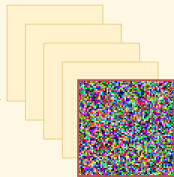**Flickered Sequence (5 frames)**

Reference ROI

Input Data

X_t

Forward Diffusion
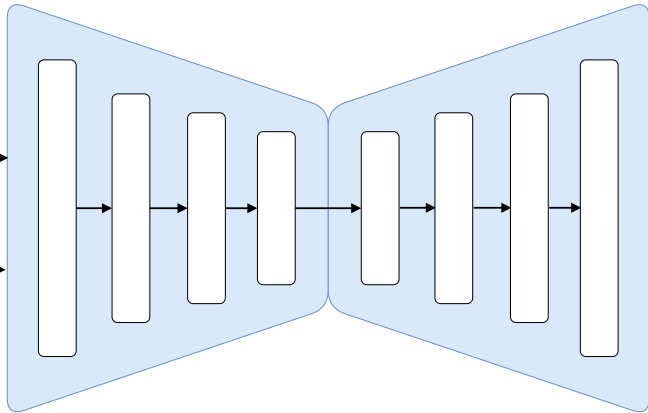
Combined Condition

Time Embed

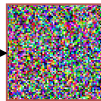Class Embed

Class-Conditioned 3D UNet

Predicted Noise $\hat{\epsilon}$

T steps