"Video of a V* dance" → Text transformer

| 🔒 Frozen | 🔓 Trainable |

Fine-tuning set

Customization D^m ∪ Video Regularization D^r

~ Sample + Add noise + Data augment

x_t → Video U-Net → x_{t-1}

Customize by tuning Text-to-Video model

Spatio-temporal Block

θ_s^c — Spatial Conv 🔒

θ_t^c — Temporal Conv 🔓

$\theta_s^{k,v}$ / θ_s^q — Spatial Attention

$\theta_t^{k,v}$ / θ_t^q — Temporal Attention