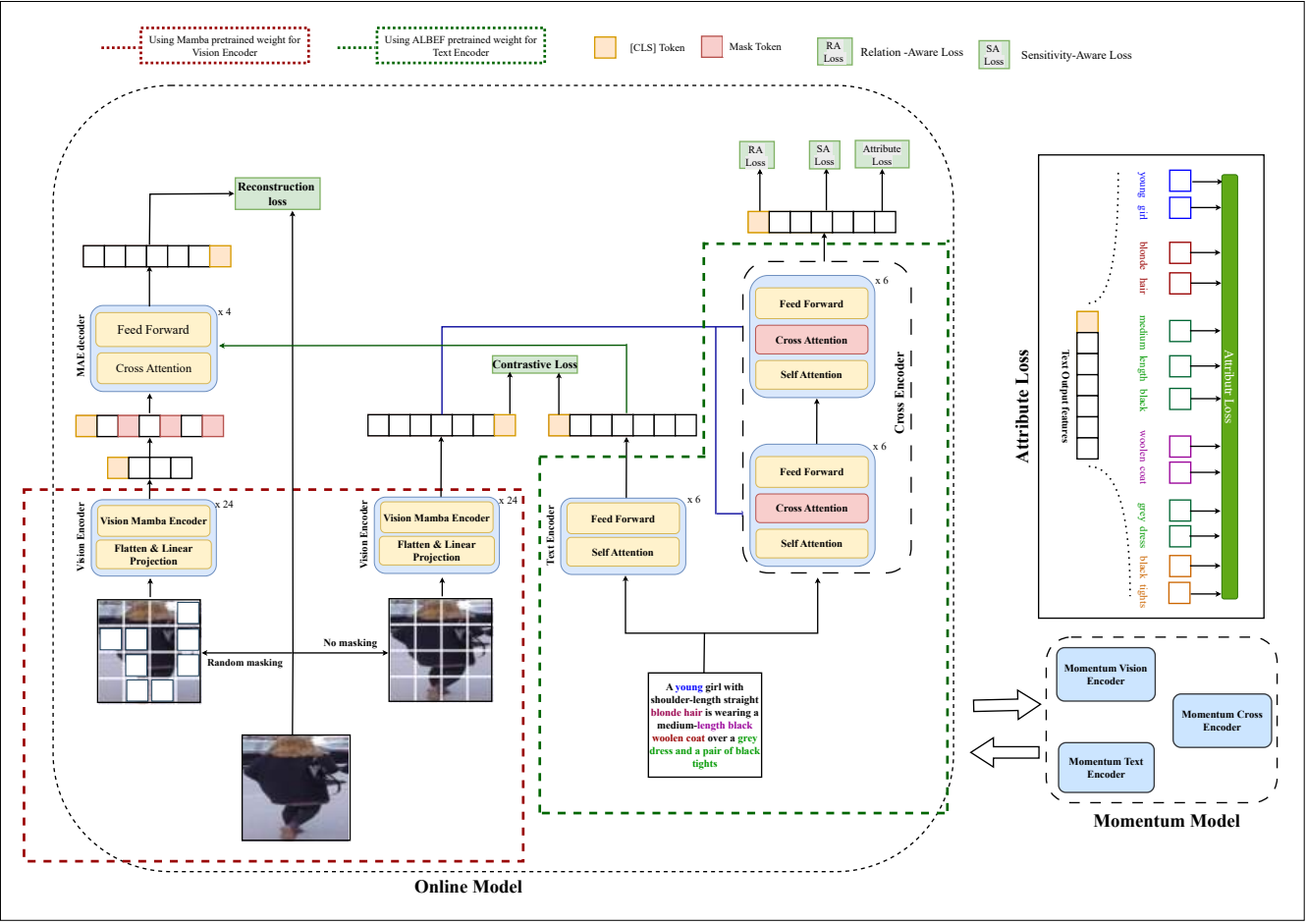
Using Mamba pretrained weight for Vision Encoder
Using ALBEF pretrained weight for Text Encoder
[CLS] Token
Mask Token
RA Loss
Relation -Aware Loss
SA Loss
Sensitivity-Aware Loss
Reconstruction loss
MAE decoder
Feed Forward
Cross Attention
x 4
Vision Encoder
Vision Mamba Encoder
Flatten & Linear Projection
x 24
Random masking
No masking
Contrastive Loss
Text Encoder
Feed Forward
Self Attention
x 6
Cross Encoder
Feed Forward
Cross Attention
Self Attention
x 6
RA Loss
SA Loss
Attribute Loss
A young girl with shoulder-length straight blonde hair is wearing a medium-length black woolen coat over a grey dress and a pair of black tights
Online Model
Attribute Loss
Text Output features
young girl
blonde hair
medium length black woolen coat
grey dress black tights
Attribute Loss
Momentum Vision Encoder
Momentum Cross Encoder
Momentum Text Encoder
Momentum Model