Output 2

Masked Language Model

MNL Task (Tokens Selection)

Semantic Embedding Vector

Position Embedding Vector

Tokenizer

Semantic Embedding

Visual Location

Visual Feature

Input Sequence

VE LE

IRM (Pairing)

Medical Image

Text

Description automatically generated

**JOINT EMBEDDING**

CNN Extraction

**VISUAL FEATURE EMBEDDING**

**LANGUAGE EMBEDDING**

1. Image report pairs and non-matching report pairs
2. Special Token
3. Images ( 512 by 512)
4. Free text report (long description)

Concatenate Block

**Pre-Training** (Mimic Open I Datasets)