

From Vision Models to Vision-Language Models

Vision Models

- Vision models achieve strong performance on tasks such as classification and segmentation, but their outputs are often limited in flexibility.
 - We increasingly want models to produce open-ended, free-form text, rather than selecting from a fixed set of classes.
 - Many visual problems also require reasoning beyond pattern matching.

Reasoning beyond Pattern Matching

- Some visual tasks are inherently difficult and require reasoning through language, not just visual pattern matching.

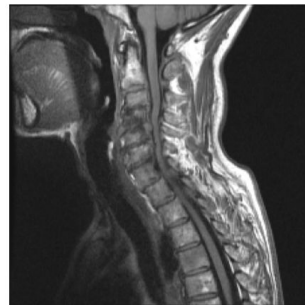
Problem 3: What is the finding in this image?

A) Stomach ulcer, B) Nerve compression,
C) Cord pathology, D) Liver cirrhosis

MedVLM-R1 Output:

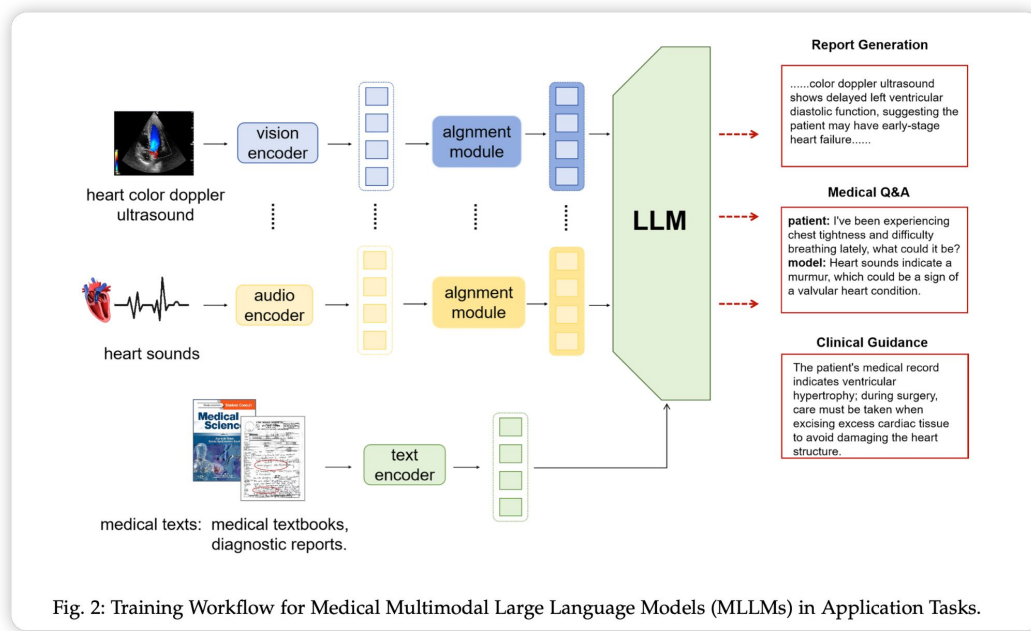
<think> To determine the correct answer, we need to identify the finding in the image provided. The MRI scan in question is likely a view of a spine, not the stomach, or the liver. Stomach ulcers, liver cirrhosis, and spinal cord pathology can be diagnosed with MRI, but the image in question shows a spinal view, not a gastrointestinal or abdominal view. Therefore, the only reasonable options left are nerve compression and cord pathology.
</think>

<answer>C</answer> **Groundtruth Answer: C**



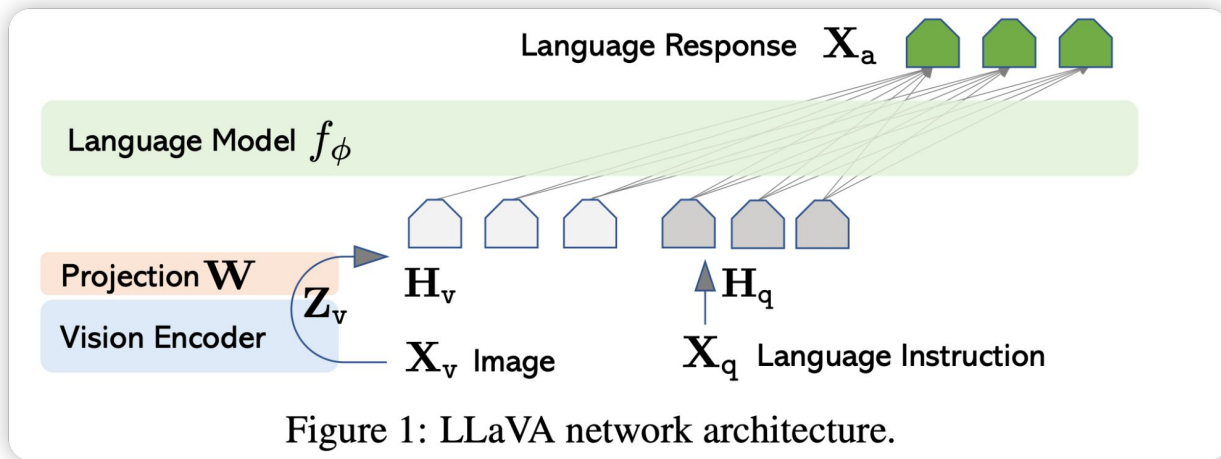
Healthcare Needs Multimodal, Generative Models

- Healthcare requires integrating multimodal inputs and generating free-form responses for models to be practically useful.



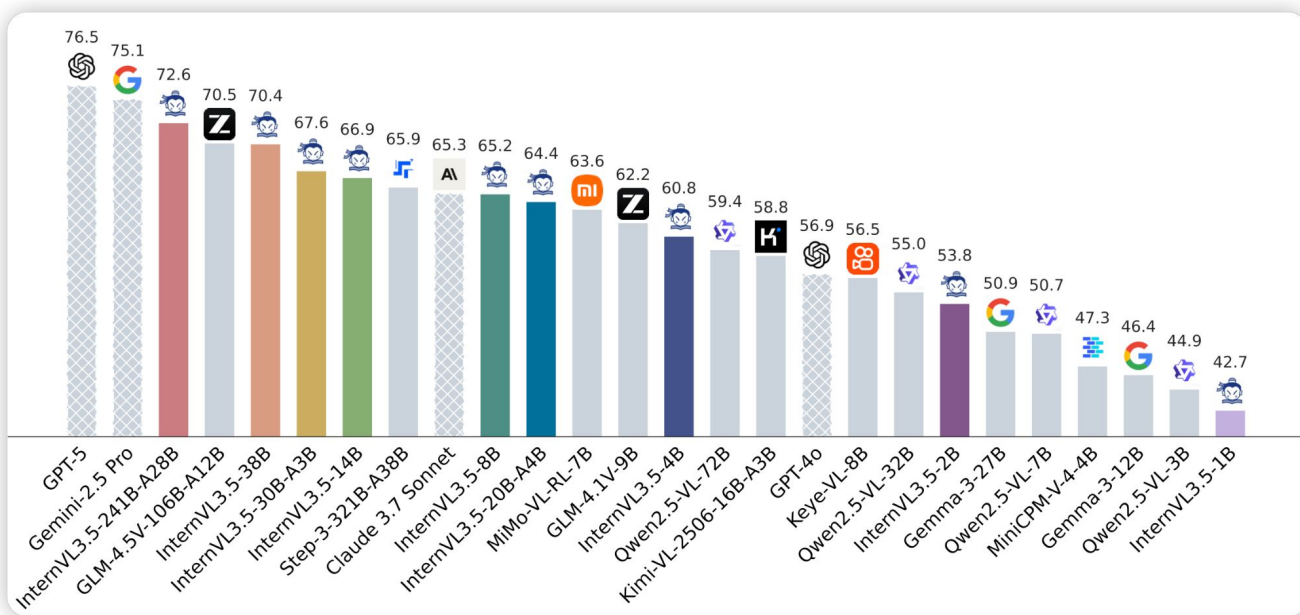
Vision-Language Model

- Visual embeddings from a CNN or ViT are converted into visual tokens and fed into the LLM input stream.



Vision-Language Model

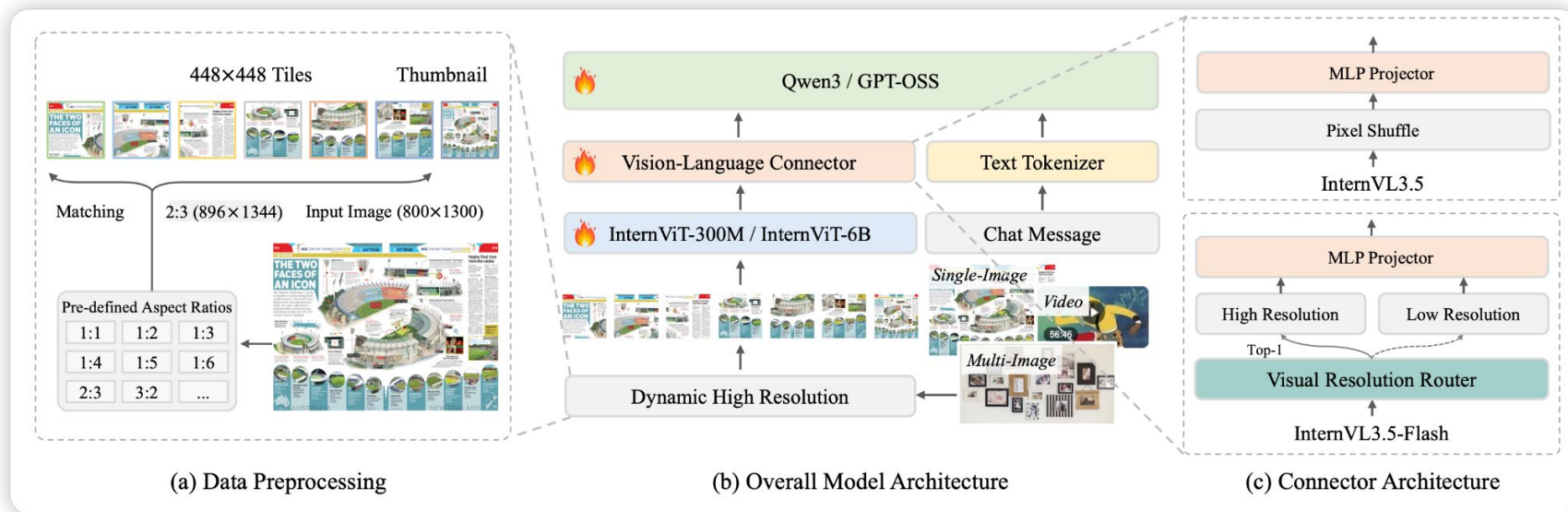
- VLMs have been rapidly adopted across both research and industry.



Training Pipeline for Vision-Language Models

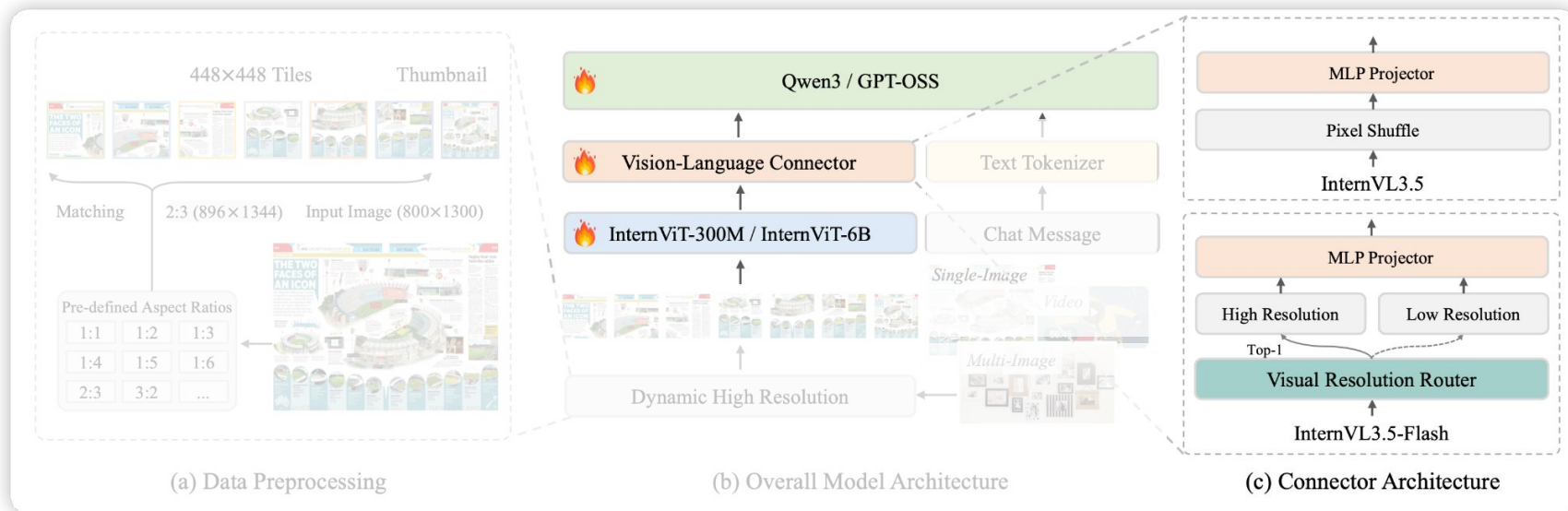
VLM Training

- Training Pipeline for Vision-Language Models



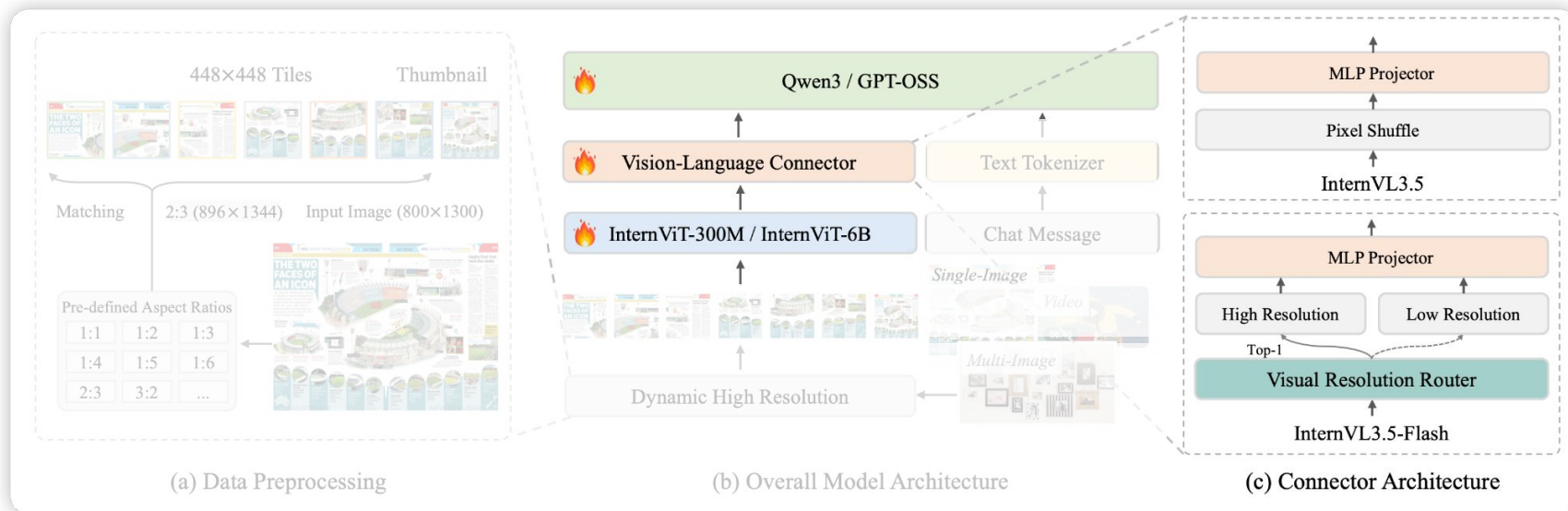
VLM Training

- Stage 1: vision-language alignment pretraining



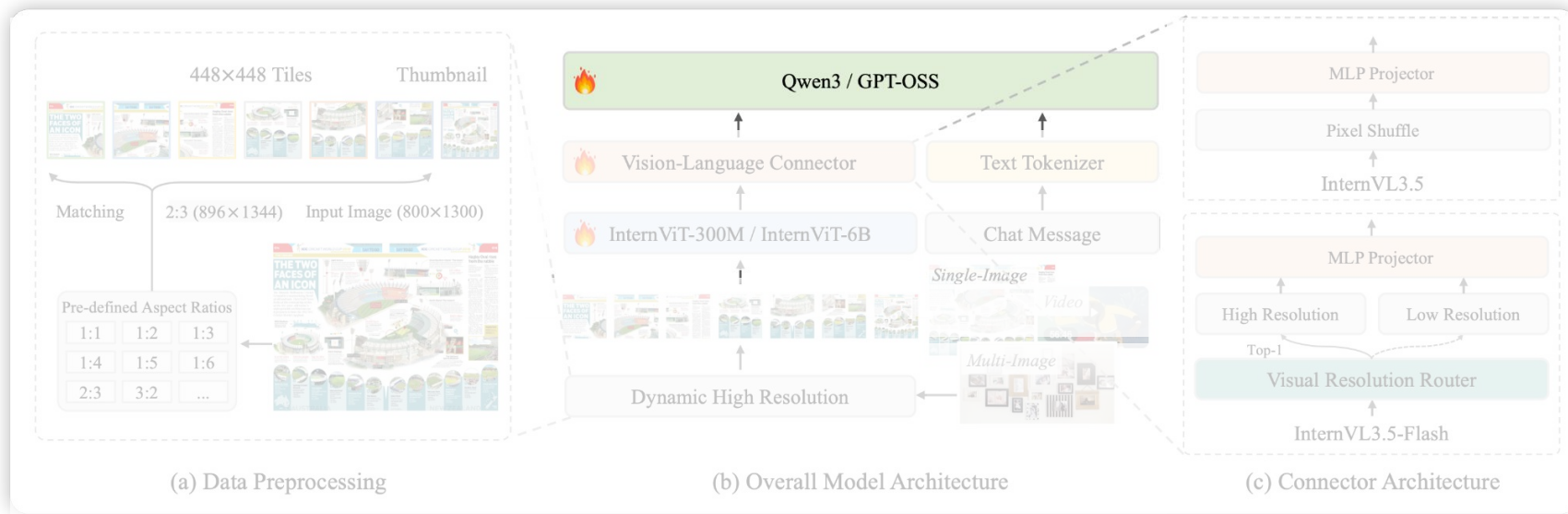
VLM Training

- Stage 2 multi-task pretraining: captioning, VQA, OCR, etc.



VLM Training

- Stage 3 Instruction fine-tuning: “Describe this image.”



VLM Training

- Stage 4 Direct Preference Optimization (DPO):

