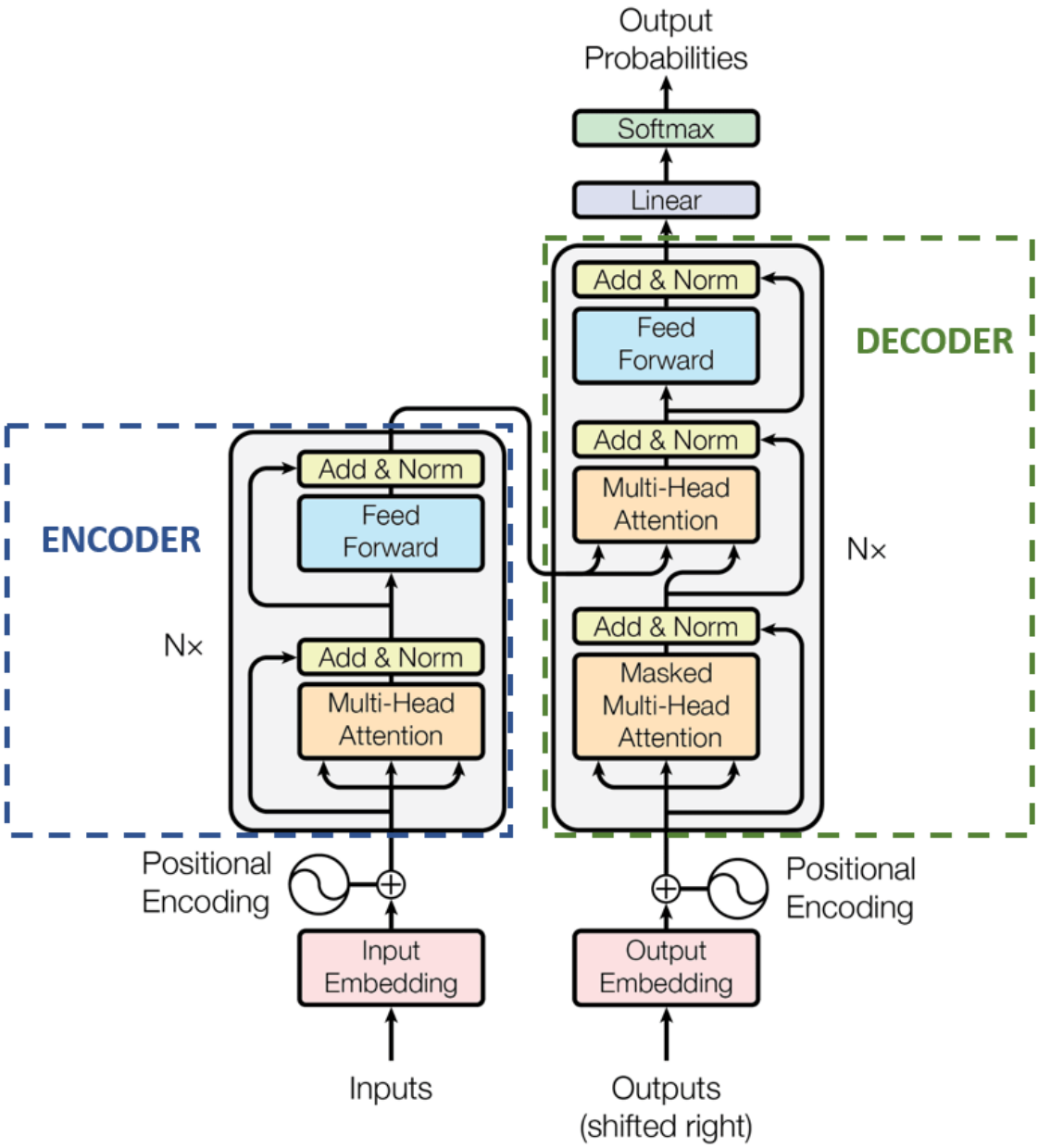
3.1 *Transformers*

A arquitetura *Transformer* original (ver Figura 1) foi proposta por Vaswani et al. (2017) para a realização da tarefa de tradução automática e é composta de um codificador e um decodificador. Este modelo, por exemplo, pode ser treinado para traduzir uma frase do inglês para o francês. Os autores introduziram um novo bloco de construção baseado em mecanismos de atenção, que são essencialmente camadas de rede neural que relacionam diferentes posições em uma única sequência para calcular sua representação. O *Transformer* é livre de convoluções, baseado exclusivamente no mecanismo de autoatenção, o que permite relacionar e preservar todas as características presentes nos dados de entrada do modelo.

Codificador: O codificador (à esquerda na Figura 1) é responsável por codificar a sequência de entrada (por exemplo, uma frase) em um vetor de comprimento fixo ou uma sequência de vetores de comprimento fixo, que pode ser usado como entrada para o decodificador. O codificador usa mecanismos de autoatenção para processar a sequência de entrada de maneira paralela, permitindo capturar dependências de longo alcance entre as palavras na sequência de entrada. O mecanismo de autoatenção permite que o modelo pondere a importância de cada palavra na sequência ao criar a codificação, o que é útil para tarefas em que as relações entre as palavras são importantes (como na tradução). A saída do codificador pode ser pensada como uma representação compacta da sequência de entrada, que resume seu significado e contexto. O decodificador da rede usa essa representação para gerar a sequência de saída de destino (VASWANI et al., 2017).

Decodificador: O decodificador (à direita na Figura 1) é um tipo de rede neural inicialmente usada para PLN. O decodificador é responsável por gerar a sequência de saída em tarefas como tradução automática e geração de texto. Ele recebe como entrada as representações codificadas do codificador e usa mecanismos de autoatenção para pesar e incorporar dinamicamente diferentes partes da entrada à medida que gera a sequência de saída passo a passo. O decodificador também possui um conjunto adicional de camadas chamado de camada de saída, que é usado para produzir as previsões finais (VASWANI et al., 2017).

Figura 1 – Arquitetura *Transformer* codificador-decodificador.



Fonte: Vaswani et al. (2017)

Autoatenção: Em uma camada autoatenção (Figura 2a), o vetor de entrada é primeiramente transformado em três vetores separados, ou seja, o vetor de *query* q, o vetor de *key* k e o vetor de *value* v com uma dimensão fixa, onde:

- *Query* (q): representa a posição atual na sequência de entrada para a qual queremos calcular uma pontuação de atenção em relação a todas as outras posições na sequência. Obtemos a *query* aplicando uma transformação linear aprendida nas *embeddings* de entrada.
- *Key* (k): é um conjunto de transformações lineares aprendidas a partir das

embeddings de entrada que são usadas para calcular as pontuações de atenção entre a *query* e todas as outras posições na sequência de entrada. Obtemos as chaves aplicando outra transformação linear aprendida nas *embeddings* de entrada.

- *Value* (v): é outro conjunto de transformações lineares aprendidas a partir das *embeddings* de entrada que são usadas para calcular uma soma ponderada dos valores em todas as posições na sequência de entrada, onde os pesos são determinados pelas pontuações de atenção entre a *query* e todas as outras posições na sequência. Obtemos os valores aplicando outra transformação linear aprendida nas *embeddings* de entrada.

Esses vetores são então agrupados em três matrizes de pesos diferentes, nomeadas W^Q, W^K e W^V. Uma forma comum de Q, K e V pode ser formulada como a Equação 3.1 para uma determinada entrada X:

$$K = W^K X, Q = W^Q X, V = W^V X, \tag{3.1}$$

onde W^Q, W^K e W^V se referem aos parâmetros treináveis. Já o mecanismo atenção escalonada do produto escalar é formulado seguindo a Equação 3.2 a seguir:

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}}V), \tag{3.2}$$

onde $\sqrt{d_k}$ é um fator de escala e uma operação *softmax* é aplicada aos pesos de atenção gerados para traduzi-los em uma distribuição normalizada.

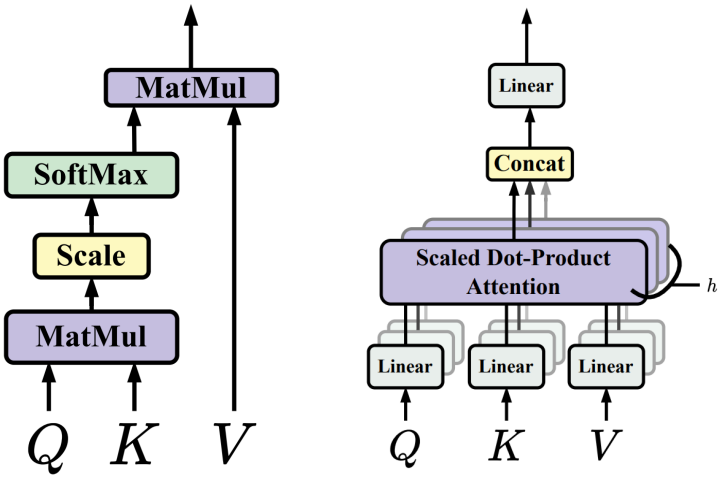
Autoatenção Multi Cabeça: O mecanismo autoatenção multi cabeça (Figura 2b) foi proposto em Vaswani et al. (2017) para modelar as relações complexas de entidades simbólicas de diferentes aspectos. Especificamente, o bloco MSA ajuda o modelo a atender conjuntamente as informações de vários subespaços de representação, pois a capacidade de modelagem do bloco de atenção de cabeça única é bastante grosseira. O processo de MSA pode ser formulado seguindo a Equação 3.3 a seguir:

$$MultiHead(Q, K, V) = [Concat(head_1, ..., head_h)]W^O, \tag{3.3}$$

onde $head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$, e W^O indica uma função de mapeamento linear para combinar a representação de várias cabeças. Vale ressaltar que h é um hiperparâmetro, referente a camadas paralelas de atenção ou cabeças, definido como $h = 8$ no trabalho original.

***Cross-Attention*:** É um mecanismo usado em tarefas que envolvem a comparação das representações de duas sequências de entrada diferentes para calcular pontuações de atenção para cada elemento em uma sequência em relação a todos os elementos

Figura 2 – (a) *Self-Attention*. (b) *Multi-head attention*. O MSA consiste em vários blocos SA (*heads*) concatenados juntos por canal como proposto em Vaswani et al. (2017)



(a) *Self-Attention* (b) *Multi-Head Self-Attention*

Fonte: Vaswani et al. (2017)

na outra sequência. Essas pontuações de atenção são usadas para ponderar as contribuições de cada elemento em uma sequência ao calcular as representações dos elementos na outra sequência.

Por exemplo, em uma tarefa de geração textual a partir de uma imagem, a imagem pode ser visualizada como a sequência de consulta e o texto como uma sequência chave para descrição da imagem. O mecanismo de atenção calculará um peso para cada conjunto de pixels na imagem de origem em relação a todas as palavras no texto de destino, permitindo que o modelo se concentre dinamicamente em diferentes partes da imagem de origem ao gerar cada palavra no texto de destino.

Os modelos *Transformers* obtiveram grande sucesso em diversas tarefas de PLN, destacando-se o BERT (*Bidirectional Encoder Representations from Transformers*) (DEVLIN et al., 2019), GPT-2 (*Generative Pre-trained Transformer*) (RADFORD et al., 2019) e T5 (*Text-to-Text Transfer Transformer*) (RAFFEL et al., 2020), que se tornaram bastante populares.

O BERT pré-treinou o *Transformer* de maneira auto-supervisionada usando o *Masked Language Model* (MLM) e *Next Sentence Prediction* (NSP) (DEVLIN et al., 2019). No MLM, 15% das palavras em uma frase são mascaradas aleatoriamente; a seguir, o modelo é treinado para prever essas palavras usando a *cross-entropy loss*[1]. Esse processo capacita o modelo a considerar o contexto bidirecional durante as previsões e a

[1] Também conhecida como *negative log-likelihood* é uma medida amplamente utilizada como medida de desempenho no treino de modelos de AM (GHEFLATI; RIVAZ, 2022; CASTRO et al., 2022; HABIB; SALEEM; LALL, 2023).

estabelecer associações entre diversas palavras. Essa capacidade é fundamental em várias tarefas de Processamento de Linguagem Natural, como a resposta a perguntas e a inferência de linguagem natural.

Uma abordagem de treinamento diferente foi usada no GPT-2, conhecida como *Causal Language Modeling* (CLM) (RADFORD et al., 2019), na qual o modelo é treinado para prever a próxima palavra dadas todas as palavras anteriores. O GPT-2 foi desenvolvido exclusivamente com camadas de decodificador do *Transformer*, onde as incorporações de tokens e as incorporações posicionais foram calculadas e adicionadas às sequências de tokens como entrada. *Multi-head Self-Attention* (MSA), rede *feedforward*, normalização de camadas e conexões residuais são aplicadas por cada camada. Esta arquitetura, especificamente, foi utilizada na criação do modelo multimodal proposto neste trabalho, devido ao seu desempenho em tarefas de geração textual. A eficácia de modelos baseados em GPT para a geração de laudos também foi explorada em trabalhos mais recentes, como o de Agarwal e Verma (2024), que combina um decodificador GPT-3 com um codificador visual, demonstrando a robustez dessa família de modelos para a síntese de texto clínico.

O DistilGPT-2 (SANH et al., 2019) é uma versão compactada popular do GPT-2 à qual possui menos parâmetros. O DistilGPT-2 é treinado no *OpenWebTextCorpus* (GOKASLAN et al., 2019) usando métodos de *Knowledge destillation*.

3.1.1 *Vision Transformer*

A arquitetura *Vision Transformer* (ViT) adapta o modelo *Transformer*, originalmente projetado para Processamento de Linguagem Natural, ao domínio da Visão Computacional. Diferentemente das abordagens baseadas em convoluções, o ViT representa a imagem como uma sequência unidimensional de patches fixos, que são linearizados e processados por camadas de atenção própria. Essa estrutura permite a modelagem explícita de dependências de longo alcance entre regiões da imagem, sem o uso de filtros convolucionais. Estudos recentes indicam que, com volume adequado de dados e capacidade computacional, o ViT pode alcançar desempenho comparável ou superior ao de arquiteturas convolucionais tradicionais em tarefas de classificação, detecção e segmentação de imagens.

Baseado na arquitetura dos *Transformers*, Dosovitskiy et al. (2020b) propuseram o modelo *Vision Transformer* conforme ilustrado na Figura 3. Quando treinado em grandes conjuntos de dados, por exemplo, JFT-300M(SUN et al., 2017), o ViT supera o estado da arte atingido anteriormente por modelos baseados em RNC como BiT (KOLESNIKOV et al., 2020). A crescente adoção e o sucesso do ViT no domínio médico são evidenciados por múltiplos estudos comparativos recentes (KANDASAMY, 2023; SHARMA et al., 2025), que analisam seu desempenho em relação às CNNs. Esses trabalhos concluem que,

embora as CNNs ainda possam ser vantajosas em cenários com dados limitados, os ViTs demonstram um desempenho superior em grandes datasets, alcançando maior precisão na classificação e, em alguns casos, com menor tempo de treinamento, o que justifica sua escolha como base para extratores visuais em modelos de geração de laudos (CHEDDI; HABBANI; NAIT-CHARIF, 2024b; AGARWAL; VERMA, 2024).

O *Transformer* padrão recebe como entrada uma sequência 1D de *embeddings* de *tokens*(WEBSTER; KIT, 1992), contudo, para lidar com imagens 2D, foi realizada a transformação da imagem em *patches* de tamanho fixo antes de ser achatada em vetores. Esses vetores são passados por uma camada de projeção linear treinável que os mapeia em N vetores com a dimensionalidade de $D \times N$, onde N é o número de patches. As saídas deste estágio são chamadas de *patch embeddings*. Para preservar as informações posicionais presentes em cada *patch*, eles adicionam *embeddings* posicionais às *patch embeddings*. Além disso, uma classe de *embedding* treinável também é anexada às *patch embeddings* antes de passar pelo codificador do *Transformer*.

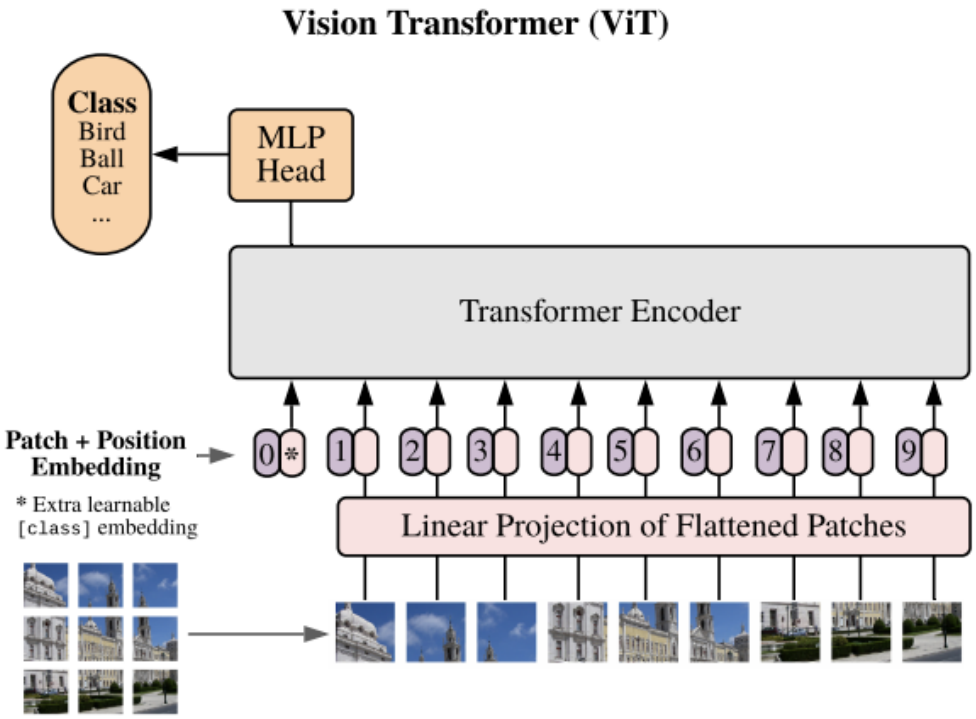
O codificador do *Transformer* é composto por vários blocos iguais, os quais são compostos por um bloco de autoatenção multi cabeça e um bloco *Multilayer Perceptron* (MLP)(ROSENBLATT, 1958) em cada bloco. As ativações são primeiro normalizadas no bloco do codificador do *Transformer* usando *LayerNorm* (LN)(BA; KIROS; HINTON, 2016) antes de ir para esses blocos. Além disso, existem conexões antes do LN que adicionam uma cópia dessas ativações às saídas correspondentes do bloco MSA ou MLP. No final, há um bloco MLP usado como cabeça de classificação que mapeia a saída para as previsões da classe. O mecanismo de autoatenção é uma característica chave dos modelos *Transformer* (AZAD et al., 2023).

3.1.2 Módulo de Memória Relacional

Nesta tese, foi incorporado ao decodificador textual um módulo de memória relacional com o objetivo de reforçar a capacidade de retenção de informações ao longo da sequência de geração, capturando dependências de longo prazo e assegurando consistência contextual. Inspirado na arquitetura proposta por Santoro et al. (2018), esse componente atua como uma memória explícita, composta por múltiplos vetores (ou *slots*) atualizados de forma iterativa durante o processo de decodificação.

A memória relacional é representada por uma matriz $M_t \in \mathbb{R}^{N \times d}$, onde N denota o número de *slots* de memória e d a dimensão de cada vetor. Em cada passo de geração t, o conteúdo da memória é atualizado com base nos vetores de consulta (Q), chave (K) e valor (V) extraídos das ativações do decodificador. O mecanismo de atualização é realizado por meio de uma atenção multi cabeça (*Multi-Head Attention*), seguida de uma projeção linear e uma camada MLP com normalização, responsável por processar as interações entre os *slots*.

Figura 3 – Arquitetura ViT e a estrutura do seu bloco codificador. *Patches* de imagens sequenciais são usados como entrada e processados por um codificador para produzir a saída de classificação final.



Fonte: Dosovitskiy et al. (2020b)

Um aspecto fundamental da atualização da memória é o uso de portões de controle, um **portão de entrada** (g_e) e um **portão de esquecimento** (g_f). Esses mecanismos determinam dinamicamente o quanto de informação nova deve ser incorporada à memória e o quanto do conteúdo anterior deve ser retido. Ambos os portões são vetores aprendidos durante o treinamento e são aplicados elemento a elemento (*element-wise*) para controlar a atualização de cada slot de memória $m_i \in M_t$, conforme:

$$m_i^{(t)} = g_f \odot m_i^{(t-1)} + g_e \odot \tilde{m}_i^{(t)}, \tag{3.4}$$

onde $\tilde{m}_i^{(t)}$ representa a nova informação gerada após a atenção e o processamento via MLP, $g_f \in [0,1]^d$ é o portão de esquecimento que regula a preservação do estado anterior, e $g_e \in [0,1]^d$ é o portão de entrada que controla a incorporação de novos conteúdos. A operação $\odot$ denota a multiplicação elemento a elemento.

O conteúdo atualizado da memória é então propagado para os blocos subsequentes do decodificador via um mecanismo de normalização condicional orientada à memória, chamado Memory-Driven Conditional Layer Normalization (MCLN). Esse

mecanismo modula dinamicamente os parâmetros de normalização (γ e β) de cada camada, com base nas representações armazenadas em M_t, por meio de duas redes MLP independentes que geram os deltas de ajuste $\Delta\gamma$ e $\Delta\beta$:

$$\gamma^* = \gamma + \Delta\gamma(M_t), \quad \beta^* = \beta + \Delta\beta(M_t) \tag{3.5}$$

com γ^* e β^* sendo os parâmetros finais utilizados na normalização da camada atual.

A integração da memória relacional ao fluxo do decodificador permite que o modelo retenha e atualize representações sem a necessidade de arquiteturas recorrentes. Isso garante uma geração textual mais coerente, contextualizada e fluida, essencial para aplicações clínicas em que a precisão semântica e a continuidade narrativa são determinantes. Enquanto esta tese foca na memória relacional para aprimorar a consistência interna do laudo gerado, outras abordagens contemporâneas buscam a consistência por meio de fontes externas, como a injeção de conhecimento suplementar de bancos de dados (HOU et al., 2025) ou a validação cruzada entre múltiplas vistas de uma mesma imagem (WANG et al., 2026), o que demonstra a relevância do desafio da coerência no estado da arte. Além disso, o uso de poucos *slots* (geralmente três) e dimensionalidade controlada permite uma incorporação eficiente sem impacto significativo no número total de parâmetros do modelo.

3.2 Métricas de Avaliação

Para avaliar a qualidade dos pré-laudos gerados pelos modelos utilizados nesse trabalho, foram calculados os escores do Bilingual Evaluation Understudy (BLEU) (PAPINENI et al., 2002), ROUGE-L (LIN, 2004), METEOR (LAVIE; AGARWAL, 2007) e SPICE (ANDERSON et al., 2016). Essas métricas são amplamente abordadas em trabalhos da área (MONSHI; POON; CHUNG, 2020; ÇALLI et al., 2021; PANG; LI; ZHAO, 2023; CHEDDI; HABBANI; NAIT-CHARIF, 2024a; AGARWAL; VERMA, 2024; WANG et al., 2026), possibilitando a comparação e o rastreamento da evolução dos trabalhos ao longo dos anos.

O BLEU é uma métrica de avaliação comumente utilizada para tarefas de PLN, particularmente tradução automática. No entanto, será utilizado neste trabalho para avaliar o problema de geração de texto. Ele mede a qualidade do texto previsto pelo modelo em comparação com o texto de referência, comparando as sobreposições de n-gramas entre os dois. Quanto maior a sobreposição entre o texto previsto e o texto de referência, maior o escore BLEU. O BLEU-1 avalia palavra por palavra individualmente, o BLEU-2 avalia combinações de duas palavras, o BLEU-3 avalia agrupamentos de três palavras e o BLEU-4 avalia combinações de quatro palavras; assim, a avaliação é mais rigorosa.

Esta métrica é calculada conforme a Equação 3.6 abaixo:

$$BLEU = BP \cdot exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \tag{3.6}$$

onde N é o tamanho de n-gramas, w são pesos positivos, p_n é a precisão modificada do n-grama e BP é a penalidade de brevidade que, segundo Papineni et al. (2002), é utilizada para penalizar com um decaimento exponencial as predições geradas que são muito menores do que o tamanho da referência e é calculada seguindo a Equação 3.7 a seguir:

$$BP = \begin{cases} 1, & \text{se } c > r \\ e^{(1\text{-}r/c)}, & \text{se } c \leq r \end{cases}, \tag{3.7}$$

onde c é o comprimento total do corpus de tradução candidato e r o comprimento de referência efetivo do corpus de teste.

Para acompanhar e avaliar a evolução de desempenho do treino do modelo multimodal foi utilizada uma função de perda denominada *cross entropy loss*. Também chamada de função de perda de entropia cruzada, ela mede a diferença entre a distribuição de probabilidade das predições de um modelo e a distribuição de probabilidade real dos dados. O objetivo durante o treino de um modelo de aprendizado de máquina é minimizar os valores dessa métrica durante a execução do treino e na validação, o que indica que o modelo está generalizando bem e conseguindo aprender com os dados. A *cross entropy loss* é calculada seguindo a Equação 3.8 abaixo:

$$L(y, \hat{y}) = -\sum y_i \log\left(\hat{y}_i\right), \tag{3.8}$$

onde y é a distribuição de probabilidade real dos dados, $\hat{y}$ é a distribuição de probabilidade das predições, y_i é o i-ésimo elemento de y e $\hat{y}_i$ é o i-ésimo elemento de $\hat{y}$.

3.3 Considerações Finais

Neste capítulo foram apresentados os métodos utilizados para desenvolver este trabalho, incluindo informações sobre as arquiteturas utilizadas na elaboração do modelo e as métricas de avaliação dos resultados. A fundamentação nos *Transformers*, tanto para o processamento de texto com arquiteturas baseadas em GPT quanto para a análise de imagens com o *Vision Transformer*, alinha esta tese com as tendências dominantes no estado da arte, conforme evidenciado por inúmeros trabalhos recentes (KANDASAMY, 2023; AGARWAL; VERMA, 2024; SHARMA et al., 2025). A escolha de um módulo de memória relacional representa uma abordagem específica para resolver o desafio da coerência textual de longo prazo, um problema também abordado por técnicas ortogonais como a injeção de conhecimento (HOU et al., 2025) e o controle explícito da geração (HOU et al., 2024). Assim, as bases teóricas aqui descritas não apenas sustentam a metodologia proposta, mas também a posicionam dentro de um diálogo ativo com as pesquisas mais atuais da área.

O capítulo seguinte descreve o modelo multimodal proposto, bem como o processo de codificação das imagens, o processo de geração textual e como eles estão interligados.

4 Metodologia

Este capítulo descreve a metodologia adotada para a geração automática de pré-laudos médicos a partir de imagens de radiografias de tórax, com ênfase em duas arquiteturas distintas desenvolvidas ao longo da pesquisa: *XRaySwinGen* e Modelo Multimodal de Memória Relacional (MMMR). As seções seguintes detalham individualmente cada arquitetura, com a apresentação de seus diagramas esquemáticos no início de cada subseção para facilitar o entendimento da evolução estrutural e das decisões técnicas.

4.1 *XRaySwinGen*

XRaySwinGen é uma arquitetura *Transformer* (VASWANI et al., 2017) modificada que realiza a análise de imagens de RX e gera pré-laudos médicos automaticamente. Em outras palavras, é uma tarefa de incorporação de sequência a sequência de ponta a ponta, na qual os pixels da imagem são sequências de entrada e o pré-laudo médico que descreve a imagem é a saída desejada. Esta abordagem foi fortemente influenciada pelo estado da arte (CHEN et al., 2020; MIURA et al., 2021; NGUYEN et al., 2021) e pelo artigo de Liu et al. (LIU et al., 2021), que teve avanços notáveis na adaptação da arquitetura direcionada a texto para a área de visão computacional. Liu et al. exploraram as discrepâncias entre ambos os domínios, destacando as amplas variações de escala das entidades visuais e a alta resolução dos pixels em imagens, em contraste com a estrutura textual baseada em palavras. A arquitetura *XRaySwinGen* é apresentada na Figura 4.

Nesta tese, a extração de características de imagens de RX é realizada com o mapeamento hierárquico de características e o Deslocamento de Janela (LIU et al., 2021) para capturar efetivamente dependências locais e globais, uma evolução da arquitetura base ViT(DOSOVITSKIY et al., 2020b). A espinha dorsal *Swin Transformer* (LIU et al., 2021) foi treinada com a base de dados imagenet-1k (DENG et al., 2009); além disso, o tamanho original dos partições de 4x4 e a configuração Swin-L foram mantidos. A geração de pré-laudos médicos é realizada por uma versão treinada baseada na arquitetura GPT com 124 milhões de parâmetros e tamanho de vocabulário 50.257, disponibilizada aos pesquisadores pela OpenAI[1] (RADFORD et al., 2019). Quando as instâncias dos modelos de codificador e decodificador pré-treinados são inicializadas, suas respectivos incorporações são conectados por meio de camadas de atenção cruzada. O *XRaySwinGen* é composto por 12 camadas de atenção cruzada.

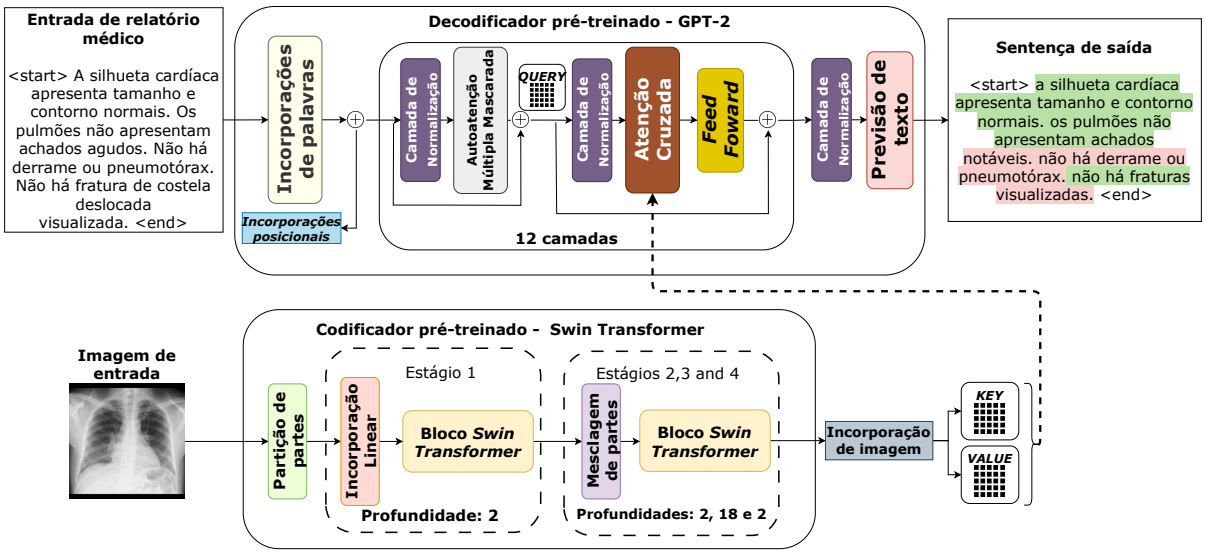
O codificador inicialmente divide uma imagem RGB de entrada em partições não sobrepostas, onde cada partição é tratada como um "*token*" e sua característica é definida a partir da concatenação dos valores brutos dos *pixels*. A dimensão das características de cada partição é de $4 \times 4 \times 3 = 48$, onde 3 se refere à quantidade de canais de imagem. Para aprimorar a percepção do modelo, a operação de deslocamento de janela é realizada. Nela, há um deslocamento na sequência de partições por um certo número de posições, permitindo que o modelo capture características globais e locais da imagem, sendo capaz de processar imagens em diferentes níveis de granularidade, o que não é possível com RNC tradicionais. As incorporações do codificador são utilizados como *KEY* e *VALUE*, enquanto no lado do decodificador, as incorporações são utilizados como *QUERY* na cabeça da camada de atenção cruzada.

Durante o treinamento, tanto a imagem de RX quanto o pré-laudo médico são as entradas do modelo. Utilizando esses dados, o modelo é direcionado a gerar a mesma legenda, levando a uma compreensão da correlação entre as palavras nos pré-laudos médicos e as imagens de RX. Ao longo de

[1] https://openai.com/blog/better-language-models/

cada iteração do treinamento, o modelo gera uma sequência de palavras como uma previsão. Esta sequência prevista é então comparada à legenda real, e a diferença resultante, conhecida como perda, é retroalimentada para o modelo. Este processo permite que o modelo aprenda e melhore sua capacidade de gerar pré-laudos mais precisos ao longo do tempo.

Figura 4 – Arquitetura do modelo *XRaySwinGen* para geração automática de pré-laudos médicos.



Fonte: Adapitado de (LIU et al., 2021) e (RADFORD et al., 2019)

A seguir são detalhadas, em duas seções, o funcionamento do codificador de imagem e o decodificador textual, tal como a integração de ambos, que resulta no modelo multimodal proposto. Na Seção 4.2, é detalhado o processo de codificação e extração de características a partir das imagens de RX, de forma que cada imagem é representada ao final por um vetor de características, o qual será compartilhado com o decodificador textual durante o treino.

Em seguida, na Seção 4.3, é mostrado o pré-processamento textual realizado, a conversão do texto médico em *tokens* tal como a codificação dos *tokens* em incorporações e, por fim, a técnica de aprendizagem por mascaramento utilizada, associada à atenção cruzada, adicionada ao modelo para processar duas informações sequenciais: o vetor de características da imagem com o vetor textual, possibilitando assim relacionar e sugerir termos específicos durante a geração textual com base nas características de imagem extraídas.

4.2 Codificador - *Swin Transformer*

Nesta tese foi usado como codificador de imagem o modelo *Swin Transformer* (LIU et al., 2021), pré-treinado com a base de dados *imagenet*-1k (DENG et al., 2009). Este modelo introduziu dois conceitos chave com relação à arquitetura ViT original: o mapeamento hierárquico de características e a abordagem de deslocamento de janelas (*Shifted Window Approach*), a qual deu origem ao nome *Swin Transformer*.

Com base no diagrama da Figura 4 parte inferior a descrição passo a passo de como o bloco *Swin Transformer* funciona é descrita a seguir:

- **Imagem de entrada**: O *Swin Transformer* inicialmente divide uma imagem RGB de entrada em partes não sobrepostas por meio do módulo Partição de Partes (PP), como o ViT (Figura 3). Cada

parte é tratada como um "*token*" e sua característica é definida a partir da concatenação dos valores brutos dos *pixels* RGB. Um tamanho de partições de 4 × 4 é usado e, portanto, a dimensão das características de cada parte é 4 × 4 × 3 = 48, onde 3 refere-se a quantidade de canais da imagem. Uma camada de Incorporação Linear é aplicada a esse valor bruto de características para projetá-lo em uma dimensão arbitrária *C*. Com o intuito de potencializar a percepção do modelo sem aumentar o número de parâmetros, o custo computacional e ainda torná-lo mais eficiente em termos de uso de memória, o método PP é usado em combinação com a operação Deslocamento de Janela (Figura 5b), ou deslocamento de janela. Nela, ocorre um deslocamento na sequência de partições por um determinado número de posições permitindo que o modelo capture características globais e locais da imagem, sendo capaz de processar imagens em diferentes níveis de granularidade, o que não é possível com RNCs tradicionais.

- **Estágio 1**: Após a execução da Incorporação Linear, vários blocos *Transformer* com mecanismos modificados de autoatenção (Bloco Swin Transformer) são aplicados nesses *tokens* de *patch*. Neste estágio, os blocos *Transformer* mantêm o número de *tokens* em (H/4 × W/4), onde H é a altura e W corresponde a largura, ambos de tamanho 224. Para produzir uma representação hierárquica, o número de *tokens* é reduzido por meio da camada Mesclagem de partes à medida que a rede se aprofunda. O primeiro bloco de Mesclagem de partes e transformação de características é indicado como Estágio 2. Esse procedimento é repetido duas vezes, nos Estágio 3 e Estágio 4.

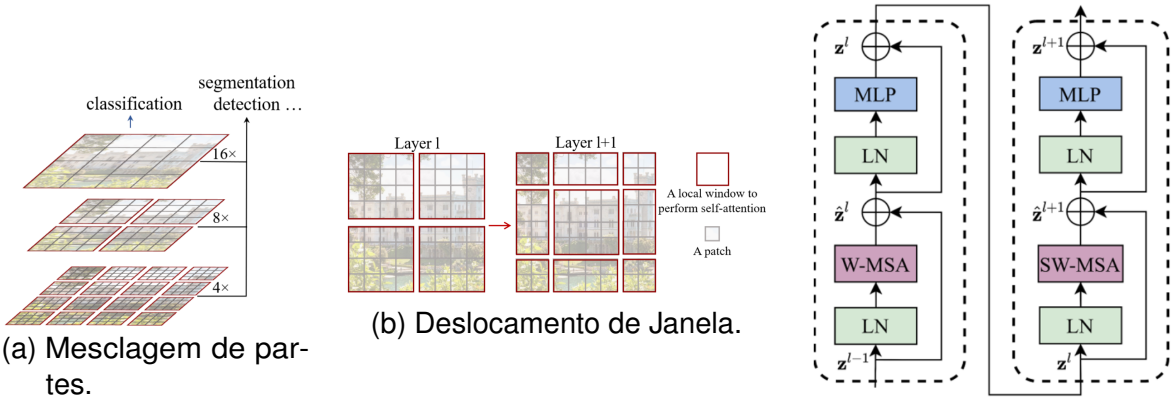
- **Estágio 2, 3 e 4**: A primeira execução da camada Mesclagem de partes, concatena as características de cada grupo de partições vizinhas 2 × 2 e aplica uma camada linear de dimensão *4C* nas características concatenadas. Isso reduz o número de *tokens* em um múltiplo de 2 × 2 = 4 (diminuição da resolução em *2×*) e a dimensão de saída é definida como *2C*, aumentando assim a profundidade. Neste aspecto, *C* refere-se ao número de canais do modelo *Swin Transformer*, ou seja, a dimensão da *embedding* do modelo o qual foi pré definido com *C = 192*. Como pode ser visto na Figura 5a, a camada Mesclagem de partes concatena quatro partições. Assim, a cada concatenação a altura e a largura da imagem são reduzidas ainda mais por um fator de *2*. No Estágio 1, a resolução de entrada é (H/4 × W/4), mas após o Mesclagem de partes, a resolução mudará para (H/8 × W/8) que será a entrada para o Estágio 2. Para o Estágio 3, a resolução de entrada será (H/16 × W/16) e para o Estágio 4, a resolução de entrada será (H/32 × W/32).

- **Bloco Swin Transformer**: O *Swin Transformer* é composto de duas subunidades, as quais consistem em um módulo MSA baseado em *Shift Window* (Figura 5b), seguida por uma MLP de 2 camadas com ativação de não linearidade *Gaussian Error Linear Unit* (GELU) (HENDRYCKS; GIMPEL, 2016), uma função de ativação de rede neural de alto desempenho que pondera as entradas por seus valores. Uma camada de normalização LN é aplicada antes de cada módulo MSA e cada MLP, e uma conexão residual é aplicada após cada módulo. A técnica Layer Norm normaliza as distribuições de camadas intermediárias. A primeira subunidade usa um módulo *Window* MSA (W-MSA), enquanto a segunda subunidade usa um módulo *Shifted Window* MSA (SW-MSA). Em cada Estágio da arquitetura *Swin Transformer*, existem dois Blocos Swin Transformer consecutivos, exceto no Estágio 3, onde existem 18 blocos em conjunto.

  W-MSA: Na W-MSA, a sequência de entrada é dividida em janelas de tamanho fixo e o mecanismo de autoatenção é aplicado separadamente a cada janela. Isso pode reduzir o custo computacional da autoatenção, uma vez que as pontuações de atenção só precisam ser calculadas dentro de cada janela, e não em toda a sequência. No entanto, esta abordagem pode levar a descontinuidades entre janelas adjacentes, o que pode ser problemático para algumas tarefas.

  SW-MSA: Para resolver o problema de descontinuidades entre janelas adjacentes, a SW-MSA foi proposta. Nessa variação de MSA, cada janela se sobrepõe às janelas adjacentes por um número

Figura 5 – Módulos da arquitetura do *Swin Transformer*.



(a) Mesclagem de partes.

(b) Deslocamento de Janela.

(c) Dois Blocos *Swin Transformer* sucessivos.

Fonte: Liu et al. (2021)

fixo de posições, chamado de tamanho de deslocamento. Isso permite que as informações fluam de maneira mais suave entre janelas adjacentes o que demonstrou melhorar o desempenho em diversas tarefas de imagem.

- **Características da Imagem**: Por fim, a saída do *Swin Transformer* após todos os Estágios será uma representação vetorial de alta dimensão da imagem de entrada, que será usada para compartilhar informações da imagem com o modelo de geração textual por meio da camada de Atenção Cruzada vista na Seção 3.1.

4.3 Decodificador - GPT-2 *Transformer*

O GPT-2 é um modelo de linguagem baseado na arquitetura *Transformer*, o qual foi originalmente treinado usando técnicas de aprendizado não supervisionado com um *corpus* (denominado *WebText* por Radford et al. (2019)) de aproximadamente 40 GB de textos coletados da *Internet*, a grande maioria em inglês. O modelo foi treinado com um objetivo simples: prever a próxima palavra, considerando todas as palavras anteriores em algum texto, sendo assim um modelo constituído apenas de blocos decodificador (Figura 1). A diversidade do conjunto de dados faz com que esse objetivo simples contenha demonstrações naturais de muitas tarefas em diversos domínios. O GPT-2 é uma expansão direta do GPT, com mais de 10 vezes o número de parâmetros e treinado com mais de 10 vezes a quantidade de dados. Nesta tese foi utilizada uma versão pré-treinada menor do GPT-2 com 124 milhões de parâmetros disponibilizada para pesquisadores pela OpenAI[2] (RADFORD et al., 2019).

Com base na Figura 4, foi feita a descrição passo a passo de como o bloco GPT-2 *Transformer* funciona.

- **Entrada de relatório médico**: A entrada textual inicialmente passa pelo processo de tokenização, que consiste na divisão do texto em palavras ou *tokens* individuais. Nesta etapa um *token* é adicionado antes dos demais *tokens* da entrada textual para indicar o início do texto ao modelo. O tokenizador do GPT-2 é baseado em *Byte-Pair Encodings* (BPE) (GAGE, 1994), que é basicamente

[2] https://openai.com/blog/better-language-models/

um vocabulário composto por caracteres e sequências de caracteres mais comuns presentes no *corpus*. Palavras pequenas ou muito comuns podem fazer parte deste vocabulário, enquanto outras são decompostas em partes menores e tratadas como mais de um *token*. Isso permite que o tokenizador possa ser utilizado para diversas línguas.

- **Incorporações de Palavras e Incorporações Posicionais**: Após o processo inicial de tokenização, o modelo utiliza as incorporações de palavras para codificar o significado de cada palavra na sequência de entrada, enquanto que as incorporações posicionais são usadas para codificar a posição de cada palavra dentro da sequência. As incorporações de palavras possuem dimensão 768 e as incorporações posicionais dimensão 1024.
- **Atenção Multi Cabeça**: A Atenção Multi Cabeça (AMC) utiliza um mecanismo de autoatenção similar ao da Figura 2a, porém, adiciona uma etapa de mascaramento a esse mecanismo, o que impede que o modelo olhe para futuras posições na sequência durante o treinamento. Isso é feito mascarando os pesos de atenção para as posições que ainda não foram processadas na sequência de entrada. Portanto, o modelo só pode atender a sequência de entrada até a posição atual, o que auxilia o modelo a gerar cada *token* mais preciso e consistente na sequência de saída.
- **Atenção Cruzada**: Para combinar as informações de ambas as modalidades (imagem e texto), o mecanismo de atenção cruzada é usado para calcular os pesos de atenção entre os *tokens* de texto e as partições da imagem. Os pesos de atenção determinam o grau em que cada *token* atende a cada *patch* da imagem e vice-versa. Esse mecanismo funciona primeiro calculando uma matriz de consulta Q, uma matriz de chave K e uma matriz de valor V para cada modalidade. Essas matrizes são usadas para calcular os pesos de atenção, que são calculados como o produto escalar entre as matrizes Q e K, seguido por uma operação *softmax* para obter as pontuações de atenção. As pontuações de atenção são usadas para ponderar a matriz de valor, produzindo a saída final do mecanismo de atenção cruzada.
- ***Feed Forward***: Uma vez que calculamos todos os pesos e obtivemos as pontuações a partir do atenção cruzada, acabamos com um vetor que representa cada *token* e contém o seu contexto apropriado. Esses vetores são então apresentados à rede neural *feed-forward* que é usada para calcular as probabilidades da próxima palavra ou *token* em uma determinada sequência, com base nas palavras anteriores e no contexto. Isso é feito por meio de uma função *softmax* que gera uma distribuição de probabilidade sobre todo o vocabulário de palavras, a partir do qual a próxima palavra mais provável pode ser selecionada.

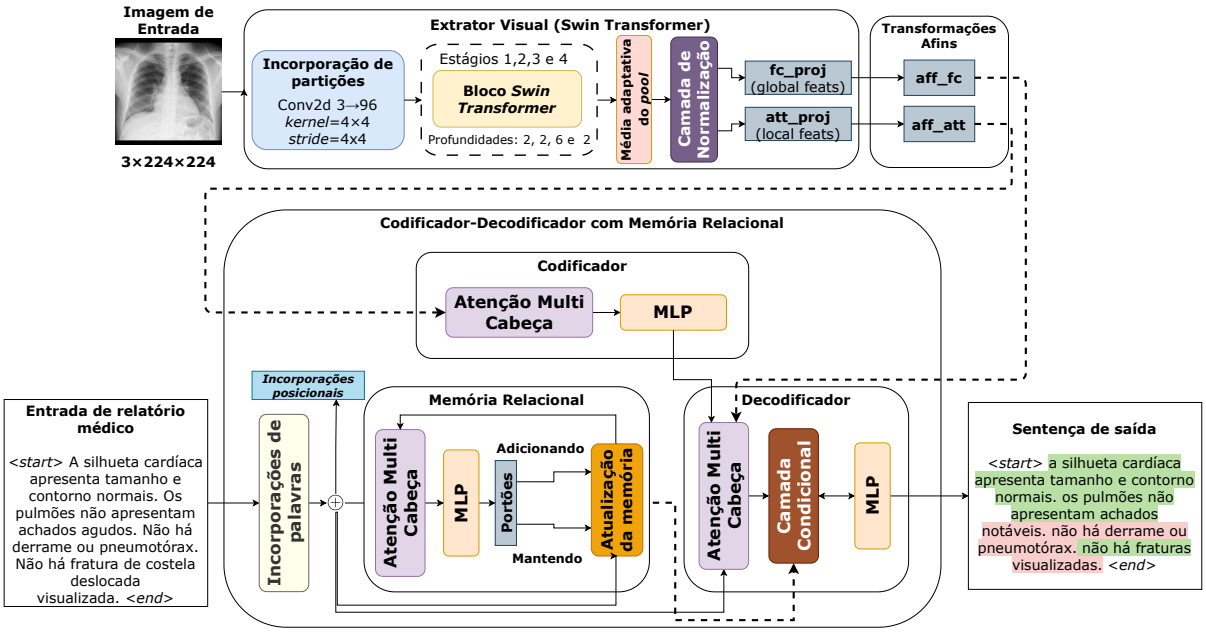
4.4 Modelo Multimodal de Memória Relacional: MMMR

A arquitetura MMMR foi concebida como uma evolução do *XRaySwinGen*, com o objetivo de mitigar limitações associadas à retenção de dependências de longo prazo e à fluência narrativa em tarefas de geração automática de pré-laudos médicos. O modelo combina um extrator visual baseado no *Swin Transformer* com um decodificador *Transformer* enriquecido por um módulo de memória relacional e camadas de normalização condicional. A arquitetura proposta é mostrada na Figura 6.

Extrator Visual com *Swin Transformer*

O extrator visual do MMMR utiliza uma implementação do *SwinTiny* da biblioteca *Hugging Face Transformers*, composta por quatro estágios hierárquicos. Cada estágio é formado por múltiplos blocos *SwinLayer*, com atenção local por janelas e deslocamento, mantendo as operações de Mesclagem de

Figura 6 – Arquitetura do modelo MMMR proposto.



Fonte: Elaborado pelo autor.

partes com reduções sucessivas de resolução: de 96 para 192, 384 e, por fim, 768 canais. O estágio final incorpora camadas lineares adicionais para projeção vetorial em duas direções distintas, otimizadas para a projeção semântica (`fc_proj`) e atenção cruzada (`att_proj`), ambas com dimensão 1024.

Decodificador com Memória Relacional

O decodificador do MMMR é uma arquitetura *Transformer* clássica com três camadas decodificadoras, mas estendida com dois componentes centrais:

- **Memória Relacional**: módulo com 3 *slots* de memória, cada um com dimensão 512, implementado com mecanismos de atenção multi cabeça (8 cabeças com projeções lineares) seguido por uma MLP não linear com ativação ReLU. A memória recebe entradas de consulta, chave e valor derivadas do embedding textual inicial, atualizando o estado interno a cada passo de geração.
- **Camada Condicional**: cada subcamada do decodificador (autoatenção, atenção cruzada e *feedforward*) é conectada a uma instância de Conexão condicional da subcamada, que aplica normalização condicional orientada à memória. A modulação dos parâmetros diferenciais de escala e deslocamento, denotados por $\Delta\gamma$ e $\Delta\beta$, é realizada por dois MLPs independentes. Cada MLP é composto por duas camadas lineares de 512 unidades com ativação ReLU, que recebem como entrada a concatenação do vetor do decodificador com o vetor da memória expandida, totalizando 1536 dimensões.

Integração Multimodal e Afins

O vetor de características extraído da imagem é transformado por duas camadas lineares distintas: affine_aa e affine_bb, ambas com dimensão de saída 2048, que são fundidas e alinhadas com os embeddings textuais para o processo de atenção cruzada. O módulo *logit* final projeta os vetores decodificados para um vocabulário de 7864 tokens, a partir dos quais são geradas as sequências textuais.

Resumo das Melhorias Arquiteturais

Comparado ao *XRaySwinGen*, o MMMR introduz ganhos substanciais em retenção de contexto e regularização da geração textual. A adição de memória relacional provê uma capacidade explícita de reter e inter-relacionar informações ao longo do tempo. Já as camadas de normalização condicional aumentam a adaptabilidade dos blocos do decodificador ao contexto multimodal. Essa busca por maior precisão e consistência clínica através de mecanismos de aprimoramento específicos está em consonância com o estado da arte, que explora desde a injeção de conhecimento externo (HOU et al., 2025) e a geração controlável por modelos de energia (HOU et al., 2024), até estratégias de auto-refinamento (CHEDDI; HABBANI; NAIT-CHARIF, 2024a) para melhorar a qualidade dos laudos. O modelo manteve-se eficiente, com 237 milhões de parâmetros, e apresentou melhoria nas métricas ROUGE-L e METEOR nas bases IU-Xray e MIMIC-CXR, como detalhado no Capítulo 5.

4.5 Considerações Finais

Este capítulo apresentou as duas arquiteturas propostas nesta pesquisa: *XRaySwinGen* e MMMR. A primeira combina o Swin Transformer com o GPT-2 e atenção cruzada para gerar textos médicos coerentes a partir de imagens. Já a MMMR introduz um módulo de memória relacional e normalizações condicionais, permitindo capturar dependências de longo prazo e enriquecer semanticamente o processo de geração textual com um número reduzido de parâmetros.

A evolução metodológica entre as arquiteturas demonstra um esforço contínuo de aprimoramento na representação visual e textual. No capítulo seguinte, serão descritos os conjuntos de dados utilizados, os protocolos de treinamento adotados, incluindo validação cruzada estratificada 3-*fold*, os hiperparâmetros definidos e as métricas utilizadas para avaliar a qualidade dos pré-laudos gerados.

5 Experimentos e Resultados

Neste capítulo são descritos os experimentos realizados para validar as arquiteturas propostas: *XRaySwinGen* e Modelo Multimodal de Memória Relacional. O objetivo é demonstrar o potencial dos modelos na geração automática de pré-laudos médicos a partir de imagens de RX. Primeiramente, é feita uma introdução sobre as bases de dados utilizadas, seguida dos detalhes do pré-processamento realizado em cada conjunto de dados; em seguida são mostrados as configurações e parâmetros utilizados no treinamento das arquiteturas e por fim são relatados os resultados obtidos seguidos de discussões.

5.1 Bases de dados

Nesta seção são descritas as bases utilizadas para treinamento e avaliação dos modelos propostos. O modelo *XRaySwinGen* foi treinado com quatro bases distintas: Base Privada, *IU Chest X-ray*, *NIH Chest X-ray* e *MIMIC-CXR-JPG*. Já o modelo Modelo Multimodal de Memória Relacional foi avaliado exclusivamente com a base *MIMIC-CXR-JPG*. Cada base apresenta características específicas quanto ao idioma, tipo de anotação e estrutura textual, impactando diretamente na forma de pré-processamento e treinamento dos modelos.

- **Base privada**: composta por 21.970 imagens no formato DICOM obtidas de hospitais privados brasileiros, com resoluções variando entre 727×692 e 4892×4020 pixels. Os pré-laudos estão em português, redigidos de forma natural por radiologistas locais. Foi necessário aplicar rotinas de anonimização com expressões regulares para remoção de dados sensíveis (ex: nomes, CPFs, código de inscrição único no CRM, idades). Após limpeza, os textos foram convertidos para letras minúsculas e tokenizados, resultando em um vocabulário de 5.438 frases distintas.
- ***IU Chest X-ray***: proposta por Demner-Fushman et al. (2016), essa base reúne imagens e laudos provenientes do Hospital da Universidade de Indiana. Contém 3.794 imagens emparelhadas com pré-laudos em inglês, escritos com maior detalhamento clínico e média de comprimento superior à Base Proposta. O pré-processamento textual consistiu apenas na normalização para letras minúsculas. O vocabulário final contém 2.827 frases distintas.
- ***NIH Chest X-ray***: proposta por Wang et al. (2017), essa base possui 112.120 imagens de tórax frontal, correspondentes a 30.805 pacientes. Os laudos originais foram substituídos por expressões textuais curtas que representam as 14 classes clínicas disponíveis (por exemplo: "No Finding", "Atelectasis", "Cardiomegaly"etc.), conforme mapeamento ilustrado na Tabela 2. Os textos apresentam comprimento limitado (1 a 2 palavras), o que limita o uso de métricas n-grama mais altas durante a avaliação dos modelos.

Tabela 2 – Mapeamento de classes *NIH Chest X-ray*.

Classe numérica	**Texto**
0	No Finding
1	Atelectasis
2	Cardiomegaly
3	Effusion
4	Infiltration
5	Mass
6	Nodule
7	Pneumonia
8	Pneumothorax
9	Consolidation
10	Edema
11	Emphysema
12	Fibrosis
13	Pleural_Thickening
14	Hernia

- ***MIMIC-CXR-JPG***: proposta por Johnson et al. (2019), contém 377.110 imagens em formato JPEG derivadas do conjunto MIMIC-CXR, originalmente armazenado em DICOM. As imagens foram convertidas utilizando o pacote OpenCV com equalização de histograma e compressão em qualidade 95, e estão acompanhadas de rótulos estruturados extraídos automaticamente dos relatórios clínicos por ferramentas como CheXpert e NegBio. As anotações cobrem 14 categorias, incluindo Atelectasis, Edema, Pneumonia e No Finding. Todos os dados foram devidamente de-identificados em conformidade com os requisitos da regulamentação HIPAA Safe Harbor. Duas versões dessa base foram usadas: (i) versão *FULL*, contendo referências ao histórico clínico; (ii) versão *No History References*, filtrada para incluir somente descrições da imagem atual. A versão *FULL* foi dividida em 354.619 imagens para treino, 2.866 para validação e 4.710 para teste, enquanto a versão filtrada possui 191.686 imagens para treino, 10.649 para validação e 10.650 para teste. Ambas versões foram submetidas à validação cruzada com três partições independentes.

Na Tabela 3 são mostradas as informações das bases de dados usadas nesta tese, como: número de amostras, assim como a proporção utilizada para treinamento, validação e teste.

Tabela 3 – Resumo das bases de dados utilizadas.

Base de Dados	**Treino**	**Validação**	**Teste**	**Total**	**Referência**
Base Privada	17.537	924	3.509	21.970	-
IU Chest X-ray	3.000	452	342	3.794	(DEMNER-FUSHMAN et al., 2016)
NIH Chest X-ray	82.197	25.596	4.327	112.120	(WANG et al., 2017)
MIMIC-CXR-JPG (*FULL*)	354.619	2.866	4.710	362.195	(JOHNSON et al., 2019)
MIMIC-CXR-JPG (*NO EXAM HISTORY*)	191.686	10.649	10.650	212.985	(JOHNSON et al., 2019)

5.2 Configurações de Treinamento

Esta seção descreve os ambientes, estratégias e hiperparâmetros utilizados para o treinamento dos modelos propostos: *XRaySwinGen* e Modelo Multimodal de Memória Relacional. As configurações são organizadas conforme a base de dados empregada, o modelo treinado e os recursos computacionais disponíveis.

5.2.1 *XRaySwinGen*: Configurações Arquiteturais e de Treinamento

O modelo *XRaySwinGen* é composto por um codificador com 24 blocos baseados em *Swin Transformer* e um decodificador com 12 blocos do modelo GPT-2. Os pesos iniciais utilizados provêm dos modelos pré-treinados *Swin Transformer Base* e GPT-2, e são ajustados via *deep fine-tuning*. Essa técnica atualiza todos os parâmetros durante o treinamento, sendo indicada quando há diferença substancial entre os domínios dos dados pré-treinados (como o ImageNet) e os dados-alvo (radiografias médicas).

Os experimentos com o *XRaySwinGen* foram conduzidos de forma individual para cada uma das seguintes bases: Base Privada, *IU Chest X-ray* e *NIH Chest X-ray*. Em todos esses casos, os modelos foram treinados sem validação cruzada, utilizando a infraestrutura do Google Colab, equipada com uma GPU NVIDIA Tesla T4 (16GB), processador Intel Xeon @2.20 GHz e 13GB de RAM.

Com a base MIMIC-CXR-JPG, o modelo *XRaySwinGen* foi treinado em ambas as versões (com e sem histórico clínico), empregando validação cruzada 3-fold. Esse treinamento foi realizado em ambiente mais robusto, com GPU NVIDIA GeForce RTX 4090 24GB, visando acelerar o processamento e suportar conjuntos maiores de dados.

A configuração de hiperparâmetros utilizada para o modelo *XRaySwinGen* foi:

- Tamanho do *batch* de treino: 6;
- Tamanho do *batch* de validação: 8;
- Número de épocas: 25;
- Taxa de aprendizagem: 2^{-5};
- Otimizador: AdamW;
- Estratégia de avaliação: por época;
- Precisão numérica: FP16 habilitada.

A escolha dos tamanhos de *batch* foi determinada empiricamente, maximizando a utilização da memória da GPU sem provocar estouros. O número de épocas foi estabelecido com base em estudos prévios que apontam boa convergência de *Transformers* em menos de 30 épocas. A taxa de aprendizagem e o otimizador AdamW seguem recomendações amplamente adotadas na literatura especializada (VASWANI et al., 2017; LOSHCHILOV; HUTTER, 2017; FAROOQ; HAFEEZ, 2020; LEE et al., 2022; CHEDDI; HABBANI; NAIT-CHARIF, 2024a; AGARWAL; VERMA, 2024).

5.2.2 MMMR: Configuração Arquitetural e Treinamento

O Modelo Multimodal de Memória Relacional foi treinado exclusivamente com a base pública *MIMIC-CXR-JPG* (JOHNSON et al., 2019), utilizando validação cruzada *3-fold* nas versões *FULL* e filtrada da base. A escolha por essa base se fundamenta em sua ampla adoção na literatura especializada,

especialmente em estudos voltados à geração automática de laudos radiológicos com abordagens multimodais (LI; WANG; LUO, 2020; YUAN et al., 2019; DELBROUCK; ZHANG; RUBIN, 2021; JING; XIE; XING, 2018; HOU et al., 2025; WANG et al., 2026; HOU et al., 2024; FYTAS et al., 2024). Além de seu grande volume de dados rotulados, a *MIMIC-CXR-JPG* oferece um cenário clínico realista, com laudos textuais autênticos e imagens de alta complexidade, o que a torna adequada para treinar e avaliar modelos com demandas de alinhamento semântico entre imagem e texto.

- Um extrator visual com 4 estágios hierárquicos *Swin Transformer*, iniciando com convoluções 4×4 e projeções com dimensões de *embedding* progressivas (96 até 768);
- Um decodificador do tipo *Transformer* com 3 camadas, integrando atenção multi cabeça, atenção cruzada e *feed-forward*;
- Camadas de normalização do tipo *Conditional LayerNorm*, parametrizadas por MLPs independentes para cálculo de $\Delta\gamma$ e $\Delta\beta$ com ativação ReLU;
- Um módulo externo de Memória Relacional (*RelationalMemory*), com mecanismos de atenção multi cabeça e atualizações paralelas ao decodificador;
- Vocabulário com 7.864 tokens e *embedding* de saída com 512 dimensões.

O treinamento foi realizado em servidor Linux com 64 núcleos lógicos e GPU NVIDIA GeForce RTX 4090 (24GB). Os hiperparâmetros utilizados no treinamento do MMMR foram definidos com base em uma combinação de boas práticas observadas na literatura e ajustes empíricos. Estudos recentes com modelos baseados em *Transformers* e mecanismos de memória explícita sugerem que configurações com dimensões intermediárias, estratégias de normalização condicional e ajustes diferenciados de taxa de aprendizado entre módulos favorecem a estabilidade e a generalização em tarefas clínicas sensíveis (LI; REN, 2020; KRALJEVIC et al., 2021; DELBROUCK; ZHANG; RUBIN, 2021; GOWDA et al., 2025). A escolha de 100 épocas, em particular, foi realizada empiricamente tendo em vista a instabilidade observada no ambiente de execução. Optou-se por configurar o treinamento com um número de épocas elevado, garantindo que o modelo tivesse tempo suficiente para convergir, mesmo em cenários com interrupções inesperadas. Essa decisão visou preservar o progresso por meio do uso contínuo de pontos de verificação (*checkpoints*), assegurando reprodutibilidade e continuidade no treinamento. A lista abaixo apresenta a configuração completa adotada:

- Tamanho do *batch*: 128;
- Número de épocas: 100;
- Taxa de aprendizagem: 5×10^{-5};
- Estratégia de avaliação: validação cruzada 3-fold;
- Precisão numérica: FP16 habilitada.

5.3 Resultados e Discussões

Os resultados quantitativos dos modelos avaliados são apresentados na Tabela 4 e 5. As métricas utilizadas (BLEU-1 a BLEU-4, ROUGE-L, METEOR e SPICE) foram calculadas com o suporte das ferramentas Stanford CoreNLP (MANNING et al., 2014) e *MS-COCO Caption Evaluation for Python 3* (CHEN et al., 2015), assegurando comparabilidade com os trabalhos mais relevantes da área.

Tabela 4 – Resultados obtidos pelos modelos *XRaySwinGen* e MMMR em diferentes bases de dados.

MODELO	BASE	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	SPICE
XRaySwinGen	Base Privada	0,731	0,691	0,672	0,661	0,748	0,741	0,659
XRaySwinGen	IU Chest X-ray	0,377	0,239	0,168	0,124	0,300	0,322	0,202
XRaySwinGen	*NIH Chest X-ray*	0,396	0,329	-	-	0,404	0,393	0,170
XRaySwinGen	*MIMIC-CXR-JPG* (FULL)	0,206	0,120	0,078	0,054	0,304	0,233	0,156
XRaySwinGen	*MIMIC-CXR-JPG* (*NO EXAM HISTORY*)	0,309	0,192	0,131	0,093	0,363	0,341	0,229
MMMR	*MIMIC-CXR-JPG* (FULL)	0,378	0,233	0,157	0,114	0,321	0,281	0,164
MMMR	*MIMIC-CXR-JPG* (*NO EXAM HISTORY*)	0,477	0,328	0,242	0,187	0,416	0,384	0,240

Os resultados consolidados para todas as bases avaliadas são mostrados na Tabela 4. A Base Privada apresenta 40% de imagens anormais e 60% normais, todas em alta definição, acompanhadas por pré-laudos detalhados que variam entre 30 e 120 palavras. Ainda que a base contenha laudos com estruturas recorrentes e alta proporção de exames normais, o modelo demonstrou desempenho competitivo nas métricas consideradas. As pontuações de BLEU, embora influenciadas pela similaridade lexical com os textos de referência, indicam que o modelo reproduz com razoável fidelidade os padrões linguísticos presentes nos laudos originais. Os valores de ROUGE-L refletem boa cobertura de segmentos textuais esperados, mas devem ser interpretados com cautela devido à natureza repetitiva de muitos laudos normais. Já os escores obtidos em METEOR e SPICE sugerem que, mesmo em um conjunto reduzido e com menor diversidade semântica, o modelo foi capaz de gerar descrições com algum grau de coerência contextual. No entanto, a avaliação automática não substitui validações clínicas qualitativas, especialmente em casos que demandam maior especificidade descritiva.

O modelo treinado com o conjunto de dados *IU Chest X-ray* resulta em pontuações menores de BLEU[1-4] e ROUGE-L do que o modelo de Chen et al., o que pode ser agravado pela distribuição da amostra. Durante o processamento do conjunto de dados, foram filtradas imagens frontais e algumas delas não tinham o pré-laudo preenchido ou vice-versa, o que resultou nos números apresentados na Tabela 3. Ainda assim, os resultados se aproximam dos reportados por trabalhos como o de Agarwal e Verma (2024), que emprega uma arquitetura *dual-transformer*, e o RG-MASR de Cheddi, Habbani e Nait-Charif (2024a), que foca no alinhamento multimodal e auto-refinamento. O desempenho é igualmente notável ao se considerar a complexidade de modelos mais recentes como o C2M-DoT (WANG et al., 2026), que explora a consistência entre múltiplas vistas da imagem. A pontuação SPICE não foi apresentada em seu trabalho, contudo, nosso resultado de 0,202 indica que o acordo entre a geração do modelo e a referência humana não é bom em termos de conteúdo semântico. Uma modificação futura no treinamento deste modelo será adquirir o banco de dados atualizado e usar imagens de RX frontais e laterais.

O modelo treinado com *NIH Chest X-ray* apresenta uma pontuação BLEU[1-2] mais elevada do que o modelo treinado com imagens da *IU Chest X-ray*. O BLEU[3-4] não foi calculado porque os pré-laudos consistem em no máximo duas palavras neste conjunto. Embora baseada em classificação, o objetivo deste experimento era utilizar uma base de dados com maior volume para compreender como o modelo se comportaria com essa quantidade de dados. O modelo aprendeu com sucesso as classes predefinidas,

conforme evidenciado pelos resultados das métricas BLEU[1-2], ROUGE-L e METEOR. Porém, por se tratar de um modelo generativo em uma base com número limitado de classes, o modelo frequentemente alucina e produz sentenças um pouco mais longas que a classe esperada; portanto, a pontuação SPICE foi baixa.

A avaliação dos modelos na base *MIMIC-CXR-JPG* (*FULL*) revela a complexidade inerente à tarefa de geração automática de laudos médicos em cenários de alta variabilidade clínica e textual. Conforme mostrado na Tabela 4, o *XRaySwinGen* apresenta desempenho estável, com BLEU-1 de 0,206, BLEU-4 de 0,054, ROUGE-L de 0,304 e METEOR de 0,233. O modelo MMMR, por sua vez, apresenta ganhos substanciais em todas as métricas: BLEU-1 de 0,378, BLEU-4 de 0,114, ROUGE-L de 0,321 e METEOR de 0,281. Estes resultados, detalhados também na Tabela 5, evidenciam que a arquitetura de memória relacional não só supera as soluções tradicionais como também eleva o patamar de qualidade do pré-laudo médico, especialmente em aspectos de coerência, abrangência e fidelidade semântica. Os ganhos são particularmente notáveis em métricas sensíveis à estrutura e conteúdo de longo alcance, como METEOR e ROUGE-L.

No conjunto de dados *MIMIC-CXR-JPG* (*NO EXAM HISTORY*), o contexto é restrito unicamente à descrição da imagem, o que permite ao modelo focar exclusivamente no relacionamento das informações visuais com o texto. Nesse cenário, conforme a Tabela 4, o *XRaySwinGen* apresenta BLEU-1 de 0,309, BLEU-4 de 0,093, ROUGE-L de 0,363 e METEOR de 0,341. O modelo MMMR, por sua vez, obtém desempenho substancialmente superior, alcançando BLEU-1 de 0,477, BLEU-4 de 0,187, ROUGE-L de 0,416, METEOR de 0,384 e SPICE de 0,240. Esses ganhos refletem a alta eficácia do mecanismo de memória relacional multimodal para alinhar evidências visuais e linguagem médica, mesmo em contextos de máxima restrição textual. Observa-se que, sem o histórico clínico, o MMMR potencializa sua capacidade de generalização visual e de geração de laudos detalhados, precisos e semanticamente consistentes, estabelecendo um novo padrão para a geração automática de relatórios em radiologia.

A Tabela 5 traz uma comparação ampliada com trabalhos do estado da arte e as variações de temperatura no modelo MMMR, evidenciando o avanço significativo alcançado pela proposta desta tese.

Tabela 5 – Comparação dos resultados dos modelos MMMR (com variação de temperatura), *XRaySwinGen* e de outros autores em bases *MIMIC-CXR* e *MIMIC-CXR-JPG*.

Base	Modelo	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	SPICE
MIMIC-CXR	R2Gen (CHEN et al., 2020)	0,353	0,218	0,145	0,103	0,277	0,142	-
MIMIC-CXR	*AlignTransformer* (YOU et al., 2021)	0,378	0,235	0,156	0,112	0,283	0,158	-
MIMIC-CXR	CvT-212DistilGPT2 (NICOLSON; DOWLING; KOOPMAN, 2023)	**0,392**	**0,245**	**0,169**	**0,124**	0,285	0,153	-
MIMIC-CXR	RGRG (TANIDA et al., 2023)	0,373	0,249	0,175	0,126	0,264	0,168	-
MIMIC-CXR	LVMed-R2 (WANG et al., 2025)	0,334	-	-	0,120	0,273	0,195	-
MIMIC-CXR	Radiology-LLaMA2 (LIU et al., 2023)	-	-	-	-	0,4427	-	-
MIMIC-CXR	Flamingo-CXR (TANNO et al., 2023)	-	-	-	0,101	0,297	-	-
MIMIC-CXR	Med-PaLM M (SINGHAL et al., 2025)	-	-	-	0,104	0,262	-	-
MIMIC-CXR-JPG (*FULL*)	*XRaySwinGen* **(Ours)**	0,206	0,120	0,078	0,054	0,304	0,233	0,156
MIMIC-CXR-JPG (*FULL*)	MMMR (T=1.0) **(Ours)**	0,378	0,233	0,157	0,114	**0,321**	**0,281**	0,164
MIMIC-CXR-JPG (*FULL*)	MMMR (T=0.5) **(Ours)**	0,357	0,219	0,147	0,105	0,316	0,269	0,163
MIMIC-CXR-JPG (*FULL*)	MMMR (T=0.1) **(Ours)**	0,346	0,209	0,138	0,097	0,312	0,263	0,159
MIMIC-CXR-JPG (*No History*)	*XRaySwinGen* **(Ours)**	0,309	0,192	0,131	0,093	0,363	0,341	0,229
MIMIC-CXR-JPG (*No History*)	MMMR (T=1.0) **(Ours)**	0,477	0,328	0,242	0,187	0,416	0,384	0,240
MIMIC-CXR-JPG (*No History*)	MMMR (T=0.5) **(Ours)**	0,442	0,301	0,220	0,169	0,415	0,368	0,248
MIMIC-CXR-JPG (*No History*)	MMMR (T=0.1) **(Ours)**	0,411	0,276	0,200	0,151	0,412	0,354	0,252

A análise comparativa dos experimentos evidencia, de modo consistente, o impacto positivo da incorporação de memória relacional no desempenho dos modelos para geração automática de laudos radiológicos. Em todos os cenários avaliados, tanto na base *MIMIC-CXR-JPG* (FULL) quanto na versão NO EXAM HISTORY, a arquitetura MMMR supera o modelo *XRaySwinGen* em todas as métricas consideradas (BLEU, ROUGE-L, METEOR e SPICE). O ganho relativo do MMMR é ainda mais expressivo na ausência de referências históricas nos exames. Esses resultados indicam que, diante de um cenário com menor disponibilidade contextual, a presença do módulo de memória relacional é capaz de estabelecer melhor

alinhamento entre as evidências visuais e a geração textual, promovendo maior coerência e detalhamento semântico.

Além disso, o modelo MMMR obteve desempenho comparável ou superior a arquiteturas de larga escala desenvolvidas por grandes centros de pesquisa. O Radiology-LLaMA2, da Stanford University, baseado no LLaMA 2 da Meta AI, destaca-se com ROUGE-L de 0,4427 na base MIMIC-CXR, embora exija infraestrutura computacional robusta e conte com bilhões de parâmetros. O Flamingo-CXR, proposto pela DeepMind, alcança BLEU-4 de 0,101 e ROUGE-L de 0,297, sendo construído sobre uma arquitetura multimodal proprietária com validação clínica especializada. Já o Med-PaLM M, desenvolvido em parceria entre o Google Health e a DeepMind, com 12 bilhões de parâmetros, registra BLEU-4 de 0,104 e ROUGE-L de 0,262, sendo amplamente reconhecido por seu desempenho em tarefas clínicas e interpretabilidade.

Mesmo diante dessas soluções com alto custo computacional, o MMMR demonstrou resultados superiores em diversos aspectos. Na *MIMIC-CXR-JPG* (FULL), atinge BLEU-4 de 0,114 e ROUGE-L de 0,321. Na configuração NO EXAM HISTORY, alcança BLEU-4 de 0,187 e ROUGE-L de 0,416. Esses resultados consolidam o MMMR como uma alternativa eficaz e viável, com apenas 237 milhões de parâmetros, capaz de produzir laudos mais coerentes, detalhados e semanticamente consistentes, mesmo em cenários com informações clínicas limitadas.

A Tabela 5 ainda sintetiza os resultados quantitativos dos experimentos realizados com o modelo MMMR, destacando o impacto da variação do parâmetro de temperatura na qualidade dos textos gerados. Observa-se que configurações intermediárias de temperatura favorecem um equilíbrio entre diversidade lexical e coerência estrutural, resultando em melhores pontuações nas métricas BLEU, ROUGE-L e METEOR. A mesma tabela também apresenta uma comparação consolidada com modelos de referência da literatura, evidenciando que o MMMR alcança desempenho expressivo mesmo com menor complexidade computacional. O modelo proposto nesta tese figura entre as melhores abordagens avaliadas, ao lado de arquiteturas reconhecidas como *R2Gen, AlignTransformer, CvT-212DistilGPT2, RGRG e LVMed-R2*, demonstrando sua capacidade de gerar laudos detalhados e semanticamente alinhados às evidências visuais, com maior economia de parâmetros e elevada generalização textual.

Em complemento e para melhor visualização da evolução entre as arquiteturas experimentadas, a Figura 7 apresenta uma visualização comparativa do desempenho dos modelos *XRaySwinGen* e MMMR sobre o conjunto MIMIC-CXR-JPG, nos cenários com e sem acesso ao histórico clínico. Observa-se que, em ambos os contextos, o modelo MMMR supera consistentemente o *XRaySwinGen* em todas as métricas avaliadas, evidenciando a contribuição do módulo de memória relacional para retenção de contexto e geração textual mais coerente. No cenário sem histórico clínico, a superioridade do MMMR se torna ainda mais expressiva, especialmente nas métricas BLEU-4 e ROUGE-L, que avaliam precisão lexical e cobertura sintática. Essa tendência reforça a capacidade do MMMR de operar com maior autonomia textual, mesmo diante de entradas visuais descontextualizadas, o que o torna mais robusto para uso clínico em situações de escassez de informação auxiliar.

Outro resultado de grande relevância prática foi a constatação de que, mesmo utilizando o *backbone Swin-Tiny*, uma arquitetura menor, eficiente e com menor número de parâmetros, o MMMR supera o *XRaySwinGen*. Isso demonstra que a adição estratégica do módulo de memória relacional compensa a menor capacidade da espinha dorsal e permite a geração de laudos automáticos de alta qualidade com menor custo computacional. Este resultado evidencia a eficiência da proposta tanto do ponto de vista de desempenho quanto de viabilidade para aplicação em cenários reais.

O ganho mais expressivo foi observado na métrica BLEU-4, em que o MMMR praticamente dobra o valor obtido pelo *XRaySwinGen* na base *FULL* e mais que dobra na versão *NO EXAM HISTORY*. Esse comportamento reflete a capacidade do módulo de memória relacional em capturar dependências de longo

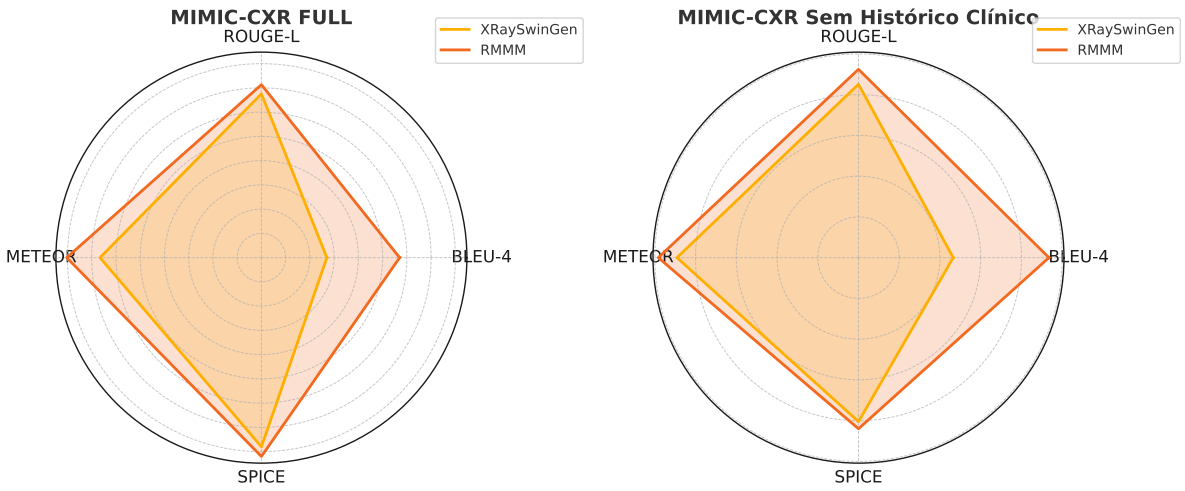


Figura 7 – Comparativo visual do desempenho entre os modelos *XRaySwinGen* e MMMR nos cenários *MIMIC-CXR-JPG* com e sem histórico clínico, considerando as métricas BLEU-4, ROUGE-L, METEOR e SPICE.

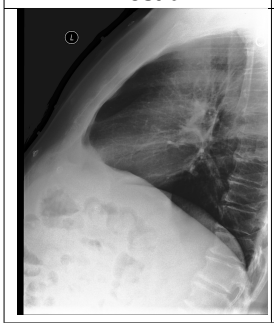
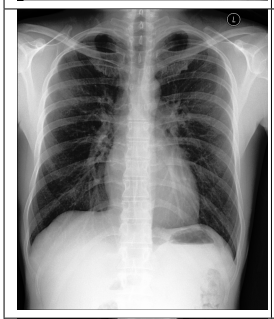
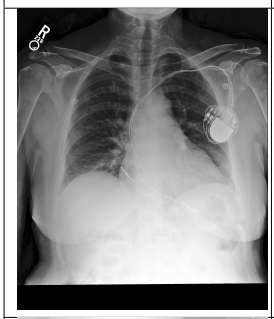
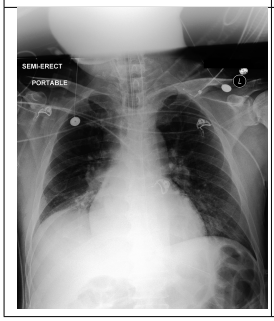
alcance e gerar frases mais extensas, precisas e coerentes. As melhorias em ROUGE-L, METEOR e SPICE indicam que a memória relacional contribui não apenas para a precisão de n-gramas, mas também para a similaridade semântica e para a representação acurada das relações entre entidades clínicas no texto. O sucesso do MMMR reside na sinergia entre o codificador visual baseado em *Swin Transformer*, a arquitetura codificador-decodificador otimizada e o módulo de memória relacional, o que possibilita a extração de características visuais relevantes, a geração de textos clinicamente ricos e a manutenção de coerência contextual ao longo do pré-laudo médico. Esses resultados reforçam o potencial do MMMR para aprimorar a eficiência diagnóstica, reduzir a sobrecarga dos radiologistas e estabelecer um novo patamar de qualidade e eficiência computacional para sistemas automáticos de geração de laudos médicos.

A análise quantitativa apresentada nas Tabelas 4 e 5, e visualmente consolidada na Figura 7, demonstra a superioridade consistente do modelo MMMR em métricas de avaliação consolidadas, superando o XRaySwinGen em todos os cenários. Contudo, para uma compreensão mais aprofundada da relevância clínica, é crucial ir além dos números. Uma análise qualitativa se faz necessária não apenas para ilustrar os acertos do modelo, mas também para identificar suas limitações e áreas que necessitam de aprimoramento. Nesse sentido, a Tabela 6 a seguir expõe exemplos concretos de pré-laudos gerados pelo MMMR em comparação direta com os laudos de referência (*ground truth*) elaborados por radiologistas. Esta comparação permite observar em detalhe a capacidade do modelo em identificar achados e a fidelidade da descrição, ao mesmo tempo que evidencia os pontos onde a geração textual ainda carece de precisão ou fluidez, oferecendo uma perspectiva prática e crítica sobre o desempenho anteriormente aferido pelas métricas.

5.4 Considerações Finais

Os resultados deste capítulo evidenciam o impacto positivo de duas contribuições principais desta tese: a introdução do janelamento hierárquico para extração visual eficiente, por meio do modelo *XRaySwinGen* baseado em *Swin Transformer*, e o avanço proporcionado pela incorporação do mecanismo

Tabela 6 – Exemplos de laudos gerados pelo modelo *MMMR* em comparação com o laudo de referência (*ground truth*).

Amostra	Laudo de Referência (Radiologista)	Laudo Gerado (RMMM)
	vistas póstero-anterior e lateral do tórax fornecidas. não há consolidação focal, derrame ou pneumotórax. a silhueta cardiomediastinal é *normal*. as estruturas ósseas visualizadas estão *intactas*. não há ar livre abaixo do hemidiafragma direito.	vistas póstero-anterior e lateral do tórax fornecidas. não há consolidação focal, derrame ou pneumotórax. a silhueta cardiomediastinal é *normal*. as estruturas ósseas visualizadas estão *intactas*. não há ar livre abaixo do hemidiafragma direito.
	o tamanho do coração é *normal*. os contornos do mediastino e do hilo são *normais*. a vasculatura pulmonar é *normal*. os pulmões estão *limpos*. não se observa derrame pleural ou pneumotórax. não há anormalidades ósseas agudas.	o tamanho do coração é *normal*. os contornos do mediastino e do hilo são *normais*. a vasculatura pulmonar é *normal*. os pulmões estão *limpos*. **cicatrizes bibasilares leves são simétricas**. não se observa derrame pleural ou pneumotórax. não há anormalidades ósseas agudas.
	dispositivo de marcapasso de câmara dupla do lado esquerdo é observado com eletrodos terminando no átrio direito e ventrículo direito. **nota-se cardiomegalia leve**. os contornos do mediastino e do hilo são *sem particularidades*. **opacidades irregulares nas bases pulmonares provavelmente refletem áreas de atelectasia**. a vasculatura pulmonar *não está ingurgitada*. não há derrame pleural, consolidação focal ou pneumotórax presente.	dispositivo de marcapasso de câmara dupla do lado esquerdo é observado com eletrodos terminando no átrio direito e ventrículo direito. **aumento leve a moderado da silhueta cardíaca é redemonstrado**. os contornos do mediastino e do hilo estão *inalterados*. a vasculatura pulmonar *não está ingurgitada*. **opacidades irregulares nas bases pulmonares provavelmente refletem áreas de atelectasia**. sem derrame pleural ou pneumotórax.
	o tubo endotraqueal termina aproximadamente 4 cm acima da carina e está posicionado adequadamente e o tubo orogástrico passa abaixo do diafragma para o estômago. uma **opacidade mal definida na base do pulmão direito na radiografia anterior foi resolvida, sugerindo que era uma aspiração ou atelectasia**. não há novas **opacidades pulmonares preocupantes**. tamanho do coração e mediastino.	a ponta do tubo endotraqueal está a aproximadamente 4,5 cm da carina. o tubo entérico passa abaixo do campo de visão. **observam-se baixos volumes pulmonares com agrupamento secundário das marcas broncovasculares**. **opacidade basilar esquerda pode ser devida a atelectasia**. a silhueta cardiomediastinal está *dentro dos limites normais*. não há anormalidades ósseas agudas.

de memória relacional, central na arquitetura MMMR. Enquanto o *XRaySwinGen* já estabelece um novo patamar na captura de padrões visuais relevantes por meio de janelamento, a evolução para o MMMR, com o acréscimo da memória relacional, potencializa o alinhamento contextual entre imagem e texto, elevando significativamente o desempenho em métricas de similaridade, precisão semântica e fluidez textual. Esses achados demonstram que a combinação entre técnicas avançadas de representação visual e mecanismos explícitos de memória é decisiva para o desenvolvimento de sistemas multimodais robustos e adaptáveis à prática clínica, consolidando um novo padrão de excelência na geração automática de laudos médicos.

6 Conclusão e Trabalhos Futuros

Esta tese apresentou uma abordagem multimodal para geração automática de pré-laudos médicos textuais a partir de radiografias de tórax, integrando técnicas de Visão Computacional e Processamento de Linguagem Natural fundamentadas na arquitetura *Transformer*. O desenvolvimento da pesquisa foi estruturado em quatro eixos principais: (i) a utilização do janelamento hierárquico no *Swin Transformer*, que viabiliza a captura eficiente de padrões visuais relevantes; (ii) a exploração de bases de dados bilíngues, com laudos em português e inglês, ampliando a abrangência da aplicação; (iii) o emprego de mecanismos de extração e alinhamento entre texto e imagem, promovendo a integração semântica entre as modalidades; e (iv) a incorporação de um módulo de memória relacional, elemento central na arquitetura MMMR, voltado à melhoria da coerência contextual e da fluidez textual ao longo da geração, superando limitações observadas em arquiteturas sequenciais anteriores.

Todos os objetivos traçados na introdução foram plenamente atingidos ao longo da tese. Inicialmente, foi concebida e implementada a arquitetura *XRaySwinGen*, que combinou com sucesso o *Swin Transformer* para codificação visual e o GPT-2 como decodificador textual, interligados por camadas de atenção cruzada. Esta integração permitiu a geração de pré-laudos médicos coerentes e semanticamente alinhados com o conteúdo das imagens, mesmo quando aplicados a diferentes bases de dados. A partir dessa base, o modelo foi expandido com a introdução do módulo de memória relacional, originando o modelo MMMR, capaz de reter contexto ao longo da geração textual. Este aprimoramento solucionou um dos principais desafios observados na fase inicial: a perda de coerência narrativa em laudos longos.

Com a nova arquitetura, observou-se uma duplicação dos escores BLEU-4 na base MIMIC-CXR-JPG (de 0,054 para 0,114), aumento expressivo em ROUGE-L (de 0,304 para 0,321) e METEOR (de 0,233 para 0,281), além de melhora na métrica semântica SPICE (de 0,156 para 0,164), evidenciando a relevância da memória explícita no processo de geração. Em cenários de maior restrição — como a versão da base sem histórico clínico, os ganhos foram ainda mais evidentes, com BLEU-4 alcançando 0,187 e ROUGE-L chegando a 0,416. Adicionalmente, o modelo manteve sua eficiência computacional, com 237M de parâmetros, mesmo ao substituir o *Swin-Base* pelo *Swin-Tiny*, provando que a evolução arquitetural proposta é eficaz mesmo com *backbones* mais leves.

Os principais obstáculos envolveram o pré-processamento textual, especialmente em laudos clínicos reais contendo linguagem inconsistente ou ruído, a escassez de bases multilíngues anotadas e os desafios de treinamento com imagens em alta resolução. Ainda assim, as estratégias adotadas (como tokenização adaptada, normalização robusta e validação cruzada 3-*fold*) possibilitaram alta reprodutibilidade dos resultados, sugerindo a viabilidade de teste da arquitetura em contextos clínicos reais.

Entre as limitações enfrentadas, destaca-se a escassez de bases de dados públicas com laudos médicos detalhados e padronizados em linguagem natural, bem como o desafio de generalização para idiomas com menor representatividade nos grandes corpora. Este desafio é amplamente reconhecido na área, motivando o desenvolvimento de modelos de pré-treinamento específicos para o domínio médico que buscam maximizar o aprendizado a partir de dados disponíveis (GOWDA et al., 2025). Outro obstáculo foi o custo computacional do treinamento com imagens em alta resolução, exigindo balanceamento criterioso entre desempenho e eficiência.

As principais lições apreendidas desta pesquisa abrangem aspectos técnicos e metodológicos centrais: (i) O pré-processamento rigoroso dos dados textuais e de imagem foi essencial para garantir consistência e qualidade na entrada do modelo; (ii) As decisões arquiteturais tiveram impacto direto na

qualidade dos laudos, especialmente no alinhamento entre imagem e texto; (iii) A inclusão de memória explícita, por meio do módulo relacional, foi decisiva para manter coerência e continuidade nos textos gerados; (iv) O uso de métricas complementares, como BLEU, ROUGE, METEOR e SPICE, mostrou-se fundamental para uma avaliação abrangente da geração textual; (v) Por fim, a eficiência computacional provou ser estratégica, uma vez que o MMMR superou modelos maiores mesmo com menos parâmetros, demonstrando que soluções enxutas podem ser mais eficazes e viáveis na prática clínica.

Como resultado adicional desta tese, foi desenvolvida uma webpage interativa capaz de receber uma imagem de RX e, a partir dela, gerar automaticamente informações textuais correspondentes ao diagnóstico médico associado. Essa webpage foi concebida como uma prova de conceito da arquitetura proposta e encontra-se atualmente hospedada na plataforma Hugging Face[1]. A Figura 8 ilustra a interface principal da aplicação em operação.

Figura 8 – *Webpage* criada para prova de conceito dos modelos *XRaySwinGen* e MMMR.



Fonte: Elaborado pelo autor

[1] Disponível em <https://huggingface.co/spaces/Gilvan/XRaySwinGen-RMMM>

6.1 Contribuições

As contribuições desta tese, detalhadas e validadas ao longo dos capítulos anteriores, representam avanços significativos no campo da geração automática de laudos radiológicos. O percurso da pesquisa culminou em uma solução robusta, eficiente e inovadora, cujas principais contribuições são sumarizadas a seguir:

- Foi projetado e validado um extrator de características visuais específico para radiografias de tórax, baseado no *Swin Transformer*. Demonstrou-se a eficácia da atenção hierárquica com janelas deslocadas para capturar, de forma integrada, tanto os detalhes finos quanto os contextos globais da imagem médica, estabelecendo uma base sólida para a subsequente tradução textual.
- Desenvolveu-se a arquitetura *XRaySwinGen*, que serviu como uma robusta linha de base multimodal. Este modelo comprovou a viabilidade de integrar eficazmente um codificador visual avançado a um decodificador de linguagem generativo (GPT-2) por meio de atenção cruzada, alcançando resultados competitivos.
- A principal contribuição metodológica desta tese foi o desenvolvimento da arquitetura inovadora MMMR (Modelo Multimodal de Memória Relacional). Este modelo solucionou a limitação crítica da retenção de contexto em laudos longos através da incorporação de um módulo de memória relacional explícito, resultando em textos com um ganho significativo de coerência, fluidez e fidelidade clínica, conforme atestado pelos resultados quantitativos.
- Aplicou-se, de forma inovadora para a tarefa de geração de laudos médicos para imagens de radiografia de tórax, a técnica de janelamento deslocado (*shifted window*) como mecanismo primário de extração de características visuais, representando uma contribuição original à área e demonstrando sua superioridade em relação a abordagens puramente convolucionais ou baseadas em ViT padrão.
- Demonstrou-se que uma arquitetura computacionalmente eficiente (237M de parâmetros) pode alcançar desempenho comparável ou superior a modelos de larga escala com bilhões de parâmetros. Esta contribuição endereça diretamente o desafio da aplicabilidade, oferecendo uma solução mais viável e escalável para implementação em ambientes clínicos com recursos computacionais limitados.
- Foi desenvolvida e disponibilizada uma prova de conceito funcional em uma interface web, que não apenas valida a aplicabilidade em tempo real dos modelos propostos, mas também serve como uma ponte entre a pesquisa acadêmica e a prática clínica, facilitando a demonstração, a validação e a futura adoção da tecnologia.

Em conjunto, estas contribuições não apenas avançam o estado da arte na geração automática de laudos, mas também o fazem de maneira pragmática e aplicável. Ao resolver o desafio da coerência textual de longo prazo por meio de um mecanismo de memória explícito e, ao mesmo tempo, manter a eficiência computacional, este trabalho estabelece um novo paradigma para o desenvolvimento de sistemas multimodais clinicamente relevantes, demonstrando que a inovação arquitetural estratégica pode superar a necessidade de modelos massivos, tornando a IA uma ferramenta mais acessível e impactante para o futuro da radiologia.

6.2 Trabalhos Futuros

Embora tenham sido obtidos resultados competitivos na geração automática de pré-laudos médicos textuais a partir de imagens de RX, ainda há espaço para avanços importantes. Algumas atividades futuras são sugeridas com o objetivo de aprimorar a robustez, generalização e aplicabilidade prática do modelo desenvolvido:

1. Ampliar a avaliação em bancos de dados multilíngues, com laudos em idiomas como espanhol, árabe e mandarim, e também em diferentes modalidades de imagem, como tomografias (CT), ressonâncias magnéticas (MRI) e ultrassom, para testar a generalização do modelo em domínios diagnósticos diversos, seguindo a linha de estudos comparativos multimodais recentes (SHARMA et al., 2025).
2. Realizar estudos com avaliação humana, conduzidos por radiologistas ou profissionais da saúde, para complementar as métricas automáticas com análises clínicas qualitativas sobre consistência, completude e utilidade dos laudos gerados.
3. Investigar técnicas de aprendizado auto-supervisionado, pseudo-rotulagem e aumento de dados (SHORTEN; KHOSHGOFTAAR, 2019), como forma de melhorar a robustez e reduzir a dependência de grandes volumes de dados anotados manualmente, explorando abordagens como as de pré-treinamento de visão-linguagem que utilizam laudos estruturados para enriquecer o contexto clínico (GOWDA et al., 2025).
4. Avaliar o impacto da arquitetura em cenários clínicos reais, com integração ao fluxo de trabalho clínico e validação com base no uso em tempo real por especialistas.
5. Realizar experimentos com variações arquiteturais no codificador, decodificador e módulos de memória alternativos, incluindo profundidade de camadas, temperatura, tamanhos de janelas do deslocamento de janelas e dimensões de *embedding*, além de análises específicas de desempenho usando apenas o codificador, por exemplo, em tarefas de classificação auxiliar, em linha com estudos que comparam a eficácia de diferentes *backbones*, como RNC e *Transformers*, para tarefas de imagem médica (KANDASAMY, 2023).

Tais direções apontam para um contínuo amadurecimento do modelo proposto, com vistas à sua consolidação como uma ferramenta de apoio confiável, escalável e clinicamente relevante.

6.3 Considerações Éticas e Limitações Clínicas

Apesar dos avanços observados, é essencial considerar os riscos associados ao uso de modelos generativos em domínios clínicos sensíveis. A possibilidade de alucinação factual, entendida como a geração de informações incorretas embora linguisticamente plausíveis, representa um desafio relevante, sobretudo em contextos diagnósticos. Para mitigar tais riscos, a fronteira da pesquisa atual explora ativamente métodos para ancorar a geração em conhecimento validado, seja através da injeção de conhecimento suplementar (HOU et al., 2025), da sinergia com sistemas baseados em regras (FYTAS et al., 2024), ou de frameworks de geração controlável que integram conhecimento médico durante a inferência (HOU et al., 2024). Além disso, a ausência de responsabilização automatizada e a necessidade de supervisão humana contínua reforçam a importância de uma integração ética e transparente dessas tecnologias. Assim, os modelos propostos devem ser empregados como ferramentas de apoio, e não como substitutos da interpretação médica qualificada, respeitando os princípios de beneficência, não maleficência e autonomia clínica.

6.4 Produção Científica

Durante o desenvolvimento desta tese, foram aceitos os seguintes artigos científicos relacionados ao tema:

- Veras Magalhães G Junior, L. de S. Santos R, H. S. Vogado L, Cardoso de Paiva A, de Alcântara dos Santos Neto P. *XRaySwinGen: Automatic medical reporting for X-ray exams with multimodal model*. Heliyon. 2024 Mar 12;10(7):e27516. doi: 10.1016/j.heliyon.2024.e27516. PMCID: PMC10979158.
- Veras Magalhães G Junior, Cardoso de Paiva A, de Alcântara dos Santos Neto P. Relational Memory Multimodal Models for Automated X-Ray Medical Reporting. In: Proceedings of the 14th International Conference on Health Informatics (HCist 2025).

Referências

AGARWAL, L.; VERMA, B. Utilizing transformer-based image descriptors for improving chest x-ray analysis. In: SPRINGER. **International Conference on Data Science and Applications**. [S.l.], 2024. p. 573–586. Nenhuma citação no texto.

ALAYRAC, J.-B.; DONAHUE, J.; LUC, P.; MIECH, A.; BARR, I.; HASSON, Y.; LENC, K.; MENSCH, A.; MILLICAN, K.; REYNOLDS, M. et al. Flamingo: a visual language model for few-shot learning. **Advances in neural information processing systems**, v. 35, p. 23716–23736, 2022. Nenhuma citação no texto.

ALFARGHALY, O.; KHALED, R.; ELKORANY, A.; HELAL, M.; FAHMY, A. Automated radiology report generation using conditioned transformers. **Informatics in Medicine Unlocked**, Elsevier, v. 24, p. 100557, 2021. Nenhuma citação no texto.

ALLAOUZI, I.; AHMED, M. B.; BENAMROU, B.; OUARDOUZ, M. Automatic caption generation for medical images. In: **Proceedings of the 3rd International Conference on Smart City Applications**. New York, NY, USA: Association for Computing Machinery, 2018. p. 1–6. Nenhuma citação no texto.

ANDERSON, P.; FERNANDO, B.; JOHNSON, M.; GOULD, S. Spice: Semantic propositional image caption evaluation. In: SPRINGER. **Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14**. [S.l.], 2016. p. 382–398. Nenhuma citação no texto.

AZAD, R.; KAZEROUNI, A.; HEIDARI, M.; AGHDAM, E. K.; MOLAEI, A.; JIA, Y.; JOSE, A.; ROY, R.; MERHOF, D. **Advances in Medical Image Analysis with Vision Transformers: A Comprehensive Review**. arXiv, 2023. Disponível em: <https://arxiv.org/abs/2301.03505>. Nenhuma citação no texto.

BA, J. L.; KIROS, J. R.; HINTON, G. E. **Layer Normalization**. [S.l.]: arXiv, 2016. Nenhuma citação no texto.

BANERJEE, S.; LAVIE, A. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: **Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization**. [S.l.: s.n.], 2005. p. 65–72. Nenhuma citação no texto.

BISWAS, S. **ChatGPT and the future of medical writing**. [S.l.]: Radiological Society of North America, 2023. e223312 p. Nenhuma citação no texto.

ÇALLI, E.; SOGANCIOGLU, E.; GINNEKEN, B. van; LEEUWEN, K. G. van; MURPHY, K. Deep learning for chest x-ray analysis: A survey. **Medical Image Analysis**, Elsevier, v. 72, p. 102125, 2021. Nenhuma citação no texto.

CASTRO, R.; PINEDA, I.; LIM, W.; MOROCHO-CAYAMCELA, M. E. Deep learning approaches based on transformer architectures for image captioning tasks. **IEEE Access**, IEEE, v. 10, p. 33679–33694, 2022. Nenhuma citação no texto.

CHEDDI, F.; HABBANI, A.; NAIT-CHARIF, H. Improving the cxr reports generation with multi-modal feature alignment and self-refining strategy. In: IEEE. **2024 3rd International Conference on Embedded Systems and Artificial Intelligence (ESAI)**. [S.l.], 2024. p. 1–7. Nenhuma citação no texto.

CHEDDI, F.; HABBANI, A.; NAIT-CHARIF, H. A multi-modal feature fusion-based approach for chest x-ray report generation. In: IEEE. **2024 11th International Conference on Wireless Networks and Mobile Communications (WINCOM)**. [S.l.], 2024. p. 1–7. Nenhuma citação no texto.

CHEN, X.; FANG, H.; LIN, T.-Y.; VEDANTAM, R.; GUPTA, S.; DOLLÁR, P.; ZITNICK, C. L. Microsoft coco captions: Data collection and evaluation server. **arXiv preprint arXiv:1504.00325**, 2015. Nenhuma citação no texto.

CHEN, Z.; SONG, Y.; CHANG, T.-H.; WAN, X. Generating radiology reports via memory-driven transformer. In: **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Association for Computational Linguistics, 2020. p. 1439–1449. Disponível em: <https://aclanthology.org/2020.emnlp-main.112>. Nenhuma citação no texto.

DELBROUCK, J.-B.; ZHANG, C.; RUBIN, D. Qiai at mediqa 2021: Multimodal radiology report summarization. In: **Proceedings of the 20th workshop on biomedical language processing**. [S.l.: s.n.], 2021. p. 285–290. Nenhuma citação no texto.

DEMNER-FUSHMAN, D.; KOHLI, M. D.; ROSENMAN, M. B.; SHOOSHAN, S. E.; RODRIGUEZ, L.; ANTANI, S.; THOMA, G. R.; MCDONALD, C. J. Preparing a collection of radiology examinations for distribution and retrieval. **Journal of the American Medical Informatics Association**, Oxford University Press, v. 23, n. 2, p. 304–310, 2016. Nenhuma citação no texto.

DENG, J.; DONG, W.; SOCHER, R.; LI, L.-J.; LI, K.; FEI-FEI, L. Imagenet: A large-scale hierarchical image database. In: IEEE. **2009 IEEE conference on computer vision and pattern recognition**. [S.l.], 2009. p. 248–255. Nenhuma citação no texto.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. [S.l.]: Association for Computational Linguistics, 2019. p. 4171–4186. Nenhuma citação no texto.

DICUONZO, G.; DONOFRIO, F.; FUSCO, A.; SHINI, M. Healthcare system: Moving forward with artificial intelligence. **Technovation**, Elsevier, v. 120, p. 102510, 2023. Nenhuma citação no texto.

DOSOVITSKIY, A.; BEYER, L.; KOLESNIKOV, A.; WEISSENBORN, D.; ZHAI, X.; UNTERTHINER, T.; DEHGHANI, M.; MINDERER, M.; HEIGOLD, G.; GELLY, S. et al. An image is worth 16x16 words: Transformers for image recognition at scale. **arXiv preprint arXiv:2010.11929**, 2020. Nenhuma citação no texto.

DOSOVITSKIY, A.; BEYER, L.; KOLESNIKOV, A.; WEISSENBORN, D.; ZHAI, X.; UNTERTHINER, T.; DEHGHANI, M.; MINDERER, M.; HEIGOLD, G.; GELLY, S.; USZKOREIT, J.; HOULSBY, N. **An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale**. [S.l.]: arXiv, 2020. Nenhuma citação no texto.

FAROOQ, M.; HAFEEZ, A. **COVID-ResNet: A Deep Learning Framework for Screening of COVID19 from Radiographs**. [S.l.]: arXiv, 2020. Nenhuma citação no texto.

FATIMA, N.; IMRAN, A. S.; KASTRATI, Z.; DAUDPOTA, S. M.; SOOMRO, A. A systematic literature review on text generation using deep neural network models. **IEEE Access**, v. 10, p. 53490–53503, 2022. Nenhuma citação no texto.

FYTAS, P.; BREGER, A.; SELBY, I.; BAKER, S.; SHAHIPASAND, S.; KORHONEN, A. Can rule-based insights enhance llms for radiology report classification? introducing the radprompt methodology. **arXiv preprint arXiv:2408.04121**, 2024. Nenhuma citação no texto.

GAGE, P. A new algorithm for data compression. **C Users Journal**, McPherson, KS: R & D Publications, c1987-1994., v. 12, n. 2, p. 23–38, 1994. Nenhuma citação no texto.

GAJBHIYE, G. O.; NANDEDKAR, A. V.; FAYE, I. Automatic report generation for chest x-ray images: A multilevel multi-attention approach. In: NAIN, N.; VIPPARTHI, S. K.; RAMAN, B. (Ed.). **Computer Vision and Image Processing**. [S.l.]: Springer Singapore, 2020. p. 174–182. Nenhuma citação no texto.

GHEFLATI, B.; RIVAZ, H. Vision transformers for classification of breast ultrasound images. In: IEEE. **2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)**. [S.l.], 2022. p. 480–483. Nenhuma citação no texto.

GOKASLAN, A.; COHEN, V.; PAVLICK, E.; TELLEX, S. **OpenWebText Corpus**. 2019. <http://Skylion007.github.io/OpenWebTextCorpus>. Nenhuma citação no texto.

GOWDA, S. N.; ZHANG, R.; GU, X.; WENG, Y.; YANG, L. Distribution-based masked medical vision-language model using structured reports. **arXiv preprint arXiv:2507.21794**, 2025. Nenhuma citação no texto.

HABIB, G.; SALEEM, T. J.; LALL, B. **Knowledge Distillation in Vision Transformers: A Critical Review**. arXiv, 2023. Disponível em: <https://arxiv.org/abs/2302.02108>. Nenhuma citação no texto.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2016. p. 770–778. Nenhuma citação no texto.

HENDRYCKS, D.; GIMPEL, K. **Gaussian Error Linear Units (GELUs)**. arXiv, 2016. Disponível em: <https://arxiv.org/abs/1606.08415>. Nenhuma citação no texto.

HOU, W.; CHENG, Y.; XU, K.; LI, H.; HU, Y.; LI, W.; LIU, J. Radar: Enhancing radiology report generation with supplementary knowledge injection. **arXiv preprint arXiv:2505.14318**, 2025. Nenhuma citação no texto.

HOU, Z.; YAN, R.; YAN, Z.; LANG, N.; ZHOU, X. Energy-based controllable radiology report generation with medical knowledge. In: SPRINGER. **International Conference on Medical Image Computing and Computer-Assisted Intervention**. [S.l.], 2024. p. 240–250. Nenhuma citação no texto.

HUANG, G.; LIU, Z.; MAATEN, L. V. D.; WEINBERGER, K. Q. Densely connected convolutional networks. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2017. p. 4700–4708. Nenhuma citação no texto.

HUANG, X.; YAN, F.; XU, W.; LI, M. Multi-attention and incorporating background information model for chest x-ray image report generation. **IEEE Access**, v. 7, p. 154808–154817, 2019. Nenhuma citação no texto.

JING, B.; WANG, Z.; XING, E. Show, describe and conclude: On exploiting the structure information of chest X-ray reports. In: **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**. [S.l.]: Association for Computational Linguistics, 2019. p. 6570–6580. Nenhuma citação no texto.

JING, B.; WANG, Z.; XING, E. Show, describe and conclude: On exploiting the structure information of chest x-ray reports. **arXiv preprint arXiv:2004.12274**, 2020. Nenhuma citação no texto.

JING, B.; XIE, P.; XING, E. On the automatic generation of medical imaging reports. In: **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Melbourne, Australia: Association for Computational Linguistics, 2018. p. 2577–2586. Nenhuma citação no texto.

JOHNSON, A. E. W.; POLLARD, T. J.; BERKOWITZ, S. J.; GREENBAUM, N. R.; LUNGREN, M. P.; DENG, C.; MARK, R. G.; HORNG, S. MIMIC-CXR: A large publicly available database of labeled chest radiographs. **CoRR**, abs/1901.07042, 2019. Disponível em: <http://arxiv.org/abs/1901.07042>. Nenhuma citação no texto.

KANDASAMY, H. P. **Enhancing chest X-ray analysis: a comparative study of deep learning models with explainable AI**. Tese (Doutorado) — University of Texas at Austin, 2023. Nenhuma citação no texto.

KAUR, P.; HARNAL, S.; TIWARI, R.; ALHARITHI, F. S.; ALMULIHI, A. H.; NOYA, I. D.; GOYAL, N. A hybrid convolutional neural network model for diagnosis of covid-19 using chest x-ray images. **International Journal of Environmental Research and Public Health**, MDPI, v. 18, n. 22, p. 12191, 2021. Nenhuma citação no texto.

KIM, S.; RIM, B.; CHOI, S.; LEE, A.; MIN, S.; HONG, M. Deep learning in multi-class lung diseases' classification on chest x-ray images. **Diagnostics**, MDPI, v. 12, n. 4, p. 915, 2022. Nenhuma citação no texto.

KOLESNIKOV, A.; BEYER, L.; ZHAI, X.; PUIGCERVER, J.; YUNG, J.; GELLY, S.; HOULSBY, N. Big transfer (bit): General visual representation learning. In: **Computer Vision – ECCV 2020**. [S.l.]: Springer International Publishing, 2020. p. 491–507. Nenhuma citação no texto.

KRALJEVIC, Z.; SHEK, A.; BEAN, D.; BENDAYAN, R.; TEO, J.; DOBSON, R. Medgpt: Medical concept prediction from clinical narratives. **arXiv preprint arXiv:2107.03134**, 2021. Nenhuma citação no texto.

KUMAR, R.; KARTHIK, K.; KAMATH, S. S. Gan-based encoder-decoder model for multi-label diagnostic scan classification and automated radiology report generation. In: **Handbook of AI-based models in healthcare and medicine**. [S.l.]: CRC Press, 2024. p. 93–109. Nenhuma citação no texto.

KUMAR, S.; CHAUBE, M. K.; ALSAMHI, S. H.; GUPTA, S. K.; GUIZANI, M.; GRAVINA, R.; FORTINO, G. A novel multimodal fusion framework for early diagnosis and accurate classification of covid-19 patients using x-ray images and speech signal processing techniques. **Computer methods and programs in biomedicine**, Elsevier, v. 226, p. 107109, 2022. Nenhuma citação no texto.

KUMAR, S.; GUPTA, S. K.; KUMAR, V.; KUMAR, M.; CHAUBE, M. K.; NAIK, N. S. Ensemble multimodal deep learning for early diagnosis and accurate classification of covid-19. **Computers and Electrical Engineering**, Elsevier, v. 103, p. 108396, 2022. Nenhuma citação no texto.

LAVIE, A.; AGARWAL, A. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In: **Proceedings of the Second Workshop on Statistical Machine Translation**. [S.l.]: Association for Computational Linguistics, 2007. p. 228–231. Nenhuma citação no texto.

LEE, H.; CHO, H.; PARK, J.; CHAE, J.; KIM, J. Cross encoder-decoder transformer with global-local visual extractor for medical image captioning. **Sensors**, MDPI, v. 22, n. 4, p. 1429, 2022. Nenhuma citação no texto.

LI, C. Y.; LIANG, X.; HU, Z.; XING, E. P. Hybrid retrieval-generation reinforced agent for medical image report generation. In: **Proceedings of the 32nd International Conference on Neural Information Processing Systems**. [S.l.]: Curran Associates Inc., 2018. p. 1537–1547. Nenhuma citação no texto.

LI, Y.; WANG, H.; LUO, Y. A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports. In: IEEE. **2020 IEEE international conference on bioinformatics and biomedicine (BIBM)**. [S.l.], 2020. p. 1999–2004. Nenhuma citação no texto.

LI, Z.; REN, J. Fine-tuning ernie for chest abnormal imaging signs extraction. **Journal of Biomedical Informatics**, Elsevier, v. 108, p. 103492, 2020. Nenhuma citação no texto.

LIN, C.-Y. ROUGE: A package for automatic evaluation of summaries. In: **Text Summarization Branches Out**. Association for Computational Linguistics, 2004. p. 74–81. Disponível em: <https://aclanthology.org/W04-1013>. Nenhuma citação no texto.

LIU, F.; WU, X.; GE, S.; FAN, W.; ZOU, Y. Exploring and distilling posterior and prior knowledge for radiology report generation. In: **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**. [S.l.: s.n.], 2021. p. 13753–13762. Nenhuma citação no texto.

LIU, Z.; LI, Y.; SHU, P.; ZHONG, A.; YANG, L.; JU, C.; WU, Z.; MA, C.; LUO, J.; CHEN, C. et al. Radiology-llama2: Best-in-class large language model for radiology. **arXiv preprint arXiv:2309.06419**, 2023. Nenhuma citação no texto.

LIU, Z.; LIN, Y.; CAO, Y.; HU, H.; WEI, Y.; ZHANG, Z.; LIN, S.; GUO, B. Swin transformer: Hierarchical vision transformer using shifted windows. In: **Proceedings of the IEEE/CVF international conference on computer vision**. [S.l.: s.n.], 2021. p. 10012–10022. Nenhuma citação no texto.

LOSHCHILOV, I.; HUTTER, F. **Decoupled Weight Decay Regularization**. [S.l.]: arXiv, 2017. Nenhuma citação no texto.

MANNING, C. D.; SURDEANU, M.; BAUER, J.; FINKEL, J. R.; BETHARD, S.; MCCLOSKY, D. The stanford corenlp natural language processing toolkit. In: **Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations**. [S.l.: s.n.], 2014. p. 55–60. Nenhuma citação no texto.

MIURA, Y.; ZHANG, Y.; TSAI, E.; LANGLOTZ, C.; JURAFSKY, D. Improving factual completeness and consistency of image-to-text radiology report generation. In: **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. Association for Computational Linguistics, 2021. p. 5288–5304. Disponível em: <https://aclanthology.org/2021.naacl-main.416>. Nenhuma citação no texto.

MOHSAN, M. M.; AKRAM, M. U.; RASOOL, G.; ALGHAMDI, N. S.; BAQAI, M. A. A.; ABBAS, M. Vision transformer and language model based radiology report generation. **IEEE Access**, IEEE, v. 11, p. 1814–1824, 2022. Nenhuma citação no texto.

MONSHI, M. M. A.; POON, J.; CHUNG, V. Deep learning in generating radiology reports: A survey. **Artificial Intelligence in Medicine**, v. 106, p. 101878, 2020. ISSN 0933-3657. Nenhuma citação no texto.

MUHARRAM, A. P.; HARYONO, H. P.; JUMA, A. H.; PUSPASARI, I.; UTAMA, N. P. Automated chest x-ray report generator using multi-model deep learning approach. **arXiv e-prints**, p. arXiv–2310, 2023. Nenhuma citação no texto.

NAWAZ, M.; NAZIR, T.; BAILI, J.; KHAN, M. A.; KIM, Y. J.; CHA, J.-H. Cxray-effdet: chest disease detection and classification from x-ray images using the efficientdet model. **Diagnostics**, MDPI, v. 13, n. 2, p. 248, 2023. Nenhuma citação no texto.

NGUYEN, H.; NIE, D.; BADAMDORJ, T.; LIU, Y.; ZHU, Y.; TRUONG, J.; CHENG, L. Automated generation of accurate & fluent medical X-ray reports. In: **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**. Association for Computational Linguistics, 2021. p. 3552–3569. Disponível em: <https://aclanthology.org/2021.emnlp-main.288>. Nenhuma citação no texto.

NICOLSON, A.; DOWLING, J.; KOOPMAN, B. Improving chest x-ray report generation by leveraging warm starting. **Artificial intelligence in medicine**, Elsevier, v. 144, p. 102633, 2023. Nenhuma citação no texto.

PANDEY, B.; Kumar Pandey, D.; Pratap Mishra, B.; RHMANN, W. A comprehensive survey of deep learning in the field of medical imaging and medical natural language processing: Challenges and research directions. **Journal of King Saud University - Computer and Information Sciences**, v. 34, n. 8, Part A, p. 5083–5099, 2022. Nenhuma citação no texto.

PANG, T.; LI, P.; ZHAO, L. A survey on automatic generation of medical imaging reports based on deep learning. **BioMedical Engineering OnLine**, BioMed Central, v. 22, n. 1, p. 1–16, 2023. Nenhuma citação no texto.

PAPINENI, K.; ROUKOS, S.; WARD, T.; ZHU, W.-J. Bleu: a method for automatic evaluation of machine translation. In: **Proceedings of the 40th annual meeting of the Association for Computational Linguistics**. [S.l.: s.n.], 2002. p. 311–318. Nenhuma citação no texto.

RADFORD, A.; WU, J.; CHILD, R.; LUAN, D.; AMODEI, D.; SUTSKEVER, I. et al. Language models are unsupervised multitask learners. **OpenAI blog**, v. 1, n. 8, p. 9, 2019. Nenhuma citação no texto.

RAFFEL, C.; SHAZEER, N.; ROBERTS, A.; LEE, K.; NARANG, S.; MATENA, M.; ZHOU, Y.; LI, W.; LIU, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. **The Journal of Machine Learning Research**, JMLRORG, v. 21, n. 1, p. 5485–5551, 2020. Nenhuma citação no texto.

RANI, S.; JAIN, A.; KUMAR, A.; YANG, G. Cchexr-attention: Clinical concept extraction and chest x-ray reports classification using modified mogrifier and bidirectional lstm with multihead attention. **International Journal of Imaging Systems and Technology**, Wiley Online Library, v. 34, n. 1, p. e23025, 2024. Nenhuma citação no texto.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. **Psychological review**, American Psychological Association, v. 65, n. 6, p. 386, 1958. Nenhuma citação no texto.

SAHNI, N.; STEIN, G.; ZEMMEL, R.; CUTLER, D. M. **The Potential Impact of Artificial Intelligence on Healthcare Spending**. [S.l.], 2023. Nenhuma citação no texto.

SANH, V.; DEBUT, L.; CHAUMOND, J.; WOLF, T. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. **CoRR**, abs/1910.01108, 2019. Disponível em: <http://arxiv.org/abs/1910.01108>. Nenhuma citação no texto.

SANTORO, A.; FAULKNER, R.; RAPOSO, D.; RAE, J.; CHRZANOWSKI, M.; WEBER, T.; WIERSTRA, D.; VINYALS, O.; PASCANU, R.; LILLICRAP, T. Relational recurrent neural networks. **Advances in neural information processing systems**, v. 31, 2018. Nenhuma citação no texto.

SHARMA, R. R.; SUNGHEETHA, A.; TIWARI, M.; PINDOO, I. A.; ELLAPPAN, V.; PRADEEP, G. Comparative analysis of vision transformer and cnn architectures in medical image classification. In: ATLANTIS PRESS. **International Conference on Sustainability Innovation in Computing and Engineering (ICSICE 2024)**. [S.l.], 2025. p. 1343–1355. Nenhuma citação no texto.

SHORTEN, C.; KHOSHGOFTAAR, T. M. A survey on image data augmentation for deep learning. **Journal of big data**, SpringerOpen, v. 6, n. 1, p. 1–48, 2019. Nenhuma citação no texto.

SINGHAL, K.; TU, T.; GOTTWEIS, J.; SAYRES, R.; WULCZYN, E.; AMIN, M.; HOU, L.; CLARK, K.; PFOHL, S. R.; COLE-LEWIS, H. et al. Toward expert-level medical question answering with large language models. **Nature Medicine**, Nature Publishing Group US New York, v. 31, n. 3, p. 943–950, 2025. Nenhuma citação no texto.

SUN, C.; SHRIVASTAVA, A.; SINGH, S.; GUPTA, A. Revisiting unreasonable effectiveness of data in deep learning era. In: **Proceedings of the IEEE international conference on computer vision**. [S.l.: s.n.], 2017. p. 843–852. Nenhuma citação no texto.

SWATHI, Y.; BAI, A. S. K. Explainable deep learning for medical image captioning in chest x-ray analysis. In: **2025 International Conference on Knowledge Engineering and Communication Systems (ICKECS)**. [S.l.: s.n.], 2025. p. 1–7. Nenhuma citação no texto.

TANIDA, T.; MÜLLER, P.; KAISSIS, G.; RUECKERT, D. Interactive and explainable region-guided radiology report generation. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2023. p. 7433–7442. Nenhuma citação no texto.

TANNO, R.; BARRETT, D. G.; SELLERGREN, A.; GHAISAS, S.; DATHATHRI, S.; SEE, A.; WELBL, J.; SINGHAL, K.; AZIZI, S.; TU, T. et al. Consensus, dissensus and synergy between clinicians and specialist foundation models in radiology report generation. **arXiv preprint arXiv:2311.18260**, 2023. Nenhuma citação no texto.

TIAN, Y.; WANG, J.; YANG, W.; WANG, J.; QIAN, D. Deep multi-instance transfer learning for pneumothorax classification in chest x-ray images. **Medical Physics**, Wiley Online Library, v. 49, n. 1, p. 231–243, 2022. Nenhuma citação no texto.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention is all you need. **Advances in neural information processing systems**, v. 30, 2017. Nenhuma citação no texto.

VEDANTAM, R.; ZITNICK, C. L.; PARIKH, D. Cider: Consensus-based image description evaluation. In: **2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2015. p. 4566–4575. Nenhuma citação no texto.

WANG, H.; YE, S.; LIN, J.; NASEEM, U.; KIM, J. Lvmed-r2: Perception and reflection-driven complex reasoning for medical report generation. **arXiv preprint arXiv:2504.02885**, 2025. Nenhuma citação no texto.

WANG, J.; YANG, Z.; HU, X.; LI, L.; LIN, K.; GAN, Z.; LIU, Z.; LIU, C.; WANG, L. Git: A generative image-to-text transformer for vision and language. **arXiv preprint arXiv:2205.14100**, 2022. Nenhuma citação no texto.

WANG, R.; XU, Z.; WANG, X.; LIU, W.; LUKASIEWICZ, T. C2m-dot: Cross-modal consistent multi-view medical report generation with domain transfer network. **Information Fusion**, Elsevier, v. 125, p. 103442, 2026. Nenhuma citação no texto.

WANG, W.; BAO, H.; DONG, L.; BJORCK, J.; PENG, Z.; LIU, Q.; AGGARWAL, K.; MOHAMMED, O. K.; SINGHAL, S.; SOM, S. et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2023. p. 19175–19186. Nenhuma citação no texto.

WANG, X.; PENG, Y.; LU, L.; LU, Z.; BAGHERI, M.; SUMMERS, R. M. ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: **2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.]: IEEE, 2017. p. 2097–2106. Nenhuma citação no texto.

WANG, X.; PENG, Y.; LU, L.; LU, Z.; SUMMERS, R. M. **TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-rays**. [S.l.]: arXiv, 2018. Nenhuma citação no texto.

WEBSTER, J. J.; KIT, C. Tokenization as the initial phase in nlp. In: **Proceedings of the 14th Conference on Computational Linguistics - Volume 4**. USA: Association for Computational Linguistics, 1992. p. 1106–1110. Nenhuma citação no texto.

WÖDLINGER, M. G. **Applications of Neural Attention for Modelling Long-Range Dependencies**. Tese (Doutorado) — Technische Universität Wien, 2025. Nenhuma citação no texto.

XIONG, Y.; DU, B.; YAN, P. Reinforced transformer for medical image captioning. In: **Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings**. Berlin, Heidelberg: Springer-Verlag, 2019. p. 673–680. Nenhuma citação no texto.

YANG, M.; TANAKA, H.; ISHIDA, T. Performance improvement in multi-label thoracic abnormality classification of chest x-rays with noisy labels. **International Journal of Computer Assisted Radiology and Surgery**, Springer, v. 18, n. 1, p. 181–189, 2023. Nenhuma citação no texto.

YOU, D.; LIU, F.; GE, S.; XIE, X.; ZHANG, J.; WU, X. Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In: **Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24**. [S.l.: s.n.], 2021. p. 72–82. Nenhuma citação no texto.

YU, F.; ENDO, M.; KRISHNAN, R.; PAN, I.; TSAI, A.; REIS, E. P.; FONSECA, E. K. U. N.; LEE, H. M. H.; ABAD, Z. S. H.; NG, A. Y. et al. Evaluating progress in automatic chest x-ray radiology report generation. **Patterns**, Elsevier, v. 4, n. 9, 2023. Nenhuma citação no texto.

YUAN, J.; LIAO, H.; LUO, R.; LUO, J. **Automatic Radiology Report Generation based on Multi-view Image Fusion and Medical Concept Enrichment**. [S.l.]: arXiv, 2019. Nenhuma citação no texto.

ZAIDI, S. Z. Y.; AKRAM, M. U.; JAMEEL, A.; ALGHAMDI, N. S. A deep learning approach for the classification of tb from nih cxr dataset. **IET Image Processing**, Wiley Online Library, v. 16, n. 3, p. 787–796, 2022. Nenhuma citação no texto.

ZHAO, G.; ZHAO, Z.; GONG, W.; LI, F. Radiology report generation with medical knowledge and multilevel image-report alignment: A new method and its verification. **Artificial Intelligence in Medicine**, v. 146, p. 102714, 2023. Nenhuma citação no texto.

ZHOU, S. K.; RUECKERT, D.; FICHTINGER, G. **Handbook of medical image computing and computer assisted intervention**. [S.l.]: Academic Press, 2019. Nenhuma citação no texto.