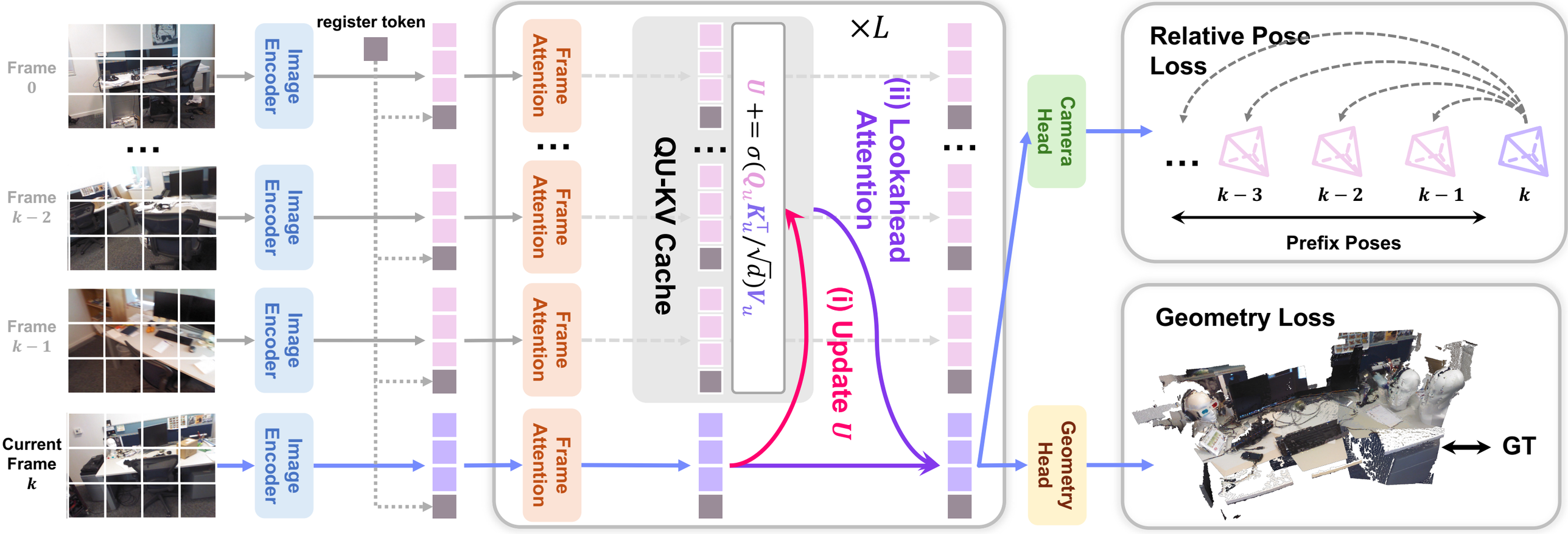

Frame 0
Frame $k-2$
Frame $k-1$
Current Frame k
Image Encoder
register token
Frame Attention
$\times L$
QU-KV Cache
$U \mathrel{+}= \sigma(Q_u K_u^\top / \sqrt{d}) V_u$
(i) Update U
(ii) Lookahead Attention
Camera Head
Geometry Head
Relative Pose Loss
$k-3$
$k-2$
$k-1$
k
Prefix Poses
Geometry Loss
GT