High frame rate video x_0 x_1 x_2 x_3 ...

Low frame rate video y_0 y_1 ...

Prior works: *synthesis middle frame based on adjoint frames*

y_0 → VFI Model → $\hat{x}_1$
y_1 →
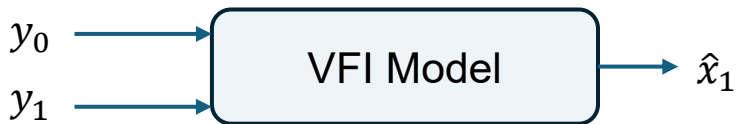
Our work: *synthesis all frames simultaneously in each video chunk*

$y_{0:T}$ → Latent Diffusion (DiT) z → Cond. VAE Decoder → $\hat{x}_{0:(2T)}$