Image Reconstruction

Masked Image Modeling (MIM) Loss

Rand Mask Ratio=75%

Shared Vision Encoder

Vision Decoder

Paired

GLA

Pretrained Text Encoder

t

Alignment

v

: Joint reconstruction (forward)

: Alignment (forward)

GLA: Global and Local Alignment

MA-CMF: Memory-Augmented Cross-Modal Fusion

......No focal consolidation s concerning for pneumonia are identified. there is no large pleural effusion or pneumothorax......

Tokenizer

Rand Mask Ratio=50%

......[MASK]focal [MASK] [MASK] concerning for [MASK] [MASK] identified. there [MASK] no large [MASK] effusion or [MASK]......

E_t

MA-CMF

Text Decoder

......No focal consolidation s concerning for pneumonia are identified. there is no large pleural effusion or pneumothorax......

Masked Language Modeling (MLM) Loss

Report Reconstruction