RESUMO I FÓRUM BRASILEIRO DE PECUÁRIA DE PRECISÃO, IA E SUSTENTABILIDADE - 1.1) PECUÁRIA DE PRECISÃO E IA

AVALIAÇÃO ZERO-SHOT DO LOCATEANYTHING-3B PARA DETECÇÃO DE BOVINOS EM IMAGENS AÉREAS DE VANT NO CERRADO DE MATO GROSSO DO SUL

Max Hiroito Tieti (ra868222@ucdb.br)

Roberto Pereira De Freitas Neto (neto2014roberto@gmail.com)

Bianca Sabka De Queiroz (biancasabka@gmail.com)

Maria Rita De Araújo (mariaritadearaujo3@gmail.com)

André Gonçalves Da Silva (instagram.andre12@gmail.com)

Heron Leal Farias (leal.heron@hotmail.com)

Magnum Johanson De Abreu (magnumjabreuu@gmail.com)

Miguel Monteiro Cruz Matos (miguelmcmatos@gmail.com)

Allan Rios Bezerra (allanrb07@gmail.com)

O monitoramento de rebanhos bovinos por imagens aéreas de VANT representa uma alternativa escalável à vistoria manual em propriedades extensivas, porém sua adoção depende de modelos capazes de detectar animais sem exigir grandes volumes de dados rotulados. Modelos de linguagem-visão (VLMs) com capacidade de grounding visual emergem como candidatos naturais para detecção zero-shot, eliminando a etapa de anotação — contudo, sua viabilidade em imagens aéreas de bovinos no contexto

brasileiro permanece sem avaliação sistemática na literatura. Enquanto abordagens supervisionadas com tiling adaptativo atingem mAP@50 superior a 80% neste mesmo dataset, sua aplicação exige coleta e anotação prévia de dados — etapa custosa e nem sempre viável em propriedades rurais. Este trabalho investiga se VLMs de grounding generalistas podem constituir uma alternativa zero-shot viável, sem necessidade de treinamento específico.

Este estudo apresenta a primeira avaliação sistemática do LocateAnything-3B (NVIDIA, 2026) para detecção zero-shot de bovinos em imagens aéreas, conduzida sobre 396 fotografias obtidas por VANT (4000×3000 px) em propriedade no Cerrado de Mato Grosso do Sul, com 2.557 anotações de referência no formato COCO. Foram realizados 94 experimentos cruzando sete estratégias de prompt, três esquemas de tiling, três resoluções máximas de entrada e três modos de geração, avaliados por mAP@[0.50:0.95], F1, precisão, revocação e IoU médio.

A melhor configuração obtida — prompt "bovine", tiling 2×2, modo slow — atingiu mAP@0.50 = 0,141 e F1 = 0,295, com revocação de 15,5%, confirmando que o modelo, mesmo otimizado, deixa de detectar a maioria dos animais. O fator de maior impacto isolado foi o tiling: a grade 2×2 superou a ausência de segmentação com efeito de tamanho enorme (Cohen's d = 2,08; ANOVA $F=26,88$, $p<0,001$). A análise combinada de resolução e tiling revelou um ponto ideal em torno de 640 px por tile — janela em que o animal é reconhecível sem perda de contexto espacial. A especificidade do prompt também foi significativa ($F=4,48$, $p=0,0013$): "bovine" superou termos genéricos com $d=1,53$, sugerindo cobertura desigual no treinamento. O controle negativo ("car", "truck") confirmou ausência de detecções espúrias massivas (mAP@0.50 = 0,02).

Os resultados indicam que o LocateAnything-3B compreende minimamente o domínio aéreo — ausência de detecções espúrias no controle negativo sugere discriminação semântica funcional — mas sua escala de treinamento, voltada a objetos terrestres de grande aparência visual, limita severamente a revocação em bovinos aéreos. Para aplicações que exijam contagem confiável, o modelo zero-shot ainda não é viável. Contudo, combinado com tiling na resolução efetiva adequada e prompts taxonomicamente específicos, constitui uma linha de base relevante e um guia de decisão para trabalhos futuros com VLMs neste domínio — sem custo de anotação e com implantação imediata.

Palavras-chave: detecção de objetos; modelos de linguagem-visão (vlm); grounding visual; detecção zero-shot; imagens aéreas de vant; pecuária de precisão.