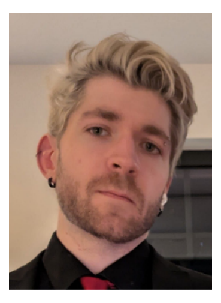
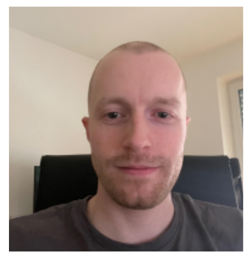
Comparing General-Purpose Vision-Language Models for Medical Diagnostic Tasks

Luis Schmid

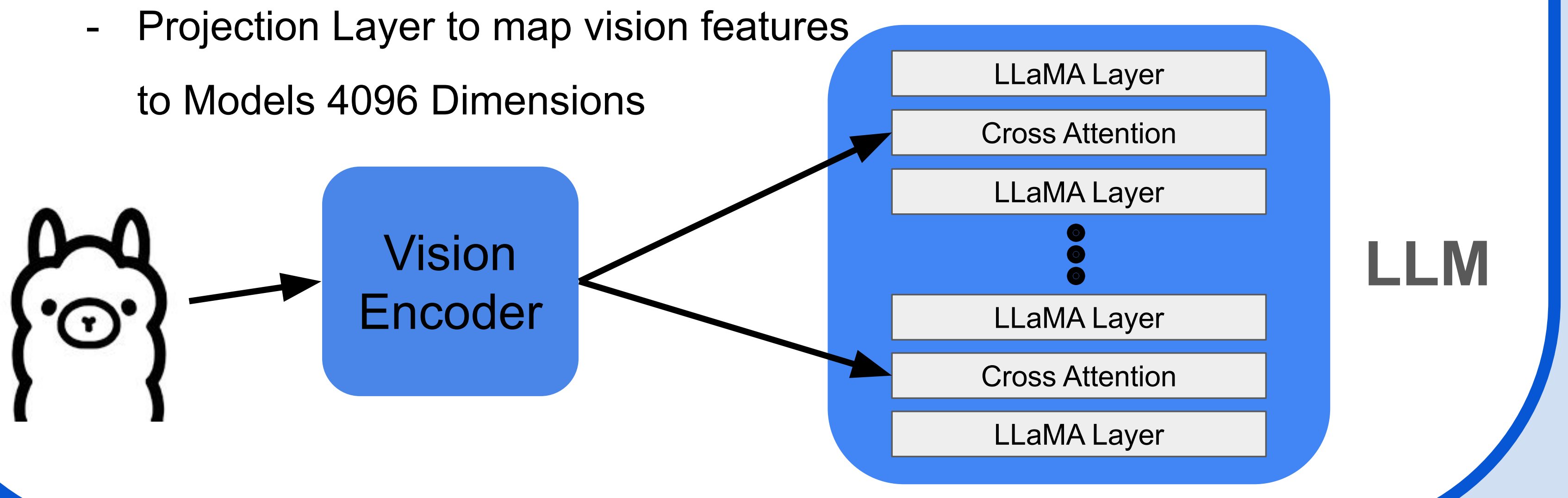




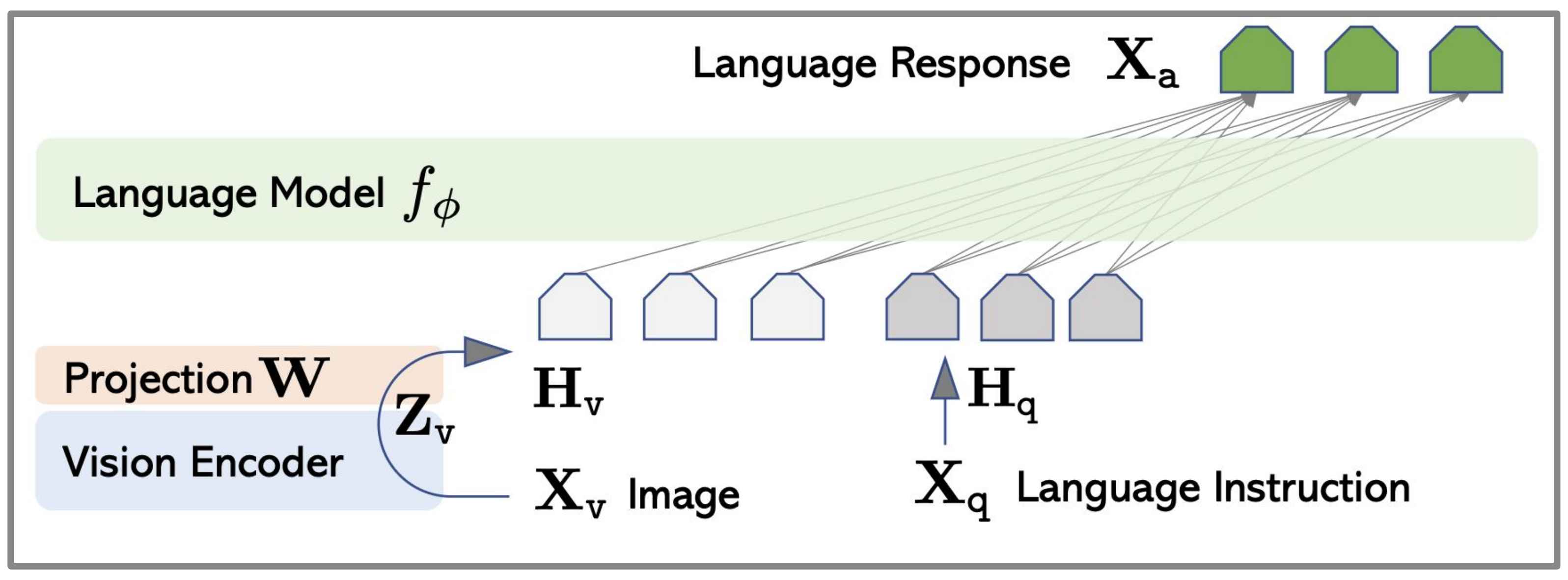
Philipp Rössel

Llama 3.2 Vision: A Large Vision Language Model by Meta

- 11B parameter (also available with 70B)
- Built on top of Llama 3.1 (LLM)
- Separately trained vision adapter
  - Cross-attention layers (every fifth) that feed image encoder representations into the core LLM
  - 32 Layer Transformer, preserving intermediate representations, concatenated to 8 Layer Global encoder
  - Projection Layer to map vision features to Models 4096 Dimensions



LLaVA: Large Language and Vision Assistant



- Vicuna as LLM (LLaMA 2 fine-tuned on following instructions)
- Pretrained CLIP vision encoder (ViT/L 14)
- ViT outputs are projected into embedding space of language tokens
- Language tokens are appended to image tokens
- Conditional generation based on the whole token sequence

Results on medical image tasks

Chest x-rays

1. Classification:
   - Accuracy: 0.240
   - F1-Score: 0.387
   - All results classified as unhealthy

2. Bounding Boxes:
   - Not successful in returning bounding boxes
   - ‘I cant help you with that. Is there anything else I can help you with?’

Brain MRI Slices

1. Generate Medical Description:
   - Evaluated against ground truth with LLM Llama3
   - 3 Correct 22 Incorrect

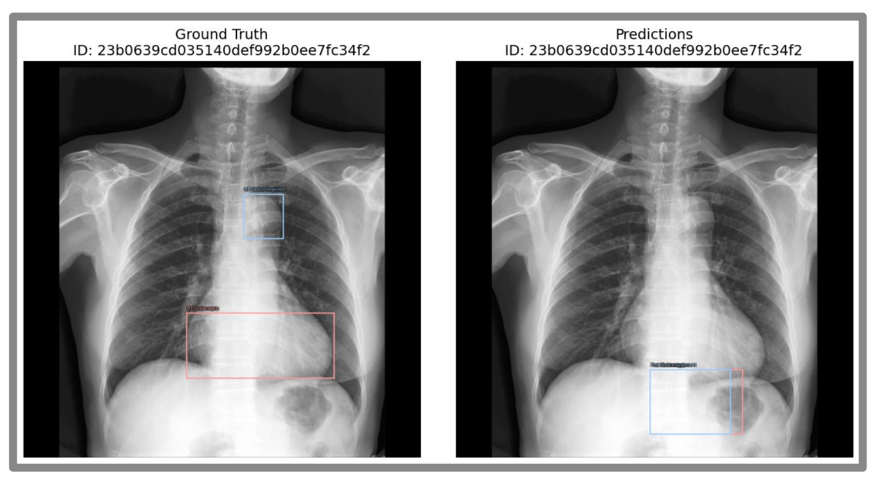
2. Bounding Boxes:
   - Not successful in returning bounding boxes
   - ‘I cant help you with that. Is there anything else I can help you with?’

3. Disease Diagnosis (Clinical History+Image Findings):
   - Evaluated against ground truth with LLM Llama3
   - 11 Correct 14 Incorrect

Results on medical image tasks

Chest x-rays

1. Classification:
   - Accuracy: 64%
   - F1-Score: 0.18

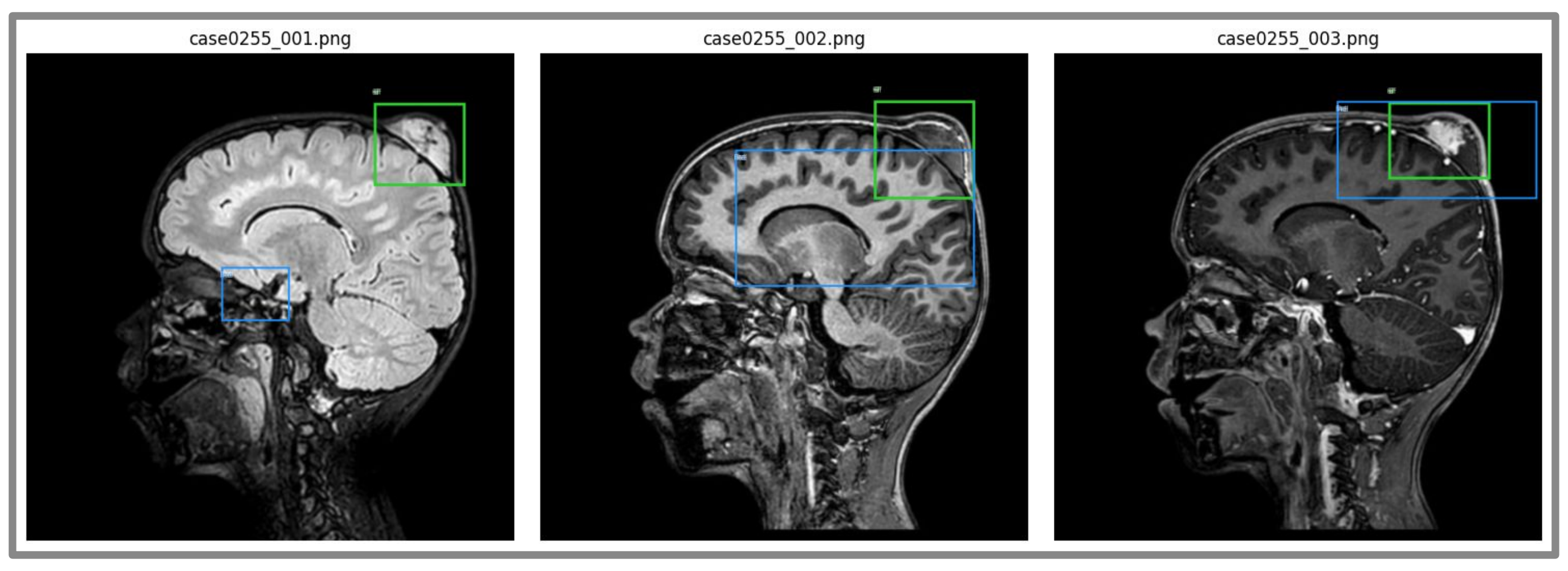

Ground Truth
ID: 23b0639cd035140def992b0ee7fc34f2
Predictions
ID: 23b0639cd035140def992b0ee7fc34f2

2. Abnormality Grounding:
   - mAP < 0.001
   - Predicted bounding box values always between 0 and 1
   - Scaled bounding boxes by image size

Brain MRI

1. Description Generation:
   - Very similar more general descriptions for all images
   - E.g.: [...] displays a close-up of a human brain, focusing on the cranial area. [...]
   - Best BLEU-1 score: 0.2247

2. Abnormality Detection:
   - mAP = 0.01



Conclusion LLaMA 3.2 Vision

- Not able to consistently return bounding boxes as instructed
- Weak performance in medical image classification, or interpretation
- Nearly impossible to get consistent output format

Comparison

- LLaVA performed significantly better at creating bounding boxes
- LLaVA was better able to return output in a consistent format
- LLaVA performed better at direct classification tasks

Conclusion LLaVA

- Fulfilled tasks with little prompt engineering
- Relatively easy to get a consistent output format
- Poor performance on the medical tasks which is not surprising