$\mathbf{x}_T$

$\mathbf{x}_t$

$\mathbf{x}_{t-1}$

$\mathbf{x}_0$

Video UNet

$\mathbf{x}_t$

$N \times 64 \times 64 \times \mathrm{C}$

ResNet Block + Downsample

$N \times 32 \times 32 \times \mathrm{C}$

ResNet Block + Downsample

$N \times 16 \times 16 \times \mathrm{C}$

ResNet Block + Attn + Downsample

$N \times 8 \times 8 \times \mathrm{C}$

ResNet Block + Attn + Downsample

$N \times 4 \times 4 \times \mathrm{C}$

ResNet Block + Attn

$N \times 4 \times 4 \times \mathrm{C}$

ResNet Block + Attn + Upsample

$N \times 8 \times 8 \times \mathrm{C}$

ResNet Block + Attn + Upsample

$N \times 16 \times 16 \times \mathrm{C}$

ResNet Block + Upsample

$N \times 32 \times 32 \times \mathrm{C}$

ResNet Block + Upsample

$N \times 64 \times 64 \times \mathrm{C}$

$\mathbf{x}_{t-1}$