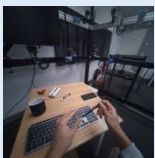
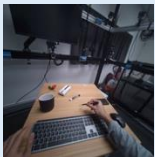
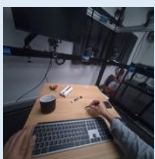
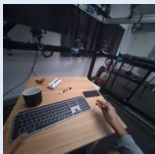
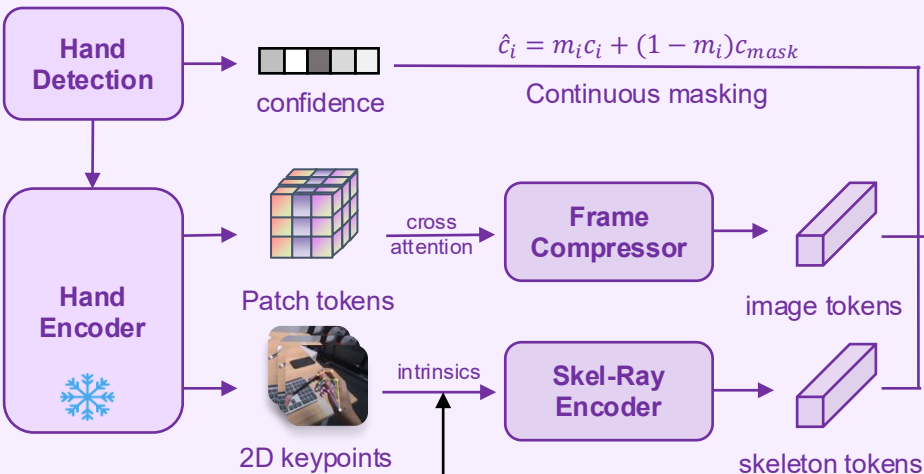


1. Input video sequence



T frames

2. Pre-processing



SLAM



$$x_0 \sim N(0, I)$$

...

x_t

Linear

Z_t

3. HandFlow denoiser

Dual-stream Transformer × L

c-stream



RoPE

z-stream

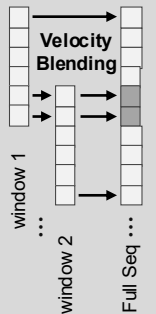


Single stream blocks × M

...

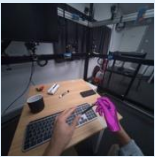
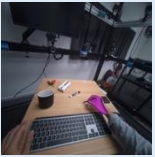
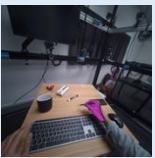
Output projection

ODE Solver



Camera to World

4. Output



T frames