

Video UNet

$$\mathbf{X}_t$$
 $\downarrow N \times 64 \times 64 \times \mathbf{C}$

ResNet Block + Downsample

$$\downarrow N \times 32 \times 32 \times \mathbf{C}$$

$$\vdots$$

$$\downarrow N \times 8 \times 8 \times C$$

ResNet Block + Attn + Downsample

$$\downarrow N \times 4 \times 4 \times C$$

ResNet Block + Attn

$$\downarrow N \times 4 \times 4 \times C$$

ResNet Block + Attn + Upsample

$$\downarrow N \times 8 \times 8 \times C$$

$$\vdots$$

$$\downarrow N \times 32 \times 32 \times C$$

ResNet Block + Upsample

$$\sqrt[4]{N imes 64 imes 64 imes C}$$
 \mathbf{X}_{t-1}