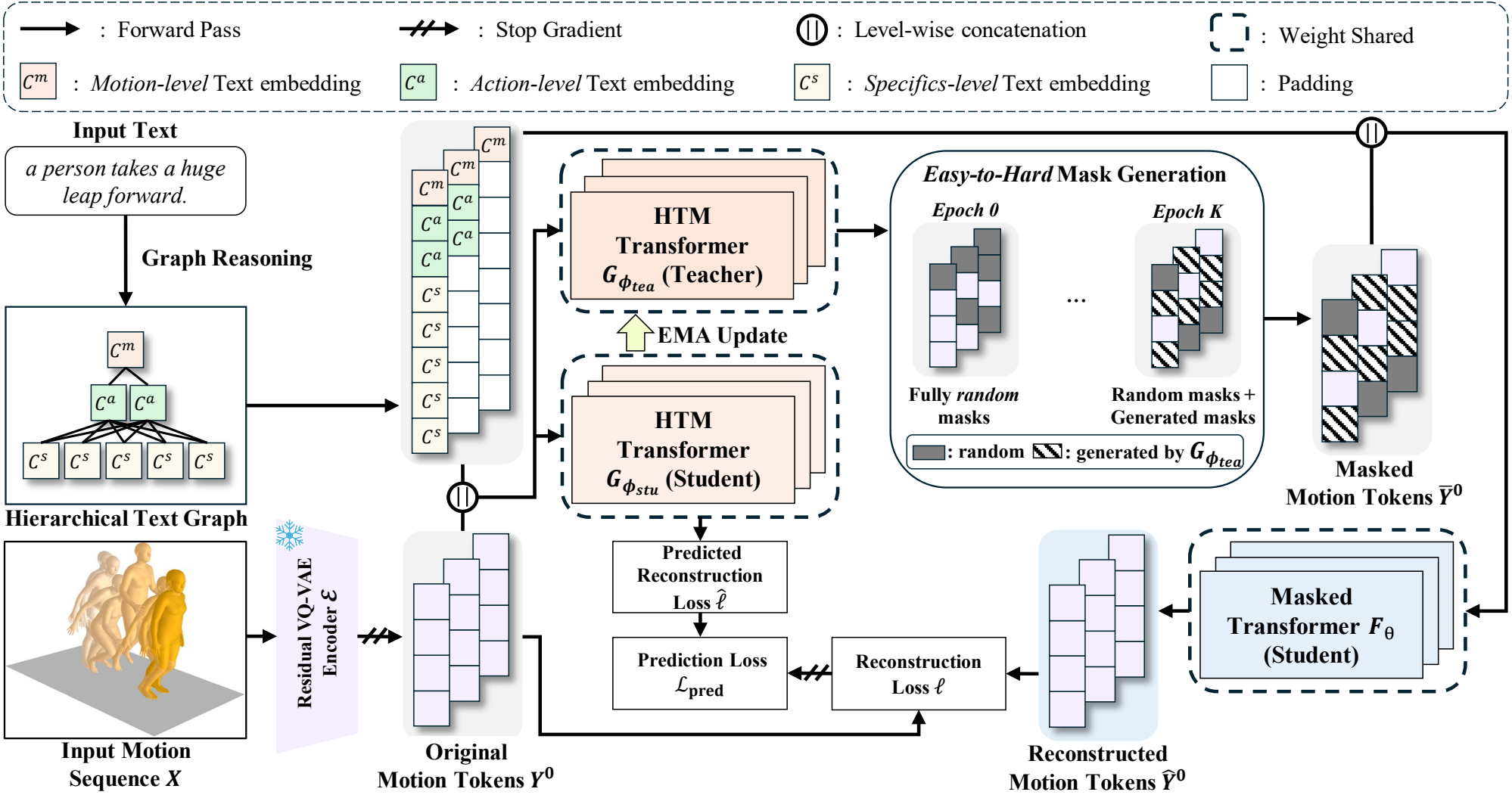
: Forward Pass
: Stop Gradient
: Level-wise concatenation
: Weight Shared
C^m : *Motion-level* Text embedding
C^a : *Action-level* Text embedding
C^s : *Specifics-level* Text embedding
: Padding
Input Text
a person takes a huge leap forward.
Graph Reasoning
Hierarchical Text Graph
Input Motion Sequence X
Residual VQ-VAE Encoder $\mathcal{E}$
Original Motion Tokens Y^0
HTM Transformer $G_{\phi_{tea}}$ (Teacher)
EMA Update
HTM Transformer $G_{\phi_{stu}}$ (Student)
Easy-to-Hard Mask Generation
Epoch 0
Epoch K
Fully *random* masks
Random masks + Generated masks
: random : generated by $G_{\phi_{tea}}$
Masked Motion Tokens $\bar{Y}^0$
Masked Transformer F_θ (Student)
Predicted Reconstruction Loss $\hat{\ell}$
Prediction Loss $\mathcal{L}_{\text{pred}}$
Reconstruction Loss ℓ
Reconstructed Motion Tokens $\hat{Y}^0$