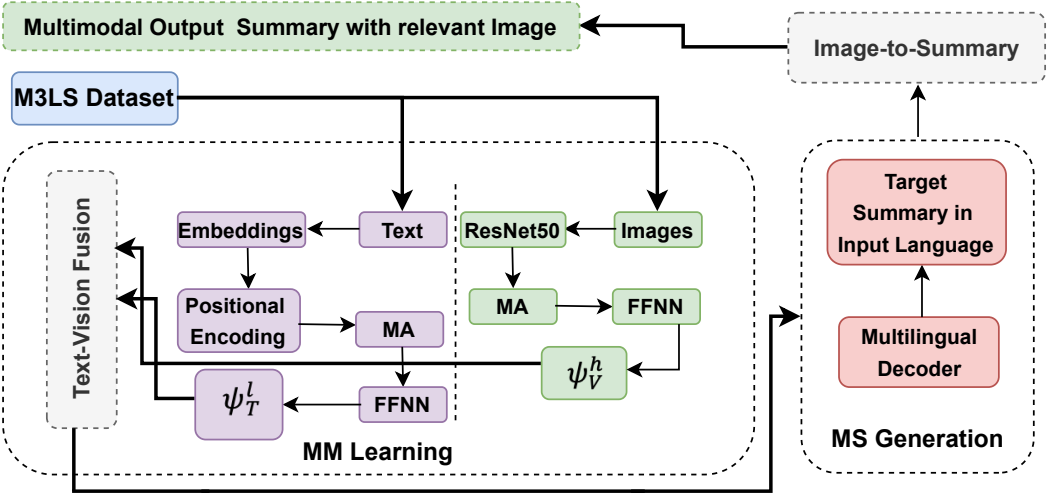
Multimodal Output Summary with relevant Image
Image-to-Summary
M3LS Dataset
Text-Vision Fusion
Embeddings
Text
ResNet50
Images
Positional Encoding
MA
MA
FFNN
ψ_V^h
ψ_T^l
FFNN
MM Learning
Target Summary in Input Language
Multilingual Decoder
MS Generation