Output latents from denoising

F0 | F1 | F2 | ... | F'-1

x DiT blocks

feedforward

cross-attention

projector

3D self-attention

patchify + concat

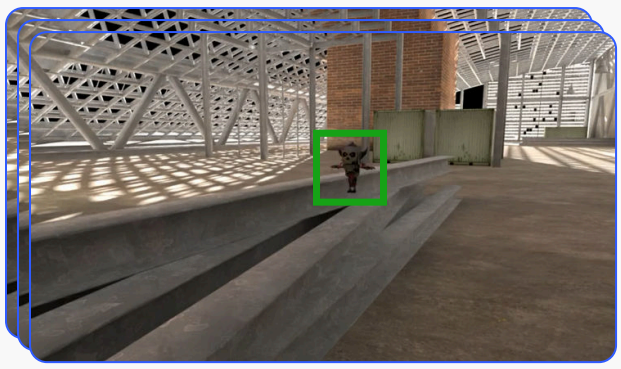
F0, F1, ..., FT-1 | F'0 | F'1 | ... | 2F'-1 | 2F'

noise ⊕ z_{trg}

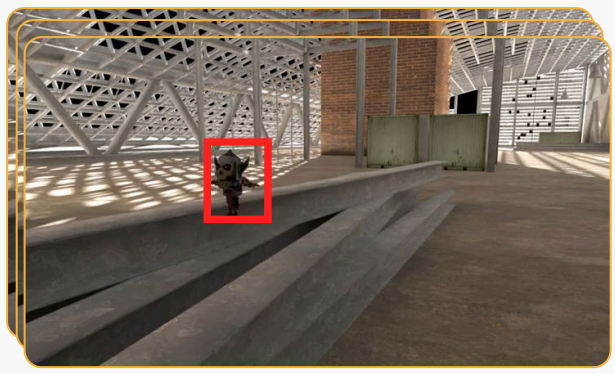
z_{src}

z_{bb}

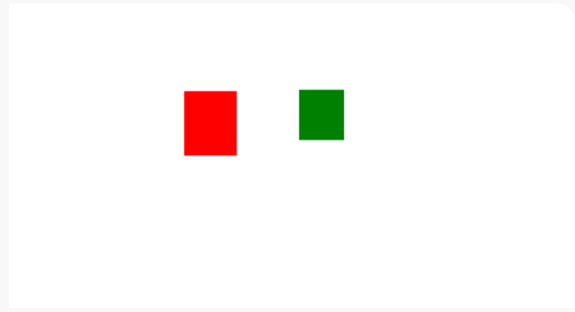
VAE | VAE | VAE



target video V_{trg}



source video V_{src}



displacement control signal I_{bb}