Tutor Inteligente Multimodal para Feedback Formativo em Radiografias de Tórax usando Modelos Visão-Linguagem

Ivan Ferreira Martins, Mathias Cesar Assis
Nádia Félix Felipe da Silva, Sergio Teixeira de Carvalho,
Luciana de Oliveira Berretta

[1] Instituto de Informática (INF) – Universidade Federal de Goiás (UFG)

{ivan.martins, mathias.assis}@discente.ufg.br,

{nadia.felix, luciana.berretta, sergiocarvalho}@ufg.br

Abstract. *Radiology training requires the development of perceptual skills to locate abnormalities and cognitive skills to interpret medical images, traditionally acquired through direct supervision by experienced specialists. This work presents a multimodal intelligent tutor designed to provide automated formative feedback for chest radiograph interpretation. The system integrates the vision–language model Qwen2-VL-7B-Instruct, spatial evaluation using Intersection over Union (IoU), and natural language processing techniques to analyze radiological findings. The evaluation was conducted through simulated student interactions using images from the VinBigData Chest X-ray dataset, in which controlled perturbations of bounding boxes reproduce common diagnostic error patterns. Results indicate that the system can distinguish localization errors and semantic inconsistencies, demonstrating the computational feasibility of the proposed architecture as an educational support tool for radiology training.*

Resumo. *A formação em radiologia exige o desenvolvimento de habilidades perceptivas para localizar alterações e de habilidades cognitivas para interpretar imagens médicas, tradicionalmente adquiridas sob supervisão direta de especialistas. Este trabalho apresenta um tutor inteligente multimodal que fornece feedback formativo automatizado na interpretação de radiografias de tórax. O sistema integra o modelo de visão-linguagem Qwen2-VL-7B-Instruct, métricas espaciais baseadas em Intersection over Union (IoU) e técnicas de processamento de linguagem natural para a análise semântica dos achados radiológicos. A avaliação foi realizada por meio de interações simuladas com imagens do conjunto de dados VinBigData Chest X-ray, nas quais perturbações controladas nos bounding boxes reproduzem padrões comuns de erro diagnóstico. Os resultados indicam que o sistema consegue distinguir erros de localização e divergências semânticas, demonstrando a viabilidade computacional da arquitetura proposta como ferramenta de apoio ao treinamento em radiologia.*

1. Introdução

A formação de radiologistas exige o desenvolvimento simultâneo de competências perceptivas, relacionadas à localização de achados em imagens médicas, e de competências cognitivas, associadas à interpretação e à classificação diagnóstica. Tradicionalmente, essas habilidades são adquiridas sob supervisão direta de especialistas durante a análise de

exames. Essa prática é conhecida como *over-the-shoulder teaching* [Twidale 2005]. O aumento da demanda por exames radiológicos e a insuficiência de profissionais qualificados têm reduzido a escalabilidade desse modelo de treinamento. Estudos recentes indicam um crescimento significativo na demanda por radiologistas em diferentes sistemas de saúde [Meşe 2024, McKee 2024].

No processo inicial de aprendizagem, dois tipos de erro são frequentemente observados. Os *search errors* ocorrem quando o estudante não direciona adequadamente a atenção visual para a região patológica. Os *recognition errors* ocorrem quando a região correta é observada, mas a anomalia não é reconhecida ou interpretada corretamente [Nawaz 2024]. Esses erros indicam que a simples indicação de acerto ou erro não é suficiente para sustentar o processo de aprendizagem. Esse tipo de indicação não torna explícito o modelo mental adotado pelo aprendiz nem orienta a revisão de estratégias perceptivas e conceituais [Pellegrino et al. 2001].

Com os avanços recentes em modelos multimodais de visão e linguagem (*Vision–Language Models* – VLMs), tornou-se possível integrar informações visuais e textuais na análise de imagens médicas [Hong 2024, Wang 2024b]. Esses modelos permitem gerar descrições estruturadas de achados radiológicos e interpretar imagens em linguagem natural, superando as limitações das abordagens baseadas exclusivamente em texto. Paralelamente, a literatura em educação médica tem enfatizado a importância do *feedback* formativo, que fornece orientações detalhadas sobre o desempenho do aprendiz e contribui para o desenvolvimento progressivo de competências diagnósticas [Hartuique 2025, Weitekamp et al. 2020].

Apesar desses avanços, a literatura ainda apresenta uma lacuna na integração estruturada entre a avaliação espacial objetiva da localização de achados, a interpretação multimodal de imagens médicas e a geração automatizada de *feedback* pedagógico. Sistemas de *Computer-Aided Diagnosis* (CAD) priorizam a acurácia diagnóstica, enquanto sistemas tutores inteligentes frequentemente recorrem a regras estáticas ou a *feedback* textual limitado, sem explorar plenamente as capacidades de interpretação visual contextualizada oferecidas por modelos multimodais recentes.

Neste contexto, este trabalho propõe um tutor inteligente multimodal para fornecer *feedback* formativo automatizado na interpretação de radiografias de tórax. A abordagem integra interpretação visual baseada em um modelo visão-linguagem, avaliação espacial da localização de achados por meio da métrica *Intersection over Union* (IoU) e análise semântica com base em técnicas de processamento de linguagem natural. O sistema foi projetado para avaliar simultaneamente a precisão perceptiva na localização de lesões e a coerência conceitual das interpretações diagnósticas, permitindo identificar diferentes tipos de erro durante o processo de aprendizagem.

As principais contribuições deste trabalho são: (i) a proposição de uma arquitetura integrada que combina interpretação multimodal, avaliação espacial objetiva e análise semântica em um pipeline único orientado ao *feedback* formativo; (ii) a demonstração da viabilidade computacional dessa arquitetura em ambientes com recursos limitados por meio de técnicas de otimização e quantização; (iii) a validação técnica do sistema por meio de simulação controlada de interação discente, permitindo a caracterização de padrões de *search errors* e *recognition errors*; e (iv) o estabelecimento de uma base metodológica para estudos futuros envolvendo usuários reais e a ampliação do escopo para

diferentes patologias radiológicas.

Este artigo está organizado da seguinte forma. A Seção 2 apresenta os trabalhos relacionados. A Seção 3 descreve a metodologia adotada e o pipeline multimodal proposto. A Seção 4 apresenta os resultados das simulações realizadas. A Seção 5 discute os principais achados e as limitações do estudo. Por fim, a Seção 6 apresenta as conclusões e as diretrizes para trabalhos futuros.

2. Trabalhos Relacionados

A evolução dos Sistemas Tutores Inteligentes (STIs) demonstra avanços significativos na aprendizagem adaptativa, oferecendo instrução personalizada e retroalimentação contínua por meio da integração de modelos do aluno, de domínio e de estratégias instrucionais [Silva 2023, Li and Wilson 2025, Sonkar 2023]. No contexto da educação médica, estudos destacam que *feedback* detalhado e metodologias ativas em ambientes visuais estimulam a percepção, a interpretação e a autorregulação em radiologia [Meşe 2024, Hartuique 2025].

Paralelamente, o campo da IA aplicada à imagem médica avançou substancialmente, com o surgimento de sistemas de *Computer-Aided Diagnosis* (CAD), eficientes na detecção, mas raramente projetados com objetivos pedagógicos [Wang 2024b, Nawaz 2024]. A introdução de modelos multimodais de grande porte, como Qwen2-VL-7B-Instruct, expandiu as possibilidades ao integrar representação visual e linguagem natural, fornecendo descrições estruturadas de achados radiológicos [Bai 2023, Meşe 2024].

Embora existam avanços significativos em sistemas CAD, modelos multimodais de visão-linguagem e STIs, a literatura evidencia uma lacuna na integração estruturada desses componentes em um *pipeline* educacional unificado para radiologia. Sistemas CAD focam predominantemente na acurácia diagnóstica, enquanto os STIs tradicionais baseiam-se em *feedback* textual ou em regras predefinidas, sem explorar plenamente as capacidades de interpretação visual contextualizada oferecidas por modelos multimodais recentes. A lacuna identificada reside na ausência de arquiteturas que integrem simultaneamente: (a) avaliação espacial objetiva da precisão perceptual do aprendiz, (b) interpretação multimodal contextualizada de achados radiológicos, e (c) geração automatizada de *feedback* formativo adaptativo, baseada em análise semântica, em um fluxo coerente orientado pelos princípios pedagógicos de *scaffolding* e de autorregulação. Este trabalho contribui ao propor e validar tecnicamente uma arquitetura que preenche essa lacuna, articulando requisitos técnicos de processamento multimodal com princípios pedagógicos aplicados à formação em radiologia.

3. Metodologia

Este estudo adota a *Design Science Research* (DSR) como abordagem metodológica, uma vez que seu foco está no desenvolvimento, na implementação e na avaliação inicial de um artefato computacional concebido para o contexto educacional.

Segundo [Kroop 2025], a DSR orienta processos de construção e análise de soluções tecnológicas voltadas à resolução de problemas reais, fornecendo diretrizes para a validação técnica e a avaliação de relevância prática. No âmbito educacional, essa abordagem tem sido amplamente aplicada ao desenvolvimento de sistemas inteligentes, especialmente aqueles destinados a apoiar a aprendizagem [Freitas 2017, Pellegrino et al. 2001]

É fundamental esclarecer que o foco central deste estudo está na validação da viabilidade técnica e arquitetural do artefato proposto, seguindo as diretrizes metodológicas da *Design Science Research* para desenvolvimento de sistemas inovadores [Kroop 2025]. Nesta fase inicial da pesquisa, o objetivo primário consiste em: (a) demonstrar a exequibilidade computacional da integração entre modelos de visão-linguagem, métricas espaciais e análise semântica em um pipeline funcional; (b) validar a capacidade do sistema de processar entradas multimodais e de gerar, de forma automatizada, *feedback* estruturado; e (c) caracterizar o comportamento do tutor em diferentes cenários de precisão perceptual e de coerência conceitual. A simulação controlada da interação discente constitui uma estratégia metodológica deliberada e amplamente validada na literatura sobre Sistemas Tutores Inteligentes, permitindo a avaliação sistemática dos componentes arquiteturais antes da incorporação de usuários reais [Freitas 2017, Pimentel et al. 2020]. Esta abordagem possibilita o controle experimental rigoroso de variáveis (níveis de deslocamento espacial, tipos de erro perceptual), a identificação de limitações técnicas e o refinamento iterativo do artefato em condições controladas. Estudos subsequentes, fundamentados na arquitetura validada nesta pesquisa, poderão investigar o impacto pedagógico do sistema em contextos educacionais autênticos com estudantes de radiologia, incluindo análises de eficácia na aprendizagem, usabilidade, aceitação e efeitos na autorregulação e na tomada de decisão clínica.

O artefato investigado neste trabalho consiste em um tutor inteligente multimodal, formado pela integração de visão computacional, processamento de linguagem natural e métricas espaciais, para a análise do desempenho discente em tarefas de radiologia. A metodologia contempla as etapas de: (i) preparação e pré-processamento dos dados; (ii) simulação controlada da interação do estudante; (iii) interpretação multimodal por meio de um modelo de visão–linguagem; (iv) avaliação espacial das anotações; (v) análise semântica dos achados; (vi) geração de *feedback* formativo; e (vii) visualização e explicabilidade. A (Figura 1) ilustra o fluxo das etapas, desde a obtenção dos dados a serem utilizados como estratégia para evitar alucinações da VLM até a visualização do resultado pelo aluno, com o *feedback*.

3.1. Dados e pré-processamento

Foram utilizadas imagens de tórax do *dataset* ***VinBigData Chest X-ray Abnormalities Detection*** [Nguyen 2020], em formato PNG com resolução normalizada de 1024×1024 *pixels* [xhlulu 2021]. As coordenadas das *bounding boxes* foram convertidas da resolução original para 1024×1024 por meio de fatores de escala, resultando em um gabarito normalizado (**bbox_real**). Para a validação técnica do protótipo, foram selecionadas amostras da classe 8 (*Nodule/Mass*) com anotações completas, garantindo a consistência estrutural e a reprodutibilidade experimental.

3.2. Simulação da Interação do Aluno

Como o estudo foca na validação técnica e arquitetural do artefato, implementou-se uma simulação controlada da marcação discente (**bbox_aluno**), estratégia amplamente adotada na prototipação de STIs para validar componentes internos antes da incorporação de usuários reais [Weitekamp et al. 2020, Sonkar 2023]. A simulação gera *bounding boxes* artificiais a partir de deslocamentos sistemáticos e aleatórios em relação à *bbox_real* [Nguyen 2020], produzindo três cenários de validação detalhados na Tabela 1. A visualização sobrepõe à imagem original duas caixas: verde (**bbox_real**) e branca

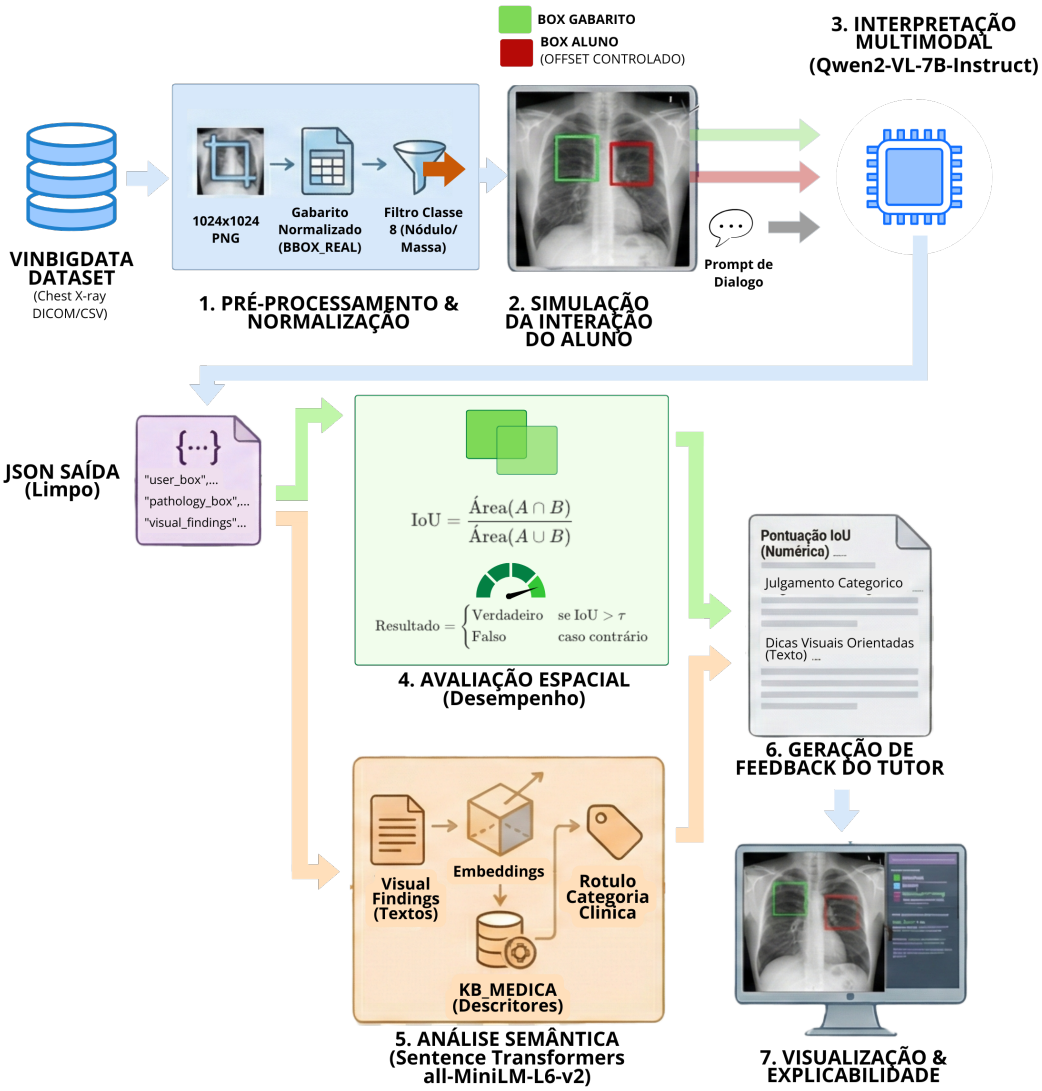

Figura 1. Arquitetura do *pipeline* proposto para o sistema de tutor inteligente multimodal. O fluxo ilustra o processamento desde a ingestão e a normalização dos dados do *dataset* VinBigData até a geração de *feedback* estruturado e a visualização explicável.

(**bbox_aluno**), permitindo avaliar o comportamento do tutor em diferentes níveis de precisão perceptual e de coerência conceitual [Castellani 2024].

3.3. Interpretação Multimodal com *Qwen2-VL-7B-Instruct*

O pipeline emprega o modelo visual-linguístico *Qwen2-VL-7B-Instruct* [Hong 2024], um VLM que integra a codificação visual e a geração de linguagem natural, para a interpretação semântica de imagens radiológicas. O modelo recebe exclusivamente a imagem original e um *prompt* estruturado, gerando uma descrição textual dos achados radiológicos observados (*visual_findings*). A inferência é configurada de forma determinística (*do_sample* = *False*) para assegurar a reprodutibilidade, e as saídas textuais são submetidas a limpeza e extração para estruturar os conteúdos relevantes [Meşe 2024].

A avaliação espacial ocorre de forma independente: as *bounding boxes* de referência (*bbox_real*) e as do aluno (**bbox_aluno**) são empregadas para calcular a métrica

Tabela 1. Tipologia dos cenários de simulação utilizados para validação das métricas espaciais e semânticas.

Tipo de Cenário	Descrição e Objetivo
Perturbações Espaciais Finas	Simulações com leve imprecisão (pequenos *offsets*). Destinam-se a avaliar a sensibilidade limítrofe do sistema e a precisão fina da métrica IoU na região de interesse.
Erros de Localização Grosseiros	Cenários de erro evidente com grandes deslocamentos. O objetivo é verificar a robustez do sistema na rejeição de falsos positivos e a qualidade do *feedback* pedagógico corretivo.
Divergência Semântica	Cenários que contêm erros na descrição textual do diagnóstico. Foram utilizados para testar a capacidade do modelo de linguagem de detectar inconsistências conceituais nos achados clínicos descritos pelo aluno.

Intersection over Union (IoU), sem participar do processo de interpretação visual. Assim, o módulo multimodal concentra-se na compreensão semântica, enquanto a geração do *feedback* educacional resulta da integração posterior entre a descrição textual e as métricas geométricas calculadas pelo sistema [Castellani 2024, Weitekamp et al. 2020].

3.4. Avaliação Espacial do Desempenho do Aluno

Para avaliar a precisão espacial da marcação discente, utilizou-se a métrica *Intersection over Union* (IoU), calculada pela Equação 1, com limiar de aceitação de 0,30 baseado em práticas de detecção de objetos [Jiang 2018, Nawaz 2024]:

$$\text{IoU} = \frac{\text{Área}(A \cap B)}{\text{Área}(A) + \text{Área}(B) - \text{Área}(A \cap B) + \epsilon} \tag{1}$$

$$\text{Score} = \begin{array}{ll} \text{Verdadeiro}, & \text{se IoU} > 0, \\ \text{Falso}, \text{caso contrário} & \end{array} \tag{2}$$

Onde $\epsilon = 10^{-6}$ é uma constante de estabilidade numérica. Essa métrica permite distinguir marcações incorretas de parcialmente corretas, fornecendo *feedback* gradual que favorece a autorregulação da aprendizagem [Pellegrino et al. 2001, Castellani 2024].

Embora o limiar de IoU de 0,30 seja inferior aos padrões rigorosos de competições de detecção de objetos, comumente definidos em 0,50, sua adoção justifica-se pelo caráter formativo do tutor. No contexto pedagógico inicial, identificar a região correta, mesmo com imprecisão nas bordas, é um marco de aprendizagem importante que merece feedback positivo, evitando a frustração do aluno por erros milimétricos.

3.5. Análise Semântica Baseada em *Sentence-Transformers*

A análise semântica interpreta, em nível conceitual, os achados visuais descritos pelo modelo multimodal, associando-os a categorias clínicas para geração de *feedback* formativo [Pellegrino et al. 2001]. Utilizou-se o modelo all-MiniLM-L6-v2 da biblioteca *Sentence-Transformers* [Reimers and Gurevych 2019] para converter a descrição textual (*visual_findings*) em um *embedding* semântico, comparado, por meio da similaridade de

cosseno, a uma base de conhecimento médica simplificada (KB_MEDICA) detalhada na Tabela 2. A categoria com maior similaridade é atribuída à instância, servindo de subsídio para a avaliação e para *feedback* educacional.

Tabela 2. Estrutura da Base de Conhecimento Médica (KB_MEDICA).

Categoria Clínica	IDs	Descritores e Palavras-Chave (Amostra)
Pneumonia/Consolidação	4, 6, 7, 11	pneumonia, consolidação, infiltrado, opacidade, *airspace disease*, mancha branca
Nódulo/Massa	2, 8, 9	nódulo, massa, tumor, câncer, lesão arredondada, *coin lesion*, calcificação
Patologia Pleural	10, 11, 12	derrame pleural, líquido, pneumotórax, ar na pleura, espessamento, *pleural effusion*
Cardiovascular	0, 3	cardiomegalia, coração grande, aorta alargada, mediastino, índice cardiotorácico
Intersticial/Fibrose	5, 13	fibrose, intersticial, faveolamento, vidro fosco, *honeycombing*, *interstitial*
Fratura/Osso	–*	fratura, quebrado, osso, costela, *fracture*, *bone*, *rib*
Normal/Sem Achados	14	normal, saudável, sem alterações, limpo, *no finding*, *clear*

* Categoria utilizada para diferenciar erros, sem ID específico correspondente no conjunto principal.

A classificação por similaridade de cosseno está alinhada a abordagens recentes que exploram representações vetoriais para interpretação em contextos educacionais e médicos [Li and Wilson 2025, Hong 2024].

3.6. Geração de *Feedback* do Tutor

O módulo de *feedback* integra informações de três componentes: (i) avaliação espacial baseada em IoU; (ii) interpretação semântica da imagem via *Qwen2-VL-7B-Instruct*; e (iii) análise semântica textual via *embeddings* de sentenças. Essas informações são consolidadas e processadas novamente pelo modelo visual-linguístico para gerar *feedback* formativo que inclui: valor numérico de IoU, indicação sobre a adequação da localização, categoria clínica predominante e orientação textual automatizada que destaca aspectos relevantes da imagem.

Ao integrar a avaliação espacial à interpretação textual, o *feedback* fornece uma visão integrada do desempenho, aproximando-se de práticas pedagógicas de especialistas e proporcionando *scaffolding* cognitivo ao longo do processo de aprendizagem [Pellegrino et al. 2001, Weitekamp et al. 2020, Castellani 2024].

3.7. Visualização e Explicabilidade

A etapa de visualização e explicabilidade tem como objetivo tornar transparentes o desempenho do estudante e os resultados das avaliações realizadas pelo sistema. Em contextos educacionais apoiados por Inteligência Artificial, especialmente em domínios complexos, como a interpretação de imagens médicas, a disponibilização de representações visuais

claras é fundamental para facilitar a compreensão do processo avaliativo e favorecer a aprendizagem [Weitekamp et al. 2020, Castellani 2024].

No protótipo desenvolvido, o módulo de visualização sobrepõe, sobre a imagem radiológica original, duas *bounding boxes*: a marcação de referência proveniente do *dataset* (**bbox_real**), exibida em cor distinta, e a marcação realizada pelo aluno, simulada no experimento (**bbox_aluno**). Essa sobreposição permite a comparação direta entre a localização correta da região de interesse e a marcação efetuada pelo estudante, evidenciando diferenças espaciais de forma intuitiva.

Além da visualização gráfica das marcações, o sistema apresenta informações textuais associadas à imagem, incluindo o valor da métrica IoU e a descrição dos achados visuais gerada pelo modelo multimodal *Qwen2-VL-7B-Instruct*. A combinação entre elementos visuais e textuais contribui para aumentar a interpretabilidade do processo avaliativo, alinhando-se às recomendações recentes da literatura sobre explicabilidade em sistemas baseados em IA [Hong 2024, Wang 2024b].

Este módulo permite que o estudante observe explicitamente como sua percepção espacial se relaciona à marcação de referência, favorecendo a reflexão sobre erros e acertos na localização das regiões de interesse. Ao mesmo tempo, a visualização auxilia professores e avaliadores na análise da coerência dos resultados gerados pelo sistema, contribuindo para a transparência e a confiabilidade do tutor inteligente.

4. Resultados

Os experimentos em ambiente controlado com o VinBigData Chest X-ray Abnormalities Detection permitiram avaliar o comportamento do tutor inteligente multimodal em diferentes cenários de marcação e de descrição de achados radiológicos, utilizando radiografias de tórax com anotações de radiologistas como referência [Nguyen 2020]. A simulação de respostas discentes, por meio da **bbox_aluno**, derivada de deslocamentos sistemáticos e aleatórios da **bbox_real**, permitiu reproduzir padrões de acerto, acerto parcial e erro grosseiro na localização de nódulos, opacidades e estruturas ósseas [Nawaz 2024]. A Figura 2 apresenta uma resposta ao teste em que o aluno acerta parcialmente a localização, mas erra a descrição da patologia.

O sistema adota uma ponderação hierárquica para o *feedback*; o sucesso na localização (*search*) é validado independentemente da precisão conceitual (*recognition*). No cenário da Figura 2, o tutor prioriza o reforço positivo do acerto espacial para, em seguida, aplicar o *scaffolding* corretivo à divergência semântica, garantindo que o aluno reconheça seu progresso perceptual antes de ser confrontado com o erro clínico.

O uso do *Qwen2-VL-7B-Instruct* para gerar descrições estruturadas, combinado ao modelo *all-MiniLM-L6-v2* para cálculo de similaridade de cosseno, permitiu mapear automaticamente os achados visuais para categorias clínicas gerais (Massa/Nódulo, Opacidade, Estruturas Ósseas e Normalidade) [Reimers and Gurevych 2019, Reimers 2020]. Esse comportamento aproxima-se do protótipo de propostas recentes que empregam grandes modelos de linguagem e de visão para a geração e o alinhamento de laudos radiológicos [Hong 2024, Li and Wilson 2025]. A Tabela 3 exemplifica o artefato final desse processamento, em que o objeto *JSON* estruturado pelo tutor é convertido em *feedback* educacional. Testes adicionais validaram a capacidade do sistema de distinguir cenários de acerto espacial com erro conceitual (IoU alto, mas categoria semântica

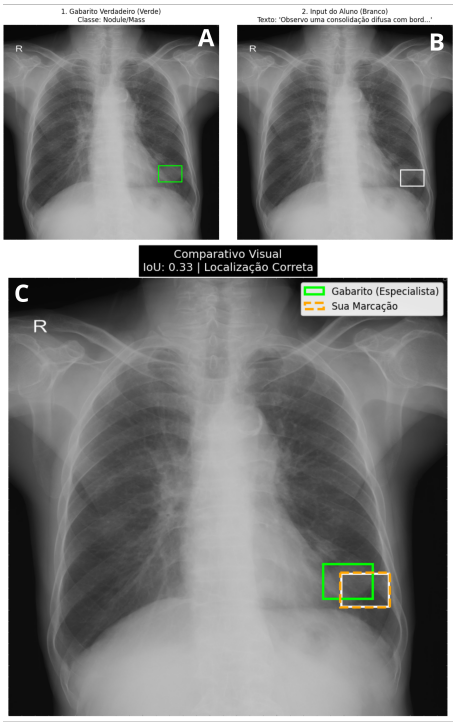

Figura 2. Cenário de acerto parcial: localização correta (IoU: 0.33, caixa laranja), mas erro semântico ao classificar "Nódulo/Massa" como "consolidação difusa".

Tabela 3. Comparativo entre os dados estruturados de inferência e o *feedback* pedagógico gerado.

Estrutura JSON (Sistema)	**Feedback do Tutor (Aluno)**
{ "result": { **"diagnosis_match"**: false, **"location_match"**: true, **"iou"**: 0.33 }, "details": { **"real_pathology"**: "Nódulo", **"student_guess"**: "Pneumonia", **"visual findings"**: "The lung pathology in the specified region appears to be a consolidation..." } }	***Avaliação do caso:*** Você identificou a localização da lesão corretamente, embora a área marcada não corresponda exatamente à do nódulo/massa. A lesão real está localizada no pulmão, mas a área marcada está mais próxima do coração. A localização da lesão foi correta, mas a interpretação foi incorreta. A lesão real é um nódulo/massa, e não uma consolidação pulmonar. Aconselho você a revisar a anatomia e a fisiologia pulmonares.

divergente) de cenários de erro perceptual com interpretação plausível em região incorreta (*search errors* vs. *recognition errors*), alinhando-se à taxonomia de erros diagnósticos descrita na literatura sobre detecção de anomalias torácicas [Nawaz 2024]. Em termos computacionais, a execução do VLM de médio porte, com técnicas de otimização de memória e quantização, demonstrou viabilidade em hardware com recursos limitados, confirmando resultados de estudos recentes sobre a aplicação de modelos visual-linguísticos otimizados em imagens médicas e em sistemas tutores inteligentes [Hong 2024, Wang 2024a, Lin et al. 2023].

5. Discussão

Os achados reforçam que a combinação entre avaliação espacial por IoU, interpretação multimodal e análise semântica permite explicitar o tipo de erro cometido pelo estudante, aproximando-se do protótipo das recomendações de *feedback* formativo estruturado na educação médica [Hartuique 2025]. Quando o sistema identifica um IoU baixo associado a uma hipótese diagnóstica incorreta, o *feedback* descreve o padrão radiológico esperado e a região correta, atuando como um mecanismo de *scaffolding* que orienta o ajuste tanto das estratégias de busca quanto das categorias conceituais utilizadas pelo discente [Silva 2023, Sonkar 2023, Weitekamp et al. 2020].

A diferença desta arquitetura não reside na criação de novos modelos, mas na orquestração inédita de VLMs e métricas geométricas para emular o ensino *over-the-shoulder*. Diferentemente de sistemas CAD puristas, o foco aqui é a transparência quanto aos erros. Contudo, reconhece-se que a natureza incremental da contribuição exige, em etapas futuras, uma análise de como essa automação impacta a curva de aprendizado real em comparação com o ensino tradicional.

Os resultados convergem com revisões que apontam o potencial de grandes modelos de linguagem e de visão para apoiar tarefas de laudo, sumarização e padronização de relatórios radiológicos [Garcia et al. 2024, Hong 2024, Xing et al. 2025]. Ao restringir o uso do VLM à descrição visual e delegar a validação espacial a métricas geométricas externas, o protótipo mitiga limitações relatadas em estudos que avaliam LLMs puramente textuais em tarefas radiológicas, como alucinações e baixa consistência na localização de achados [Hasani et al. 2024, Garcia et al. 2024].

As principais limitações deste estudo incluem: (i) a validação técnica restrita a uma única classe patológica (Nódulo/Massa), o que pode mascarar desafios de localização em patologias difusas; (ii) a dependência de simulações sintéticas que, embora validadas para testes de estresse da arquitetura, não capturam a variabilidade de comportamento e a linguagem natural de estudantes reais; e (iii) a necessidade de experimentos controlados com usuários para medir a eficácia pedagógica e a aceitação da ferramenta.

Ainda assim, a presente investigação permanece em nível de prova de conceito, com interação discente simulada e escopo restrito a classes, o que limita inferências sobre o impacto real na aprendizagem e a generalização para cenários clínicos mais complexos [Hevner et al. 2004, Lin et al. 2023]. Estudos futuros com estudantes e residentes, que incorporem medidas de autorregulação, engajamento com o *feedback* e desempenho em avaliações práticas, são necessários para validar o impacto pedagógico do sistema [Hartuique 2025, Lin et al. 2023]. Também é necessário investigar como estratégias como o *feedback* multimodal, os ciclos *feed-up*/*feedback*/*feed-forward* e a participação ativa do aprendiz podem ser implementadas no próprio tutor, seguindo recomendações para o desenho de sistemas de avaliação formativa centrados no estudante [Li and Wilson 2025, Hartuique 2025]. Os resultados indicam que a integração de um tutor inteligente multimodal ao *e-learning* em radiologia acompanha tendências mais amplas da IA na educação médica, ao combinar ambientes visuais ricos e agentes conversacionais para personalização do estudo [Meşe 2024, Lin et al. 2023].

6. Conclusão

Este estudo demonstra a viabilidade de um tutor inteligente multimodal para apoio à formação em radiologia, apresentando contribuições em três dimensões complementa-

res. Do ponto de vista técnico, a pesquisa propõe e valida uma arquitetura que integra modelos de visão-linguagem de médio porte (*Qwen2-VL-7B-Instruct*), métricas espaciais objetivas (IoU) e análise semântica automatizada (*sentence embeddings*) em um pipeline unificado. Os resultados indicam a viabilidade computacional da abordagem em ambientes com recursos limitados, por meio do uso de técnicas de quantização e de otimização de memória.

O trabalho também contribui ao propor e implementar uma estratégia de simulação controlada de interação discente que permite a validação sistemática de componentes arquiteturais em condições experimentais controladas. Essa estratégia permite caracterizar padrões de erro perceptual (*search errors*) e conceitual (*recognition errors*) antes da incorporação de usuários reais. A abordagem está alinhada às boas práticas de desenvolvimento de Sistemas Tutores Inteligentes.

Outra contribuição refere-se à dimensão educacional do sistema, que demonstra capacidade de gerar *feedback* formativo, explicativo e adaptativo. Esse mecanismo torna visível o modelo mental do aprendiz ao articular a avaliação espacial da precisão perceptual à análise semântica da coerência conceitual. Dessa forma, aproxima-se do modelo de supervisão clínica *over-the-shoulder* e alinha-se aos princípios pedagógicos de *scaffolding* cognitivo e de autorregulação da aprendizagem.

Com um escopo restrito, as limitações são reconhecidas. Restrição a uma classe patológica e a ausência de validação com usuários reais não comprometem a validade das contribuições apresentadas, mas delimitam claramente o estágio atual da pesquisa como uma prova de conceito técnico-arquitetural. Trabalhos futuros deverão expandir o escopo para múltiplas patologias radiológicas e conduzir estudos empíricos com estudantes de radiologia, investigando sistematicamente o impacto do *feedback* multimodal no desempenho diagnóstico, na autorregulação da aprendizagem e na tomada de decisão clínica em ambientes educacionais autênticos.

Referências

Bai, J. e. a. (2023). Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Castellani, A. M. e. a. (2024). Uso de inteligência artificial em sistemas de tutores inteligentes. *Revista de Ensino, Educação e Ciências Humanas*, 24(4):507–512.

Freitas, L. G. C. e. a. (2017). Design science research methodology enquanto estratégia metodológica para a pesquisa tecnológica. *Revista Espaços*, 38(6):7–20.

Garcia, B. T., Westerfield, L., Yelemali, P., Gogate, N., Rivera-Munoz, E. A., Du, H., Dawood, M., Jolly, A., Lupski, J. R., and Posey, J. E. (2024). Improving automated deep phenotyping through large language models using retrieval augmented generation. Repository: Genetic and Genomic Medicine.

Hartuique, H. C. O. C. e. a. (2025). A influência do feedback formativo no desenvolvimento da autorregulação da aprendizagem na formação médica. *Saúde Coletiva (Barueri)*, 15(94):15399–15424.

Hasani, A. M., Singh, S., Zahergivar, A., Ryan, B., Nethala, D., Bravomontenegro, G., Mendhiratta, N., Ball, M., Farhadi, F., and Malayeri, A. (2024). Evaluating the performance of generative pre-trained transformer-4 (GPT-4) in standardizing radiology reports. 34(6):3566–3574.

Hevner, A. R., March, S. T., Park, J., and Ram, S. (2004). Design science in information systems research1. *Management Information Systems Quarterly*, 28(1):75–106.

Hong, W. e. a. (2024). Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*.

Jiang, B. e. a. (2018). Acquisition of localization confidence for accurate object detection. In *Computer Vision – ECCV 2018*, pages 816–832, Cham. Springer.

Kroop, S. (2025). Artifact validity in design science research (dsr): A comparative analysis of three influential frameworks. In *Design Science Research*. Springer-Verlag, Berlin.

Li, M. and Wilson, J. (2025). Ai-integrated scaffolding to enhance agency and creativity in education: A systematic review. *Information*, 16(7):519.

Lin, C.-C., Huang, A. Y. Q., and Lu, O. H. T. (2023). Artificial intelligence in intelligent tutoring systems toward sustainable education: a systematic review. 10(1):41.

McKee, J. (2024). Workforce trends in radiologic technology. American Society of Radiologic Technologists (ASRT).

Meşe, e. a. (2024). Educating the next generation of radiologists: A comparative report of chatgpt and e-learning resources. *Diagnostic and Interventional Radiology*, 30(3):163–174.

Nawaz, U. e. a. (2024). Classification of thoracic abnormalities from chest x-ray images with deep learning. *International Journal of Advanced Computer Science and Applications*, 15(4).

Nguyen, H. Q. e. a. (2020). Vindr-cxr: A large-scale benchmark dataset for computer-aided diagnosis in chest radiography. *Scientific Data*.

Pellegrino, J. W., Chudowsky, N., and Glaser, R. (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. National Academy Press, Washington, DC.

Pimentel, M., Filippo, D., and Santoro, F. M. (2020). Design science research: fazendo pesquisas científicas rigorosas atreladas ao desenvolvimento de artefatos computacionais projetados para a educação. *Informática na Educação: Teoria & Prática*.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*.

Reimers, N. e. a. (2020). Sentence-transformers: Multilingual sentence embeddings using bert. *arXiv preprint*.

Silva, C. S. e. a. (2023). Sistemas tutores inteligentes na aprendizagem por competências: Uma revisão sistemática da literatura. In *SBIE*, Porto Alegre. SBC.

Sonkar, S. e. a. (2023). Class: A design framework for building intelligent tutoring systems based on learning science principles. In *Findings of ACL: EMNLP 2023*, Singapore.

Twidale, M. B. (2005). Over the shoulder learning: Supporting brief informal learning. *Computer Supported Cooperative Work (CSCW)*, 14(6):505–547.

Wang, S. e. a. (2024a). Interactive computer-aided diagnosis on medical image using large language models. *Communications Engineering*, 3:133.

Wang, W. e. a. (2024b). Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Weitekamp, D., Harpstead, E., and Koedinger, K. R. (2020). An interaction design for machine teaching to develop ai tutors. In *Proceedings of the CHI Conference*. ACM.

xhlulu (2021). Vinbigdata: Process and resize to png (1024x1024). Kaggle. Acesso em: 10 out. 2025.

Xing, Q., Song, Z., Zhang, Y., Feng, N., Yu, J., and Yang, W. (2025). Mca-rg: Enhancing llms with medical concept alignment for radiology report generation.