Metodologia para Geração de Datasets de Segmentação Multimodal a partir de Atributos Geométricos e Prompts Sintéticos

Alexandre Arantes Naves[1], **Ricardo Augusto Pereira Franco**[2]

[1]Escola de Engenharia Elétrica, Mecânica e de Computação
Universidade Federal de Goiás (UFG)
Goiânia – GO – Brazil

[2]Instituto de Informática
Universidade Federal de Goiás (UFG)
Goiânia – GO – Brazil

alexandrearantes32@gmail.com, ricardofranco@ufg.br

Abstract. *This paper presents a methodology to mitigate the scarcity of multimodal datasets, essential for advancing Vision-Language Segmentation Models. The proposal focuses on converting existing computer vision datasets into a multimodal format. Through an automated process, textual descriptions are generated from visual annotations. Crucial object attributes, including size, location in the image, are combined to create prompts. The ability to systematically produce this large-scale multimodal data from already annotated features contributes to significantly accelerating the research of semantic segmentation models that understand the interaction between vision and language.*

Resumo. *O artigo apresenta uma metodologia para mitigar a escassez de datasets multimodais, essenciais para o avanço dos Vision-Language Segmentation Models. A proposta centraliza-se na conversão de datasets de visão computacional já existentes para um formato multimodal. Através de um processo automatizado gera descrições textuais a partir das anotações visuais. Atributos cruciais do objeto, incluindo seu tamanho, localização na imagem, são combinados para a criação de prompts. A capacidade de produzir sistematicamente esses dados multimodais em larga escala a partir de recursos já anotados contribui para acelerar significativamente a pesquisa de modelos de segmentação semântica que compreendem a interação entre visão e linguagem.*

1. Introdução

A segmentação de imagens representa uma das tarefas mais críticas e desafiadoras no campo da visão computacional [Hao et al. 2020]. Sua aplicação é fundamental em uma vasta gama de cenários, desde a navegação autônoma de veículos, que depende da identificação precisa de obstáculos e vias, até inovações em aplicações domésticas e, de forma crucial, no diagnóstico médico por imagem. Por um longo período, as Redes Neurais Convolucionais (CNNs) [LeCun et al. 2015] foram consideradas como a arquitetura de estado da arte para essas tarefas, devido à sua excepcional capacidade de extrair hierarquias de características espaciais.

No entanto, nos últimos anos, o cenário tecnológico tem sido transformado pela ascensão dos modelos baseados em *Transformers* [Islam et al. 2024] . Originalmente desenvolvidos para o processamento de linguagem natural, sua aplicação à visão computacional, por meio dos *Vision Transformers* (ViT) [Dosovitskiy et al. 2021], emergiu como uma poderosa alternativa à abordagem convolucional. Mais recentemente, uma nova fronteira se abriu com as variações multimodais dessas arquiteturas [Radford et al. 2021], que são capazes de processar simultaneamente *inputs* de imagem e texto. Essa sinergia permite que o modelo seja guiado por *prompts* textuais, possibilitando o surgimento dos *Vision-Language Segmentation Models* (VLSM) [Rao et al. 2022]. Apesar do enorme potencial dessa abordagem, sua aplicação prática enfrenta um gargalo significativo: a escassez de *datasets* que contenham anotações multimodais, ou seja, imagens pareadas com descrições textuais detalhadas.

Para avaliar a eficácia da metodologia proposta por este trabalho, foi escolhido um cenário de alta relevância médica e social: a segmentação de pólipos colorretais. Os pólipos são precursores do câncer colorretal, sendo o segundo tipo de câncer que mais causa mortes em todo os EUA [Siegel et al. 2025] e estima-se que será o segundo câncer mais incidente no Brasil no ano de 2025 [de Oliveira Santos et al. 2023].

A detecção e segmentação precisas durante exames de colonoscopia são cruciais para a prevenção e o diagnóstico precoce da doença. A escolha desse cenário não se deve apenas à sua importância clínica, mas também às características intrínsecas dos pólipos, que apresentam uma grande variedade de tamanhos, formas e localizações no cólon. Essa diversidade oferece um ambiente de teste robusto e desafiador, ideal para uma avaliação rigorosa e detalhada da capacidade do método proposto em gerar descrições textuais úteis e precisas para a segmentação multimodal.

A Figura 1 apresenta exemplos contendo pólipos de tamanhos variados e posições diferentes nas imagens.

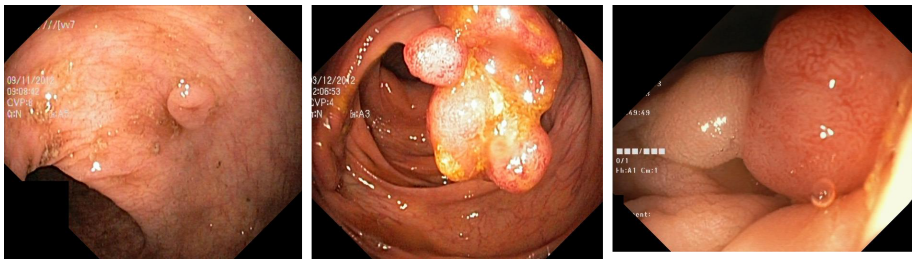

Figura 1. Exemplos de pólipos colorretais.

Portanto, as principais contribuições deste artigo são apresentadas abaixo:

1. Gerar anotações multimodais (imagem-texto) de forma automática a partir de anotações de máscaras de segmentação semântica;
2. Analisar o desempenho de modelos multimodais do estado da arte para a segmentação de pólipos.

Dessa forma, o objetivo principal deste artigo é propor uma metodologia para converter *datasets* de visão computacional já existentes, que possuem anotações de máscaras de segmentação e caixas delimitadoras, em um *dataset* com formato multimodal. A abordagem proposta utiliza atributos espaciais, como a localização e o tamanho relativo

dos objetos de interesse, para gerar automaticamente *prompts* pré-definidos estruturados e descritivos de alguns atributos que servirão como entrada de texto para modelos de segmentação multimodal.

Este trabalho está organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados à segmentação de imagens com modelos multimodais visão-linguagem; na Seção 3 é descrita a metodologia proposta para converter datasets de visão para o formato multimodal (visão-linguagem); os resultados da aplicação da metodologia são discutidos na Seção 4; a Seção 5 apresenta as principais conclusões deste trabalho.

2. Trabalhos relacionados

Dentre os trabalhos relacionados à segmentação semântica guiada por linguagem (*Vision-Language Semantic Segmentation*), destacam-se os modelos CLIPSeg [Lüddecke and Ecker 2022] e CRIS [Wang et al. 2022], que representam o estado da arte em arquiteturas do tipo *Vision-Language Segmentation Models* (VLSM). Esses modelos são capazes de segmentar objetos em imagens a partir de descrições textuais arbitrárias, uma capacidade conhecida como segmentação *zero-shot*. Essa abordagem elimina a necessidade de treinar ou refinar o modelo para cada nova classe de objeto. A aplicação dessas arquiteturas no domínio médico tem sido explorada em alguns trabalhos [Poudel et al. 2024], [Wang et al. 2025] e [Li et al. 2023], apresentando desempenho considerável em diversas tarefas como segmentação de manchas pulmonares em raios-x, análise de lesões de pele e úlceras de pés e de pólipos colorretais. Poundel et al. [Poudel et al. 2024] também propôs uma estrutura de prompts textuais para esse tipo de modelo a qual utiliza-se de trabalho manual e *Large Language Models* para a geração das descrições textuais.

Além da segmentação, o uso de *machine learning* para a classificação de patologias em imagens médicas é uma área de pesquisa consolidada, como demonstrado pelos autores em [Amin et al. 2022], no contexto de tumores cerebrais. Adicionalmente, a capacidade de diferenciar e classificar objetos com base em atributos físicos, como o tamanho, também foi abordada em pesquisas recentes [Li and Iskander 2022].

Contudo, a geração de datasets multimodais se apresenta como um desafio na atualidade. O custo financeiro e o tempo necessário para a realização da anotação de imagens no formato visual e textual são altos. Dessa forma, a metodologia proposta neste artigo tem como objetivo automatizar a geração da anotação textual para imagens, visando aprimorar as métricas para a tarefa de segmentação semântica.

3. Métodos

Nesta seção, será apresentada a metodologia proposta para a geração automática de *prompts* a partir de anotações de segmentação semântica.

A metodologia é composta pelas seguintes etapas: 1) seleção de *dataset*; 2) análise da área de interesse da máscara de segmentação; 3) definição do tamanho dos objetos utilizando algoritmos de agrupamento; 4) criação dos *prompts*; 5) treinamento de modelos de VLSM; 6) e avaliação da performance. A Figura 2 mostra o fluxograma das etapas definidas pela metodologia proposta neste trabalho.

As principais etapas da metodologia proposta serão descritas nas subseções a seguir.

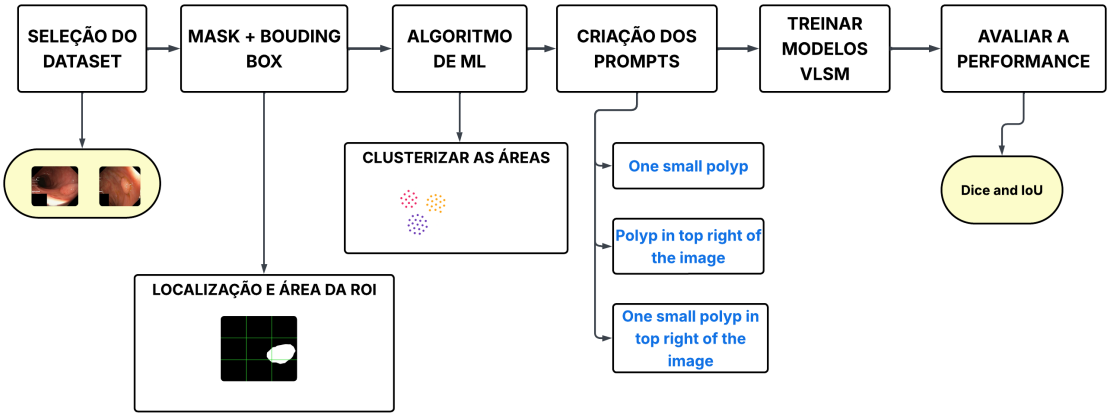

Figura 2. Fluxograma contendo as etapas da metodologia proposta.

3.1. Captura da localização do objeto

A primeira etapa é, basicamente, a definição do *dataset* a ser utilizado. Em seguida, executa-se a segunda etapa, a análise da área da máscara de segmentação para capturar a localização do objeto.

Esta etapa se inicia com a aplicação de uma grade de 3x3 sobre a imagem, dividindo-a em nove quadrantes iguais, conforme apresentado na Figura 3. Esta grade é utilizada com o objetivo de criar um sistema de coordenadas espaciais, permitindo a localização do objeto de interesse.

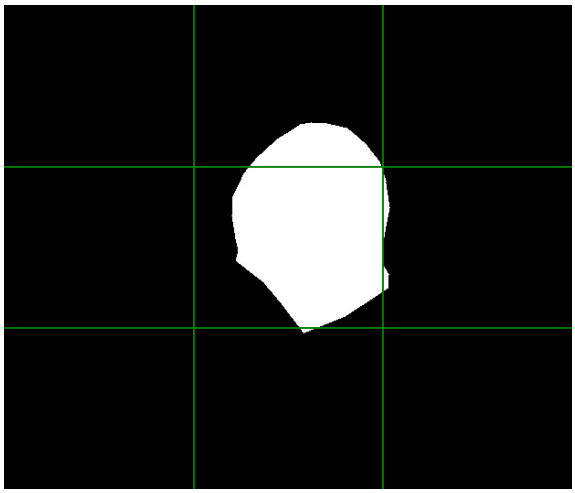

Figura 3. Divisão da imagem para o cálculo das áreas.

Para cada anotação de segmentação presente na imagem, será gerada uma caixa delimitadora (*bounding box*), com o objetivo de separar o objeto de interesse do fundo e outros objetos que possam existir. Isso é realizado definindo valor igual a zero para todos os *pixels* que se encontram externos à *bounding box* da respectiva anotação. Essa etapa garante que a análise subsequente se concentre exclusivamente em um objeto de interesse por vez, eliminando ruídos e informações contextuais irrelevantes.

Em seguida, realiza-se a análise da máscara de segmentação associada. Realiza-se o cálculo dos *pixels* da máscara em cada um dos nove quadrantes da imagem. O

quadrante que contiver a maior quantidade de *pixels* da máscara é identificado como a região de predominância do objeto na imagem. Esta informação de localização (e.g., ”canto superior esquerdo”, ”centro”, ”borda direita inferior”) é um dos elementos-chave para a composição dos *prompts* a serem gerados a posteriori.

3.2. Classificando o tamanho das áreas

Em paralelo à definição da localização do objeto, é calculada a área percentual que a máscara do objeto ocupa em relação à área total da imagem. Este cálculo fornece uma medida quantitativa da região do objeto dentro da imagem na qual está inserido.

Após o processamento do conjunto de imagens, as áreas percentuais calculadas para cada objeto são classificadas utilizando um algoritmo de *clusterização* não supervisionado. Este algoritmo classifica os dados em relação à área percentual em três categorias distintas, que foram consideradas como ”pequena”, ”média”e ”grande”proeminência do objeto. Cada objeto na imagem é, então, rotulado com sua localização predominante na grade e a categoria de área atribuída pelo algoritmo de *clusterização*. Foram utilizados para fim de avaliação os seguintes algoritmos: *K-Means* [Lloyd 1982], *Gaussian Mixture Models* [Reynolds 2015] e *Hierarchical cluster* [Sokal et al. 1958].

3.3. Geração dos *prompts*

A etapa final consiste na geração de *prompts* de texto. As informações extraídas são combinadas, sendo elas: a localização predominante do objeto em uma posição da grade e a classificação de tamanho da área relativa. Um sistema baseado em templates é então utilizado para gerar uma descrição coesa e informativa dos *prompts*.

Foram definidas as gerações de três *prompts* distintos que poderão ser adaptados e utilizados em função da aplicação específica:

1. *Prompts* com a combinação da localização com a descrição da anomalia;
2. *Prompts* com a combinação dos atributos de quantidade e tamanho com a descrição da anomalia;
3. *Prompts* com a combinação dos atributos de quantidade, tamanho e localização com a descrição da anomalia.

Essa abordagem sistemática permite a criação de descrições ricas e detalhadas, automatizando a análise e a interpretação de conteúdo visual de forma escalável e eficiente. É importante ressaltar que a definição dos *prompts* a serem gerados fica a critério do usuário, podendo criar, modificar ou remover os *prompts* definidos neste trabalho. Pode-se notar, também, que a geração de prompts pode ser feita em qualquer idioma, desde que os modelos de linguagem tenham sido treinados para aquele idioma específico.

3.4. Modelos treinados

Duas arquiteturas de VLSM foram selecionadas para analisar sua eficácia na metodologia proposta, sendo elas as arquiteturas CRIS [Lloyd 1982] e CLIPSeg [Lüddecke and Ecker 2022]. Ambas as arquiteturas foram treinadas com as três variações de *prompts* definidas na metodologia, juntamente com o nome da anomalia a ser segmentada.

1. : quantidade de pólipos + descrição de localização (ex.: one polyp in the right of the image);

2. : quantidade de pólipos descrição de tamanho + pólipo (ex.: one small polyp);
3. : quantidade de pólipos descrição de tamanho + pólipo seguido de localização (ex.: one small polyp in the right of the image);

3.5. Métricas de avaliação

Para avaliar os resultados das segmentações geradas pelos modelos, as métricas escolhidas foram: *Intersection over Union* (**IoU**) e **Dice**; ambas as métricas baseadas em sobreposição que quantificam a similaridade entre duas áreas.

A métrica **IoU** é calculada através da seguinte equação:

$$\text{IoU} = \frac{1}{N}\sum_{i=1}^{N}\frac{M_i \cap P_i}{M_i \cup P_i}, \tag{1}$$

onde M_i corresponde à máscara original da i-ésima imagem; P_i corresponde à máscara predita pelo modelo para a mesma imagem; e N corresponde ao número total de imagens analisadas.

O valor de **Dice** pode ser obtido por meio da seguinte equação:

$$\text{Dice}_i = \frac{2 \cdot |M_i \cap P_i|}{|M_i| + |P_i|}. \tag{2}$$

4. Resultados e discussão

Nesta seção serão discutidos os resultados obtidos por meio da aplicação da metodologia proposta para a tarefa de segmentação de pólipos em imagens de colonoscopia.

O *dataset* escolhido foi o Kvasir-SEG [Jha et al. 2020] que possui 1.000 imagens de pólipos. A divisão do conjunto de imagens em conjunto de treino, validação e teste foi realizada da seguinte forma: 800/100/100. O sistema computacional utilizado foi um notebook com processador i7-13650HX, 16 GB de memória RAM e uma GTX 4060 com 8GB VRAM para realizar o treinamento dos modelos de *machine learning*.

4.1. Classificação por *machine learning*

Os primeiros pontos a serem destacados são a distribuição da classificação de localização dos pólipos e de seus respectivos tamanhos nas imagens, para, posteriormente, gerar os prompts relacionados às imagens. Pode-se observar, na Tabela 1, a distribuição de tamanhos realizada pelos três algoritmos utilizados.

Tabela 1. Resultado da classificação do tamanho dos pólipos utilizando os algoritmos *K-means*, *Gaussian Mixture Model* (GMM) e *Hierarchical*

Classificação	*K-means*	GMM	*Hierarchical*
Small	677	670	880
Medium	311	334	168
Large	83	67	23

Nota-se que há uma tendência dos modelos em realizar a classificação dos pólipos como pequenos nos três casos, com uma acentuação maior feita pelo algoritmo *Hierarchical*. Este modelo apresentou um padrão de classificação que difere dos demais modelos (*K-means* e GMM), que tiveram quantidades mais próximas de imagens em cada classe, sendo a maior diferença observada na classe "média"com 23 imagens a mais classificadas pelo GMM.

A distribuição de localização, apresentada na Tabela 2, demonstra uma concentração de pólipos na faixa central da imagem, principalmente no centro da imagem. Isso indica que os especialistas que realizam o exame centralizam os pólipos ao serem detectados para depois realizarem a captura das imagens.

É importante ressaltar, também, que um fator que contribui para a baixa concentração de pólipos classificados na posição *bottom left* é que essa região normalmente é escolhida por alguns aparelhos colonoscópios para mostrar informações relevantes ao especialista durante o exame.

Tabela 2. Resultado da contagem dos pólipos por região obtida pela metodologia proposta

Localização	Quantidade
Center	474
Right	180
Bottom	141
Top	103
Left	73
Bottom right	45
Top right	33
Top left	17
Bottom left	5

4.2. Avaliação dos resultados

Após a definição do tamanho e da localização do pólipo, o próximo passo é realizar a geração da máscara dos pólipos.

Para gerar as máscaras utilizando os modelos CRIS e CLIPSeg, dois experimentos foram realizados para simular duas situações: 1) situação em que se desconhece toda a informação a respeito do pólipo na imagem; 2) situação na qual o especialista tem a informação sobre o pólipo e utiliza essa informação para gerar a máscara do pólipo. Isso implica, na prática, em gerar a máscara sem nenhum *prompt* ou com o auxílio do *prompt* igual ao utilizado para treinar os modelos CRIS e CLIPSeg.

Os resultados obtidos estão descritos na Tabela 3 e eles se referem apenas às imagens do conjunto de teste. Há também uma comparação com o melhor resultado atingindos por [Poudel et al. 2024]. Os resultados serão discutidos nas Seções 4.3 e 4.4. Na coluna [Poudel et al. 2024] é feito uma comparação dos resultados obtidos por este trabalho com o artigo [Poudel et al. 2024], apesar dos prompts definidos utilizarem mais descrições como forma e cor é possível fazer uma comparação entre os resultados, visto que segue-se uma estrutura de treinamento e validação semelhante.

Tabela 3. Comparação de métricas de segmentação para diferentes algoritmos e configurações.

	K-means		GMM		Hierarchical		[Poudel et al. 2024]	
Configuração	**Dice**	**IoU**	**Dice**	**IoU**	**Dice**	**IoU**	**Dice**	**IoU**
CRIS - sem prompt								
CRIS - P1	0,808	0,710	0,808	0,710	0,808	0,710	-	-
CRIS - P2	0,849	0,773	0,800	0,710	**0,856**	**0,778**	-	-
CRIS - P3	0,759	0,674	0,838	0,765	0,705	0,606	-	-
CRIS								
CRIS - P1	0,886	0,822	0,886	0,822	0,886	0,822	0,864*	-
CRIS - P2	0,911	0,851	**0,912**	**0,852**	0.894	0.829	0,843	-
CRIS - P3	0,909	0,847	0,907	0,848	0,896	0,837	0,854	-
CLIPSeg - sem prompt								
CLIPSeg - P1	**0,810**	**0,723**	**0,810**	**0,723**	**0,810**	**0,723**	-	-
CLIPSeg - P2	0,753	0,653	0,761	0,662	0,746	0,649	-	-
CLIPSeg - P3	0,760	0,675	0,776	0,691	0,798	0,705	-	-
CLIPSeg								
CLIPSeg - P1	0,857	0,781	0,857	0,781	0,857	0,781	0,890*	-
CLIPSeg - P2	**0,886**	**0,814**	0,884	0,811	0,879	0,806	0,883	-
CLIPSeg - P3	0,872	0,801	0,878	0,804	0,868	0,794	0,890	-

* - melhor métrica do artigo [Poudel et al. 2024]

4.3. Experimento 1: Performance dos modelos sem *prompt*

No primeiro experimento, após a realização dos treinamentos conforme a metodologia apresentada, os modelos CRIS e CLIPSeg foram avaliados no conjunto de teste sem a adição de informações textuais, para que esses modelos realizassem a tarefa de segmentação de pólipos.

É importante ressaltar que, durante o treinamento, o *prompt* P1 combina apenas atributos de localização, portanto, a *clusterização* por *machine learning* não alterou os resultados para esses casos. Assim, o valor obtido para esse *prompt* é igual para os três algoritmos apresentados na Tabela 3.

Na arquitetura CRIS, o melhor resultado obtido foi com o *prompt* P2 que utilizou apenas a descrição de tamanho da área do pólipo a ser segmentada e com a classificação feita pelo algoritmo *Hierarchical Clustering*.

Já nos resultados obtidos pelo CLIPSeg, observa-se desempenho inferior quando comparado com os resultados do modelo CRIS. Contudo, o mesmo padrão dos resultados com os prompts se repetiu, isto é, utilizar apenas a informação a respeito do tamanho do pólipo obteve melhores valores nas métricas. Tal situação demonstra que os modelos conseguiram obter mais características relevantes quando utilizadas apenas as informações relacionadas ao tamanho relativo do pólipo na imagem.

4.4. Experimento 2: Performance dos modelos com *prompt*

No Experimento 2, foi realizada a segmentação de imagens com os mesmos modelos do Experimento 1, porém, além da imagem a ser segmentada, também foi adicionada como

entrada do modelo a informação textual gerada em relação ao pólipo. Essas informações são de acordo com a *prompt* usado para treinamento, por exemplo, se o modelo foi treinado apenas com atributos de localização, então as informações textuais passadas são apenas de localização.

Espera-se, nesse caso, que a performance seja melhor em relação ao Experimento 1, visto que serão fornecidas informações pelo (*prompt*) juntamente com a imagem como entrada para o modelo de segmentação.

Observa-se na Tabela 3 que o modelo CLIPSeg apresentou melhor perfomance geral nos resultados apresentados no artigo [Poudel et al. 2024]. Contudo, nos experimentos realizados neste trabalho, essa característica não se repetiu. O modelo CRIS apresentou desempenho melhor quando comparado com o modelo CLIPSeg.

Pode-se observar, também, que o resultado obtido pelo modelo CRIS utilizando o *prompt* P2 juntamente com o algoritmo GMM, foi o melhor resultado dentre todos os experimentos realizados com 0,912 de Dice e 0,852 de IoU. Este resultado demonstra que ao adicionar o *prompt* P2 (descrição de tamanho + pólipo) como entrada do modelo juntamente com a imagem, obtém-se melhor resultado na tarefa de segmentação de pólipo.

A Figura 4 apresenta uma comparação visual das máscaras geradas pelos Experimentos 1 e 2 e com a máscara anotada por um especialista. Nota-se que as máscaras geradas quando fornecido a *prompt* como entrada apresentam melhores resultados quando comparadas com a máscara original.

4.5. Análise dos Resultados

Os resultados apresentados na Tabela 3 indicam que pode-se considerar a informação de 'tamanho' do pólipo sendo mais relevante em relação a informação de 'localização' do pólipo. Esse fato ocorre devido a quase todos os resultados obtidos para as métricas de Dice e IoU terem sido maiores para os prompts P2 e P3, exceto pelos resultados com CLIPSeg sem *prompt*.

Ao realizar a clusterização das imagens em relação ao tamanho, observa-se que os resultados das métricas ao utilizar os algoritmos *k-means* e GMM ficaram bastante próximos. Esse fato se justifica ao analisar a divisão das imagens em *small*, *medium* e *large* por esses algoritmos. Nota-se, pequenas alterações entre as classes, isto é, houve uma variação de apenas 7 imagens classificadas como classe *small*, 23 imagens para a classe *medium* e 16 imagens para a classe *large*. Assim, os resultados obtidos utilizando *k-means* e GMM foram próximos, com vantagem para o algoritmo GMM.

Já em relação aos resultados dos modelos segmentadores, observa-se vantagem ao realizar o teste com as imagens e seus respectivos *prompts* gerados. Em todos os casos, os resultados obtido com a adição de *prompts* resultaram em métricas superiores àquelas obtidas pela realização da segmentação sem a inclusão do *prompt*.

Nota-se, também, que os resultados obtidos pela metodologia de geração de *prompt* automática obteve métricas maiores que o trabalho [Poudel et al. 2024] em 66,7% dos casos. Analisando os resultados com o modelo CRIS com *prompt* gerado, todos os valores para as métricas de Dice e IoU foram superiores às métricas obtidas em [Poudel et al. 2024]. Dentre todos os modelos avaliados, o modelo CRIS com o formato de *prompt* P2 é o recomendável para uso, visto que ele obteve as melhores métricas com

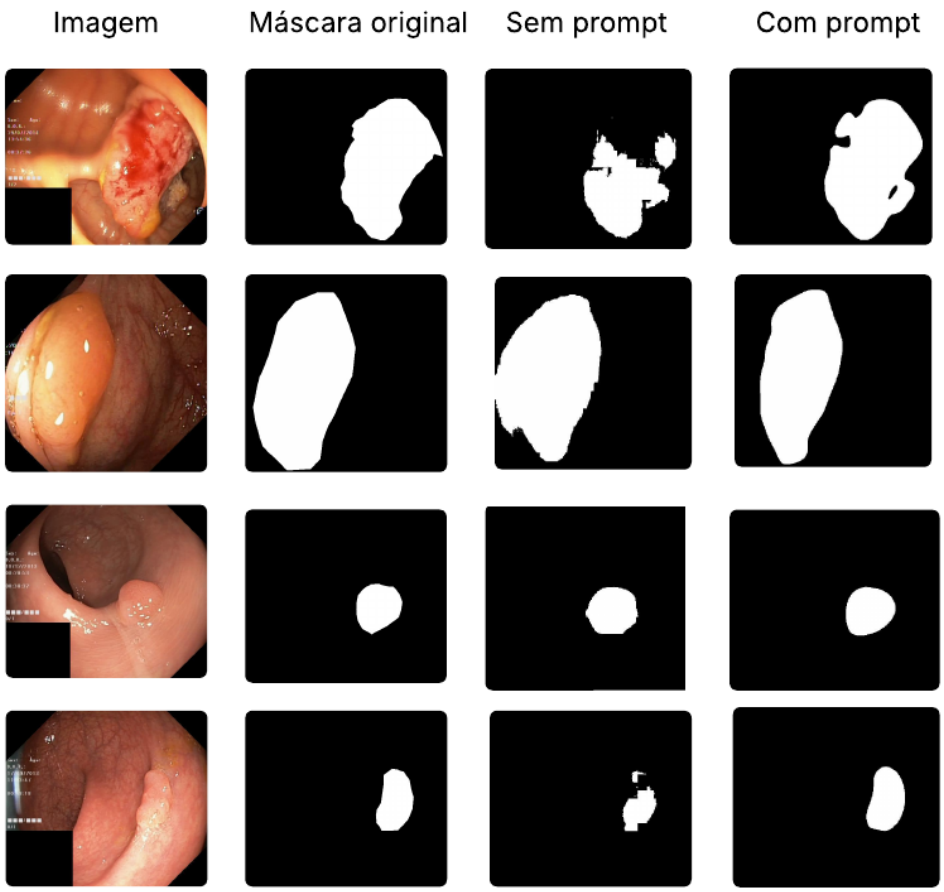

Figura 4. Comparativo das máscaras original (anotada por especialista) e as máscaras geradas com *prompt* e sem *prompt*

um valor de 0,912 de Dice e 0,852 de IoU.

Por fim, é importante observar que os *prompts* em [Poudel et al. 2024] foram gerados por especialistas, enquanto a proposta deste artigo é realizar a geração automática de *prompts*. Portanto, mesmo os modelos utilizando *prompts* gerados automaticamente obtiveram resultados superiores a prompts feitos por especialistas. Este fato demonstra a qualidade dos *prompts* gerados pela proposta deste artigo e que a geração automática pode remover vieses da anotação realizada por diferentes especialistas.

5. Conclusão

A avaliação da metodologia revela uma dualidade nos resultados: por um lado, ela é capaz de atingir um desempenho de alta qualidade, mas por outro, exibe uma considerável variabilidade. Essa sensibilidade ao contexto também se reflete na escolha do algoritmo de clusterização, pois todos os três foram capazes de alcançar a performance máxima sob diferentes combinações de arquitetura e *prompts*.

Apesar disso, o *Gaussian Mixture Models* (GMM) emergiu como a abordagem

mais robusta, apresentando a maior consistência e os resultados mais promissores de forma agregada. Porém em cenários com distribuição de classe desequilibrada como este e que nenhuma informação será passada para a geração das máscaras, o mais indicado aparenta ser algoritmos que extrapolem a classe mais comum como o Hierarchical clustering.

6. Trabalhos futuros

Para trabalhos futuros, pretende-se aplicar a metodologia em outros datasets de pólipos colorretais, além de mais cenários médicos, também visa a possibilidade de expandir a metodologia para um cenário multiclasses e testar outros algoritmos de clusterização. Aplicações de *VQA* em modelos multimodais gerais com os prompts sintéticos geradas também é uma possibilidade.

7. Agradecimentos

Este trabalho foi parcialmente financiado pelo CNPq e pelo projeto Inteligência Artificial Aplicada na Detecção de Anomalias em Vídeos no Apoio à Tomada de Decisão, apoiado pelo Centro de Excelência em Inteligência Artificial (CEIA), Embrapii, ZSCAN e SEBRAE, com recursos financeiros do processo nº PEIA-2501.0109.

References

Amin, J., Sharif, M., Haldorai, A., Yasmin, M., and Nayak, R. S. (2022). Brain tumor detection and classification using machine learning: a comprehensive survey. *Complex & intelligent systems*, 8(4):3161–3183.

de Oliveira Santos, M., de Lima, F. C. d. S., Martins, L. F. L., Oliveira, J. F. P., de Almeida, L. M., and de Camargo Cancela, M. (2023). Estimativa de incidência de câncer no brasil, 2023-2025. *Revista Brasileira de Cancerologia*, 69(1).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Hao, S., Zhou, Y., and Guo, Y. (2020). A brief survey on semantic segmentation with deep learning. *Neurocomputing*, 406:302–321.

Islam, S., Elmekki, H., Elsebai, A., Bentahar, J., Drawel, N., Rjoub, G., and Pedrycz, W. (2024). A comprehensive survey on applications of transformers for deep learning tasks. *Expert Systems with Applications*, 241:122666.

Jha, D., Smedsrud, P. H., Riegler, M. A., Halvorsen, P., de Lange, T., Johansen, D., and Johansen, H. D. (2020). Kvasir-seg: A segmented polyp dataset. In *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II*, page 451–462, Berlin, Heidelberg. Springer-Verlag.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Li, L. and Iskander, M. (2022). Use of machine learning for classification of sand particles. *Acta Geotechnica*, 17(10):4739–4759.

Li, Z., Li, Y., Li, Q., Wang, P., Guo, D., Lu, L., Jin, D., Zhang, Y., and Hong, Q. (2023). Lvit: language meets vision transformer in medical image segmentation. *IEEE transactions on medical imaging*, 43(1):96–107.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137.

Lüddecke, T. and Ecker, A. (2022). Image segmentation using text and image prompts. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7076–7086.

Poudel, K., Dhakal, M., Bhandari, P., Adhikari, R., Thapaliya, S., and Khanal, B. (2024). Exploring transfer learning in medical image segmentation using vision-language models. In Burgos, N., Petitjean, C., Vakalopoulou, M., Christodoulidis, S., Coupe, P., Delingette, H., Lartizien, C., and Mateus, D., editors, *Proceedings of The 7nd International Conference on Medical Imaging with Deep Learning*, volume 250 of *Proceedings of Machine Learning Research*, pages 1142–1165. PMLR.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., and Lu, J. (2022). Denseclip: Language-guided dense prediction with context-aware prompting. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18061–18070.

Reynolds, D. (2015). Gaussian mixture models. In *Encyclopedia of biometrics*, pages 827–832. Springer.

Siegel, R., Kratzer, T., Giaquinto, A., Sung, H., and Jemal, A. (2025). Cancer statistics, 2025. *CA: A Cancer Journal for Clinicians*, 75(1):10–45.

Sokal, R. R., Michener, C. D., et al. (1958). A statistical method for evaluating systematic relationships.

Wang, Y., Su, J., Li, X., and Nakahara, E. (2025). Medlangvit: A language–vision network for medical image segmentation. *Electronics*, 14(15):3020.

Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., and Liu, T. (2022). Cris: Clip-driven referring image segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11676–11685.