

Input Video

Video Gen. Encoder

Scene Token F

Query Decoder

 = Frozen
 = Trainable

4D Reconstruction & Tracking

t_{cam} t_{src} t_{tgt}

2D Pos.+RGB Patch

Continuous Time Query

$e_{\text{src}}(t_{\text{src}})$ $e_{\text{tgt}}(t_{\text{tgt}})$ $e_{\text{cam}}(t_{\text{cam}})$

Q
Query q

3D Points

3D Tracks

Camera Poses

