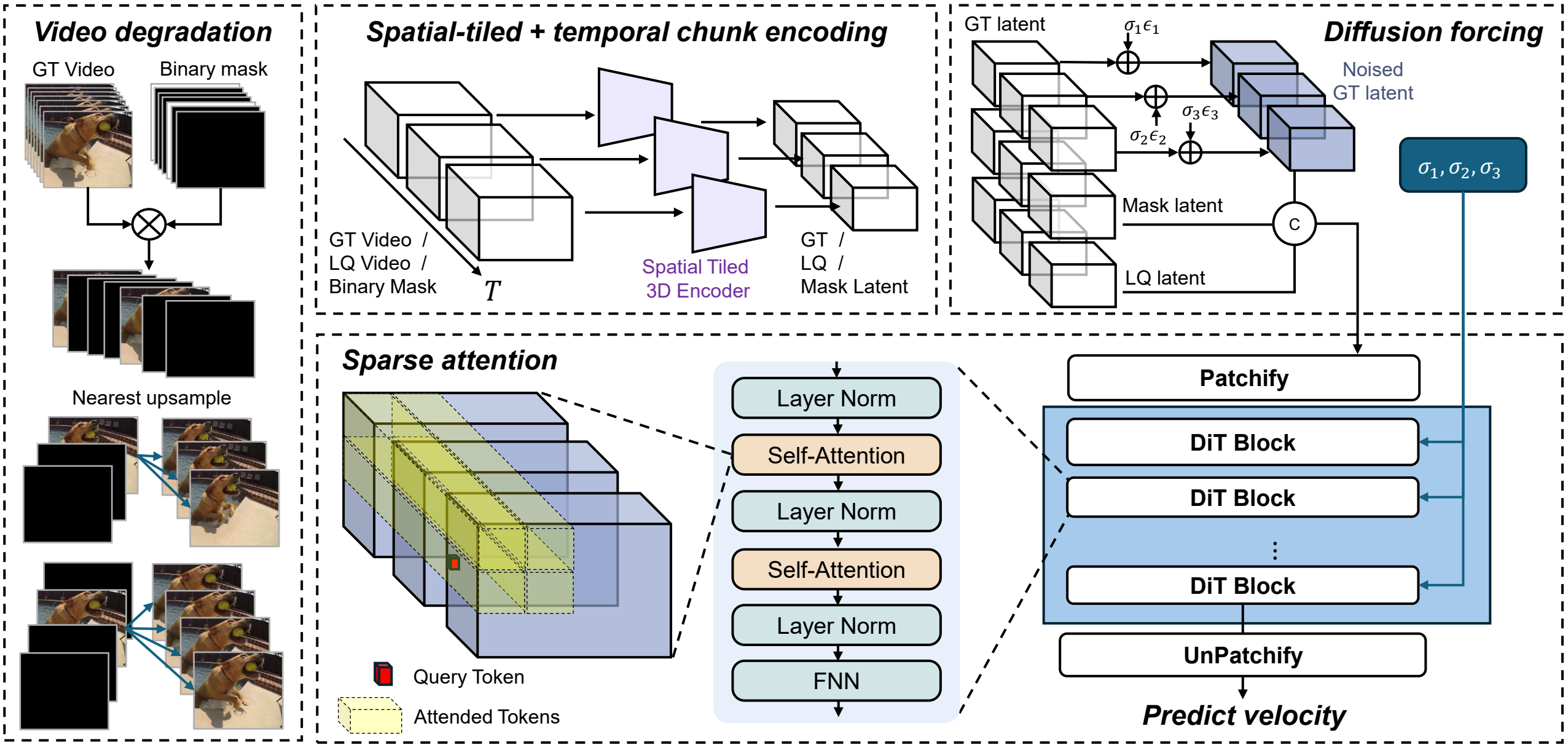
Video degradation
GT Video
Binary mask
Nearest upsample
Spatial-tiled + temporal chunk encoding
GT Video /
LQ Video /
Binary Mask
T
Spatial Tiled
3D Encoder
GT /
LQ /
Mask Latent
Diffusion forcing
GT latent
$\sigma_1\epsilon_1$
$\sigma_2\epsilon_2$
$\sigma_3\epsilon_3$
Noised
GT latent
$\sigma_1, \sigma_2, \sigma_3$
Mask latent
C
LQ latent
Sparse attention
Query Token
Attended Tokens
Layer Norm
Self-Attention
Layer Norm
Self-Attention
Layer Norm
FNN
Patchify
DiT Block
DiT Block
DiT Block
UnPatchify
Predict velocity