A Details and hyperparameters

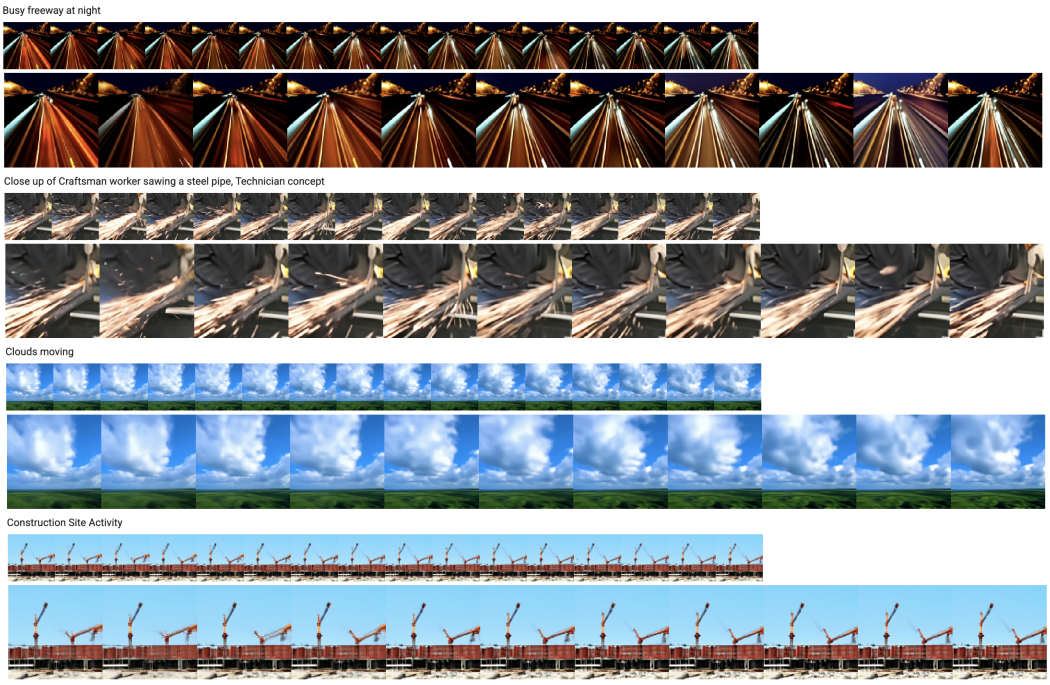

Figure 5: More samples accompanying Fig. 2.

Here, we list the hyperparameters, training details, and compute resources used for each model.

A.1 UCF101

Base channels: 256
Channel multipliers: 1, 2, 4, 8
Blocks per resolution: 2
Attention resolutions: 8, 16, 32
Attention head dimension: 64
Conditioning embedding dimension: 1024
Conditioning embedding MLP layers: 4
Diffusion noise schedule: cosine
Noise schedule log SNR range: $[-20, 20]$
Video resolution: 16x64x64 frameskip 1
Weight decay: 0.0

Optimizer: Adam ($\beta_1 = 0.9, \beta_2 = 0.99$)
Learning rate: 0.0003
Batch size: 128
EMA: 0.9999
Dropout: 0.1
Training hardware: 128 TPU-v4 chips
Training steps: 260000
Joint training independent images per video: 8
Sampling timesteps: 256
Sampling log-variance interpolation: $\gamma = 0.1$
Prediction target: $\boldsymbol{\epsilon}$

A.2 BAIR Robot Pushing

Base channels: 128
Channel multipliers: 1, 2, 3, 4
Blocks per resolution: 3
Attention resolutions: 8, 16, 32
Attention head dimension: 64
Conditioning embedding dimension: 1024
Conditioning embedding MLP layers: 2
Diffusion noise schedule: cosine
Noise schedule log SNR range: $[-20, 20]$
Video resolution: 16x64x64 frameskip 1
Weight decay: 0.01
Reconstruction guidance weight: 50

Optimizer: Adam ($\beta_1 = 0.9, \beta_2 = 0.999$)
Learning rate: 0.0002
Batch size: 128
EMA: 0.999
Dropout: 0.1
Training hardware: 128 TPU-v4 chips
Training steps: 660000
Joint training independent images per video: 8
Sampling timesteps: 256 (+256 Langevin cor.)
Sampling log-variance interpolation: $\gamma = 0.0$
Prediction target: $\mathbf{v}$
Data augmentation: left-right flips

A.3 Kinetics

Base channels: 256
Channel multipliers: 1, 2, 4, 8
Blocks per resolution: 2
Attention resolutions: 8, 16, 32
Attention head dimension: 64
Conditioning embedding dimension: 1024
Conditioning embedding MLP layers: 2
Diffusion noise schedule: cosine
Noise schedule log SNR range: $[-20, 20]$
Video resolution: 16x64x64 frameskip 1
Weight decay: 0.0
Reconstruction guidance weight: 9

Optimizer: Adam ($\beta_1 = 0.9, \beta_2 = 0.99$)
Learning rate: 0.0002
Batch size: 256
EMA: 0.9999
Dropout: 0.1
Training hardware: 256 TPU-v4 chips
Training steps: 220,000
Joint training independent images per video: 8
Sampling timesteps: 128 (+128 Langevin cor.)
Sampling log-variance interpolation: $\gamma = 0.0$
Prediction target: $\mathbf{v}$

A.4 Text-to-video

Small 16x64x64 model

Base channels: 128
Channel multipliers: 1, 2, 4, 8
Blocks per resolution: 2
Attention resolutions: 8, 16, 32
Attention head dimension: 64
Conditioning embedding dimension: 1024
Conditioning embedding MLP layers: 4
Diffusion noise schedule: cosine
Noise schedule log SNR range: $[-20, 20]$
Video resolution: 16x64x64 frameskip 1
Weight decay: 0.0

Optimizer: Adam ($\beta_1 = 0.9, \beta_2 = 0.99$)
Learning rate: 0.0003
Batch size: 128
EMA: 0.9999
Dropout: 0.0
Training hardware: 64 TPU-v4 chips
Training steps: 200000
Joint training independent images per video: 0, 4, 8
Sampling timesteps: 256
Sampling log-variance interpolation: $\gamma = 0.3$
Prediction target: $\boldsymbol{\epsilon}$

Large 16x64x64 model

Base channels: 256
Channel multipliers: 1, 2, 4, 8
Blocks per resolution: 2
Attention resolutions: 8, 16, 32
Attention head dimension: 64
Conditioning embedding dimension: 1024
Conditioning embedding MLP layers: 4
Diffusion noise schedule: cosine
Noise schedule log SNR range: $[-20, 20]$
Video resolution: 16x64x64 frameskip 1,4
Weight decay: 0.0

Optimizer: Adam ($\beta_1 = 0.9, \beta_2 = 0.99$)
Learning rate: 0.0003
Batch size: 128
EMA: 0.9999
Dropout: 0.0
Training hardware: 128 TPU-v4 chips
Training steps: 700000
Joint training independent images per video: 8
Sampling timesteps: 256
Sampling log-variance interpolation: $\gamma = 0.3$
Prediction target: $\boldsymbol{\epsilon}$

Large 9x128x128 model

Base channels: 128
Channel multipliers: 1, 2, 4, 8, 16
Blocks per resolution: 2
Attention resolutions: 8, 16, 32
Attention head dimension: 128
Conditioning embedding dimension: 1024
Conditioning embedding MLP layers: 4
Diffusion noise schedule: cosine
Noise schedule log SNR range: $[-20, 20]$
Video resolution: 9x128x128 frameskip 1
Weight decay: 0.0

Optimizer: Adam ($\beta_1 = 0.9, \beta_2 = 0.99$)
Learning rate: 0.0002
Batch size: 128
EMA: 0.9999
Dropout: 0.0
Training hardware: 128 TPU-v4 chips
Training steps: 800000
Joint training independent images per video: 7
Sampling timesteps: 256
Sampling log-variance interpolation: $\gamma = 0.3$
Prediction target: $\boldsymbol{\epsilon}$