Multi-scale Feature Maps

Visual features flow in VLEF

Linguistic features flow in VLEF

Multi-scale Deformable
Self Attention

Image Feature Maps

Visual-Language Early
Fusion

Image

Ft

Linguistic features

X 4