Reunião de Alinhamento
Introdução

Alysson Machado de Oliveira Barbosa

alysson.barbosa@ee.ufcg.edu.br

Universidade Federal de Campina Grande – UFCG
Programa de Pós-Graduação em Engenharia Elétrica – PPgEE

Orientadores:

Prof. Dr. Wamberto J. L. de Queiroz

wamberto@dee.ufcg.edu.br

Profa. Dra. Luciana Ribeiro Veloso

luciana.veloso@dee.ufcg.edu.br

24 de setembro de 2026

Sumário

1. Aprendizado Multimodal
2. Proposta de Investigação
3. Revisão Bibliográfica e Bases de Dados
4. Conclusão

Aprendizado Multimodal

O aprendizado multimodal busca modelar e relacionar informações provenientes de diferentes modalidades, explorando suas relações e complementaridades.

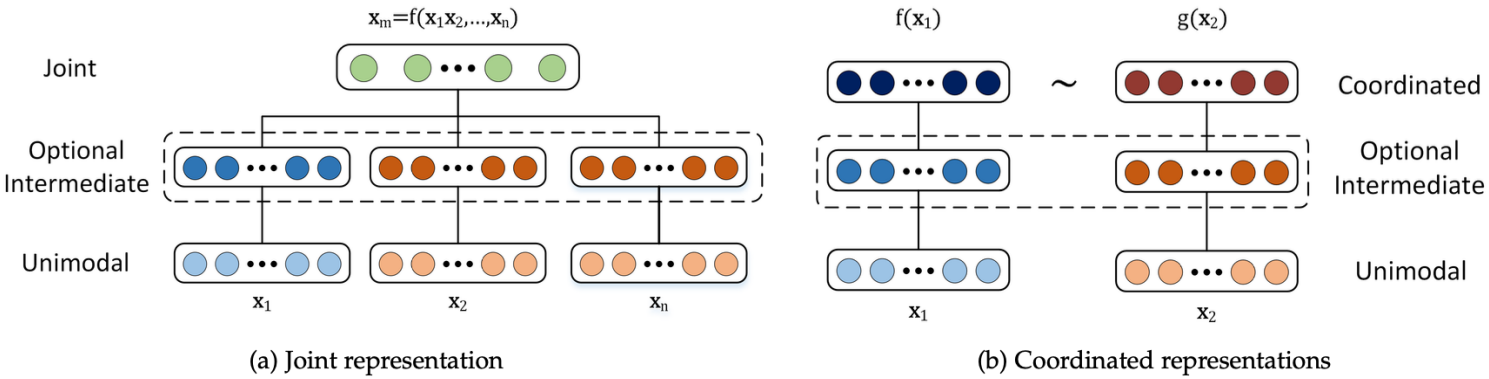


Figura 1: Estratégias de representação em aprendizado multimodal. Fonte: Baltrušaitis et al. (2019, fig. 1).

Duas perspectivas

- **Representação conjunta:** as modalidades contribuem para um mesmo espaço de representação.
- **Representações coordenadas:** espaços distintos são aproximados por relações de similaridade.

Fusão Multimodal em *Deep Learning*

Diferentes modalidades podem ser processadas por redes especializadas e posteriormente integradas em uma representação multimodal.

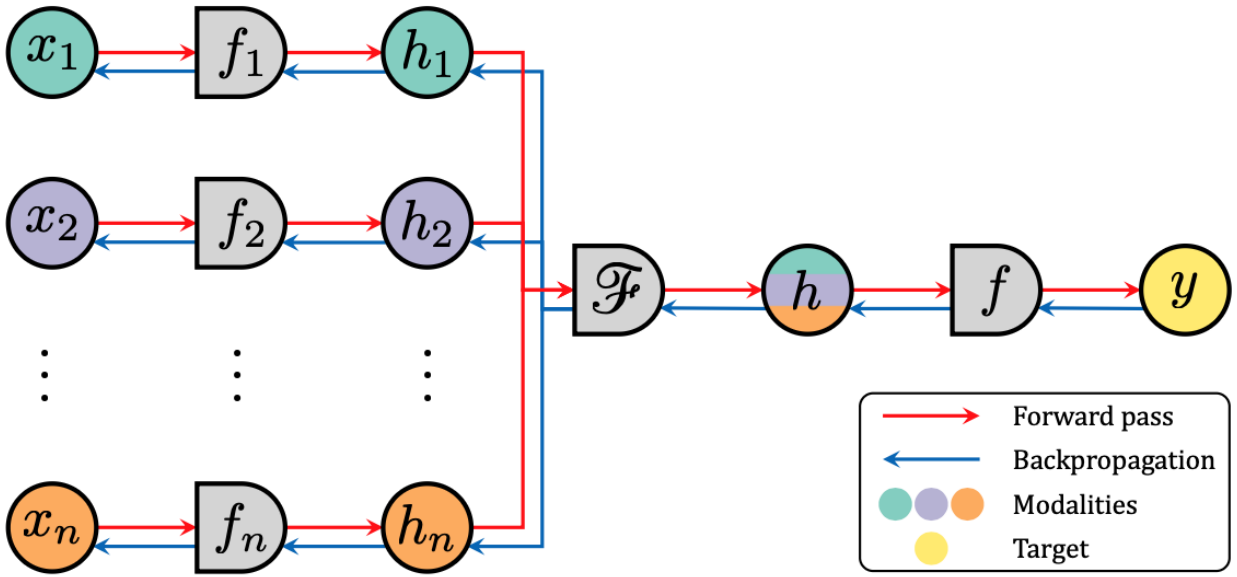
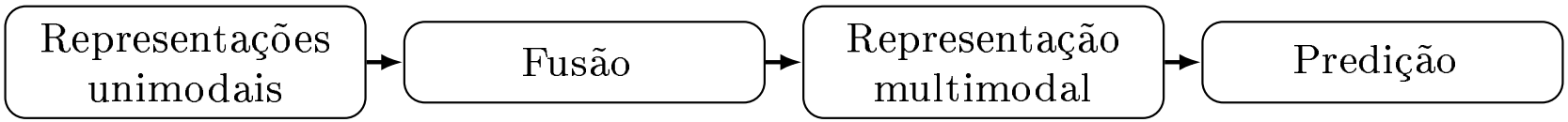


Figura 2: Fusão intermediária em aprendizado multimodal.
Fonte: Guarrasi et al. (2025, fig. 3).



Alinhamento Visual–Textual

Integrar duas modalidades não garante que suas representações sejam semanticamente correspondentes.

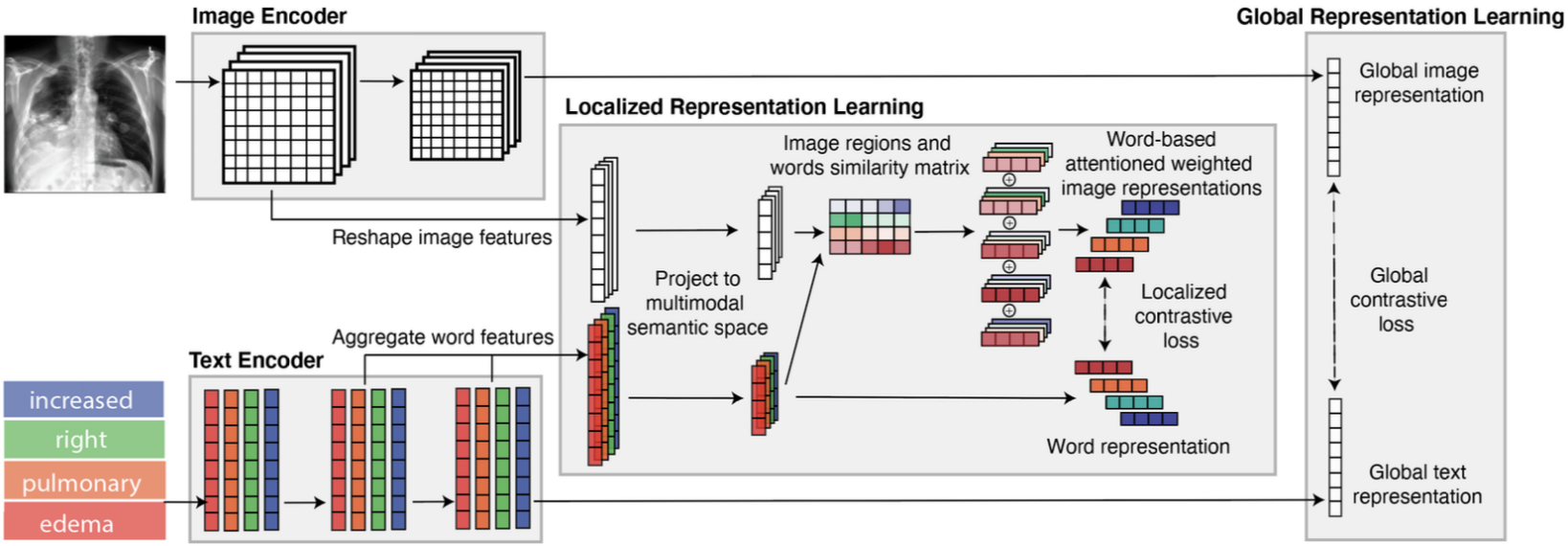


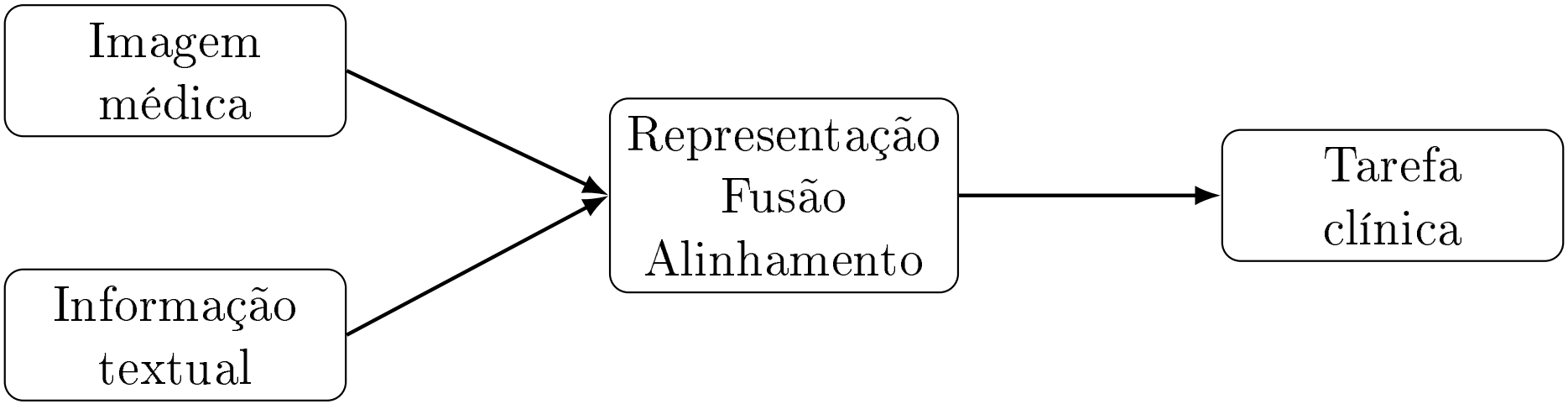
Figura 3: Aprendizado de representações globais e locais no GLoRIA. Fonte: Huang et al. (2021, fig. 2).

Alinhamento global
Imagem ↔ relatório

Alinhamento local
Região ↔ conceito

Problema a ser Pesquisado

- Imagens médicas e informações textuais fornecem informações complementares sobre uma mesma condição clínica.
- Disponibilizar diferentes modalidades a um modelo não garante que suas informações sejam adequadamente integradas.



Diferentes estratégias de integração podem produzir representações multimodais distintas, afetando a correspondência entre as informações visuais e textuais.

Pergunta da Pesquisa

Pergunta principal

Como diferentes estratégias de fusão e alinhamento entre imagem e texto influenciam o desempenho e a coerência multimodal de modelos de aprendizado profundo aplicados à análise de imagens médicas?

Questões investigativas iniciais

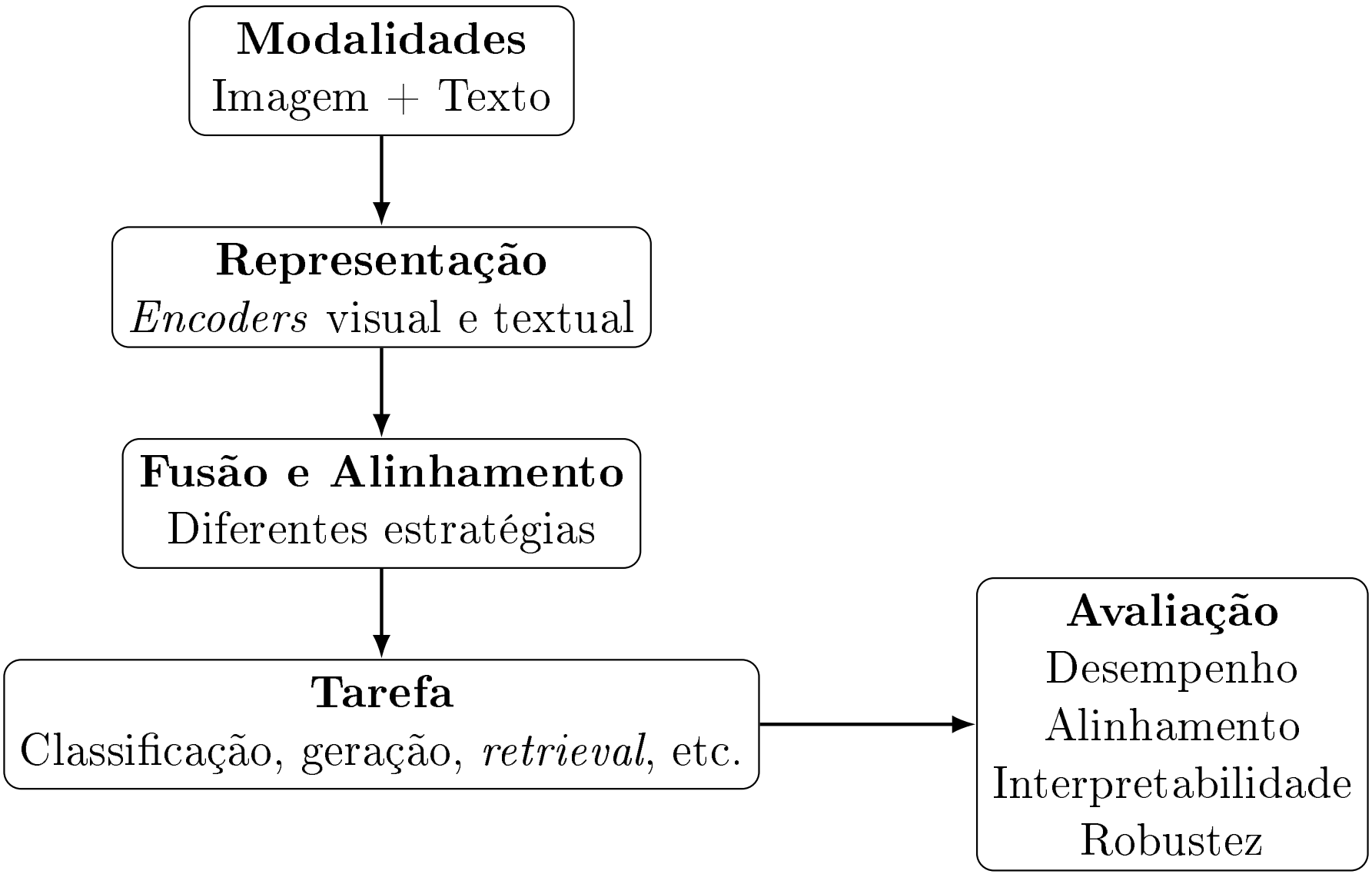
- Como diferentes pontos e mecanismos de fusão afetam o desempenho do modelo?
- Estratégias explícitas de alinhamento multimodal favorecem uma maior correspondência entre evidências visuais e informações textuais?
- Os benefícios da multimodalidade permanecem quando comparados, de forma controlada, com modelos unimodais e diferentes configurações de integração?

Pergunta da Pesquisa

Hipótese inicial

O desempenho final, isoladamente, pode não ser suficiente para caracterizar a qualidade da integração entre as modalidades.

Proposta Inicial de Investigação



Proposta Inicial de Investigação

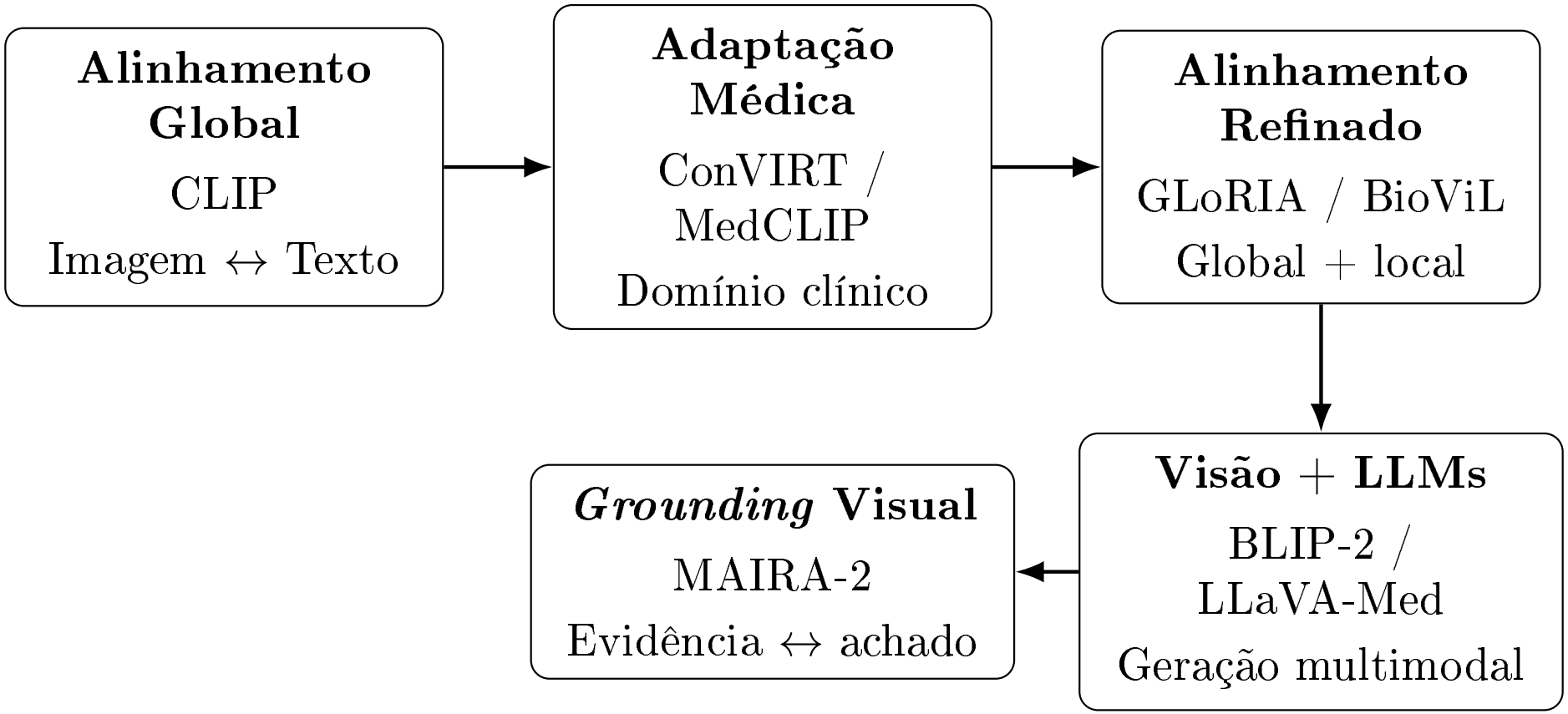
Contribuição pretendida

Investigar quais ganhos estão associados à representação, à fusão e ao alinhamento das modalidades, em vez de avaliar exclusivamente o desempenho global de diferentes modelos.

Revisão Bibliográfica

Datas	Artigos	Título	Motivo da Seleção
08/10/26	Artigo 1	Multimodal Machine Learning: A Survey and Taxonomy	Base conceitual inicial, define vários tópicos sobre o tema que ajudará com o entendimento dos próximos artigos.
08/10/26	Artigo 2	A Systematic Review of Intermediate Fusion in Multimodal Deep Learning for Biomedical Applications	Destaca a questão da fusão intermediária, arquiteturas de integração, pontos de fusão e formalização das estratégias biomédicas.
08/10/26	Artigo 3	Vision-language foundation models for medical imaging: a review of current practices and innovations	Atualização do panorama médico antes de entrar nos modelos individuais. É uma revisão de 2025 focada especificamente em VLMs para imagens médicas.
22/10/26	Artigo 4	Learning Transferable Visual models From Natural Language Supervision	Traz ideias sobre *dual encoders* , espaço latente compartilhado, similaridade entre imagem e texto, *contrastive learning* e *zero-shot* .
22/10/26	Artigo 5	Contrastive Learning of Medical Visual Representations from Paired Images and Text	Adaptação da aprendizagem contrastiva de imagem-texto no domínio médico. Usa os próprios laudos como supervisão para representações visuais.
05/11/26	Artigo 6	GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-efficient Medical Image Recognition	Aborda a questão do alinhamento não só de imagem-texto, mas também de regiões da imagem e palavras.
05/11/26	Artigo 7	Making the Most of Text Semantics to Improve Biomedical Vision-Language Processing	Mostra que não basta apenas alinhas *embeddings* , é importante adequar a semântica clinica do texto. Introduz representação radiológica especializada e alinhamento multimodal.
05/11/26	Artigo 8	MedCLIP: Contrastive Learning from Unpaired Medical Images and Text	Apresenta limitações do CLIP no domínio médico, com falsos negativos, escassez de pares e uso de conhecimento semântico para melhorar a função do alinhamento.
19/11/26	Artigo 9	BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models	Evolui o uso para modelo de visão + LLM generativo. Apresenta o Q-Former como mecanismo de ponte entre *encoder* visual e LLM congelado.
19/11/26	Artigo 10	LLaVa-Med: Training a Large Langugage-and-Vision Assistant for Biomedicine in One Day	Mostra como alinhamento e *instruction tuning* são levados para o domínio médico. O modelo primeiro aprender o vocabulário médico e depois instruções multimodais.
19/11/26	Artigo 11	MedGemma Techinical Report	Estado mais recente dos *foundation models* médicos. Serve como referência de arquitetura e capacidades contemporâneas.
03/12/26	Artigo 12	Integração de Contexto Clínico e Imagens de Raios-X do Tórax para Geração Automática de Laudos Radiológicos	Estudo com ResNet-50 + BioClinicalBERT + fusão multimodal + Transformer sobre MIMIC-CXR.
03/12/26	Artigo 13	MAIRA-2: Grounded Radiology Report Generation	Mostra onde a informação textual está sustentada na imagem, contribuindo para a questão da explicabilidade.
17/12/26	Artigo 14	Visual Alignment of Medical Vision-Language Models for Grounded Radiology Report Generation	O trabalho ataca explicitamente a deficiência de *cross-modal-alignment* e busca reduzir alucionações ligando achados linguísticos às evidências visuais.
17/12/26	Artigo 15	Multimodal Explainable Artificial Intelligence: A Comprehensive Review of Methodological Advances and Future Research Directior	Ajuda a construir a dimensão da interpretabilidade multimodal, apresentando a temático de um ponto de vista mais generalista.
17/12/26	Artigo 16	A Reality Check of Vision-Language Pre-training in Radiology: Have We Progressed Using Text?	O artigo questiona resultados de *vision-language pre-training* em radiologia, compara com pré-treinamento unimodal e aponta problemas em avaliações *zero-shot* .

Evolução das Abordagens na Literatura



Bases de Dados

Bases de Dados	Modalidades	Uso Inicial na Pesquisa	Motivos da Seleção
MIMIC-CXR	Radiografias de tórax e laudos radiológicos.	Experimentos de fusão e alinhamento entre imagem e texto.	Base pública diretamente alinhada ao problema multimodal entre imagem e texto e já considerada no plano de trabalho.
PadChest-GR	Radiografias de tórax e informação textual com associação a evidências/regiões da imagem.	Exploração de alinhamento mais localizado, *grounding* e avaliação complementar ao MIMIC-CXR.	Complementa a análise ao permitir investigar a correspondência entre achados textuais e evidências visuais.

Conclusão

- A multimodalidade possibilita explorar informações complementares entre imagem e texto, mas depende de estratégias adequadas de representação, fusão e alinhamento.
- A literatura apresenta diferentes formas de estabelecer interações e correspondências entre modalidades, desde representações globais até alinhamentos locais.
- A proposta inicial consiste em avaliar estratégias de fusão e alinhamento multimodal aplicadas à análise de imagens médicas.
- O aprofundamento da revisão bibliográfica subsidiará a definição das bases de dados, tarefas, arquiteturas e métricas utilizadas no protocolo experimental.

Referências

- Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(2), 423–443.
- Guarrasi, V., Aksu, F., Caruso, C. M., Di Feola, F., Rofena, A., Ruffini, F., & Soda, P. (2025). A systematic review of intermediate fusion in multimodal deep learning for biomedical applications. *Image and Vision Computing, 158*, 105509.
- Huang, S.-C., Shen, L., Lungren, M. P., & Yeung, S. (2021). GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 3942–3951).

Obrigado!

alysson.barbosa@ee.ufcg.edu.br



Universidade Federal
de Campina Grande