# Data

Primary Bone Tumor Radiograph Dataset (Stanford)

- **3,746 images** (1,879 normal / 1,867 tumor)
- Foot, knee, ankle, hip
- Benign & malignant tumors; labels + boxes + masks

GRAZPEDWRI-DX – Pediatric Wrist Trauma

- **20,327** pediatric wrist X-rays
- Trauma-focused; classification + boxes + masks + multi-views

FracAtlas – Musculoskeletal Fractures

- **4,083 images** (717 fractures, 922 fracture instances)
- Upper & lower extremities; fracture labels + boxes + masks

Simple vs. Comminuted Fracture Dataset (Mendeley)

- **16,061 images** (original + augmented)
- Labels: simple vs comminuted

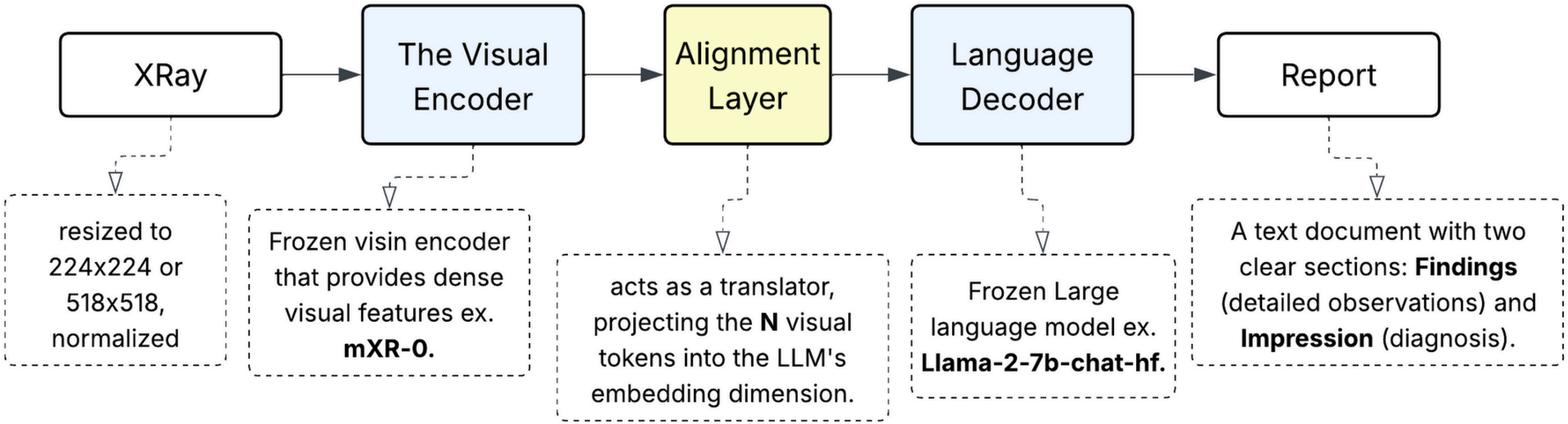
HBFMID – Human Bone Fractures (X-ray + MRI)

- **510 X-rays + 131 MRI images**
- Elbow, forearm, humerus, shoulder, femur, tibia/fibula

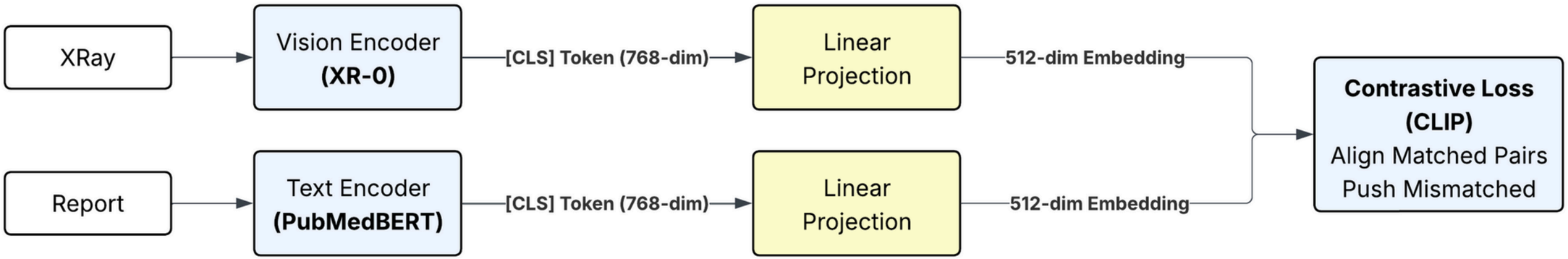
Multi-Region Bone Fracture Dataset

- **10,580 radiographs**
- fracture vs non-fracture

# End-to-End Vision-Language Pipeline



# mXR-0: Multimodal Pretraining Architecture



# Models

| Model Name | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE-L | CIDEr |
|---|---|---|---|---|---|---|
| mXR-0 (combined) | 42.8 | 28.6 | 20.7 | 15.7 | 36.8 | 42.2 |
| XR-0 (combined) | 41.0 | 27.1 | 19.5 | 14.7 | 36.9 | 40.1 |
| DeepMedix-R1 (Findings) | 39.58 | 25.35 | 18.34 | 13.62 | 37.53 | - |
| LLaVA-Rad>>MIMIC-CXR | _ | _ | _ | _ | 0.291 | _ |
| LLaVA-Rad>>Open-I | _ | _ | _ | _ | 0.347 | _ |
| LLaVA-Rad>>CheXpert | _ | _ | _ | _ | 0.298 | _ |

# Models

| Model Name | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE-L | METEOR | CIDEr |
|---|---|---|---|---|---|---|---|
| R2Gen (2020)>>IU X-Ray | 0.470 | 0.304 | 0.219 | 0.165 | 0.371 | 0.187 | _ |
| R2Gen (2020)>>MIMIC-CXR | 0.353 | 0.218 | 0.145 | 0.103 | 0.277 | 0.142 | _ |
| R2Gen-Mamba>>IU X-Ray | 0.487 | 0.315 | 0.229 | 0.175 | 0.384 | 0.195 | _ |
| R2Gen-Mamba>> MIMIC-CXR | 0.362 | 0.226 | 0.154 | 0.112 | 0.287 | 0.150 | _ |
| MambaXray-VL>>CheXpert Plus | 0.348 | 0.330 | 0.241 | 0.185 | 0.276 | 0.157 | 0.139 |

# Evaluation Matrices

## BLEU-1, 2, 3, 4 (Bilingual Evaluation Understudy)

precision-based metric that counts the overlap of words between he generated report and the reference report.(how much of the output is correct)

## ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation)

measures the Longest Common Subsequence between the generated report and the reference. It looks for the longest sequence of words that appear in both texts in the same relative order, even if they are not consecutive.(how much of the reference was captured)

## CIDEr (Consensus-based Image Description Evaluation)

CIDEr is a specialized metric designed specifically for the evaluation of image captioning systems. Unlike BLEU that treat all words equally or rely purely on n-gram precision, CIDEr measures the agreement between a candidate caption and a set of human reference captions based on human consensus.

## GREEN (Generative Radiology Report Evaluation and Error Notation)

GREEN is a clinically focused metric that uses a fine-tuned Large Language Model (LLM) to evaluate radiology reports. Unlike the previous metrics which only match text patterns, GREEN analyzes the meaning to identify and explain clinically significant errors (e.g., False Positives, False Negatives, wrong location, or wrong severity). It provides a score (0 to 1) that aligns closely with expert radiologist preferences.