

COMPACT VISION-LANGUAGE MODEL FOR EXPLAINABLE CLASSIFICATION OF NEOPLASTIC LESIONS IN GI ENDOSCOPY

Keisuke Tomiyasu

UEG Week Berlin 2025

Poster

Oesophagus

October 6, 2025

A hybrid AI system combining CNN-based lesion detection with a compact vision-language model generates structured explanations for upper gastrointestinal endoscopy images while improving diagnostic accuracy for neoplastic versus non-neoplastic lesions.

KEY POINTS

- The system integrates a DETR-RT detector with a lightweight Gemma-3-4B vision-language model, trained using distillation from larger models (Gemini-2.5-Pro and GPT-4.1) on 1247 training images from 197 clinical cases.
- On a held-out test set of 313 gastric images, the system achieved 67% sensitivity and 85% specificity for distinguishing neoplastic from non-neoplastic lesions, representing 3% and 2% improvements respectively over the detector alone.
- Structured explanations achieved 100% completeness across consistent fields, with component-wise accuracy of 85% for organ identification, 100% for light color, 75% for lesion type, and 83% for JGES classification.
- The system maintains clinically practical inference speed with CNN-based detection at approximately 10ms per input and VLM generation at 92.8 tokens per second, suitable for post-procedure evaluation.
- This work addresses the black-box limitation of most AI endoscopy systems by providing explanatory outputs based on morphological and contextual features without sacrificing diagnostic performance.

VIEW THIS CONTENT ON GUTFLIX

<https://gutflix.eu/search/d/413426e8-a813-11f0-8ab8-0242ac140006>



AI-generated summary

This summary was generated by an AI large language model based on the content transcript. It is for informational purposes only and should not be considered a substitute for clinical judgment. Always rely on your professional expertise and the full clinical context when making clinical decisions.