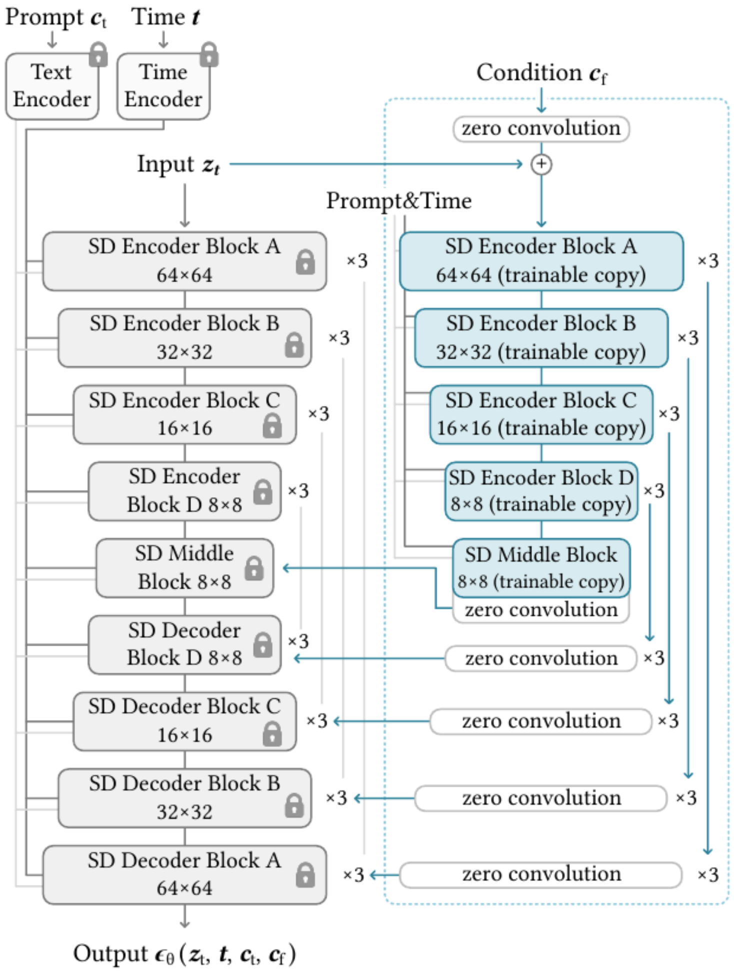
Control Net for Image and Video Generation

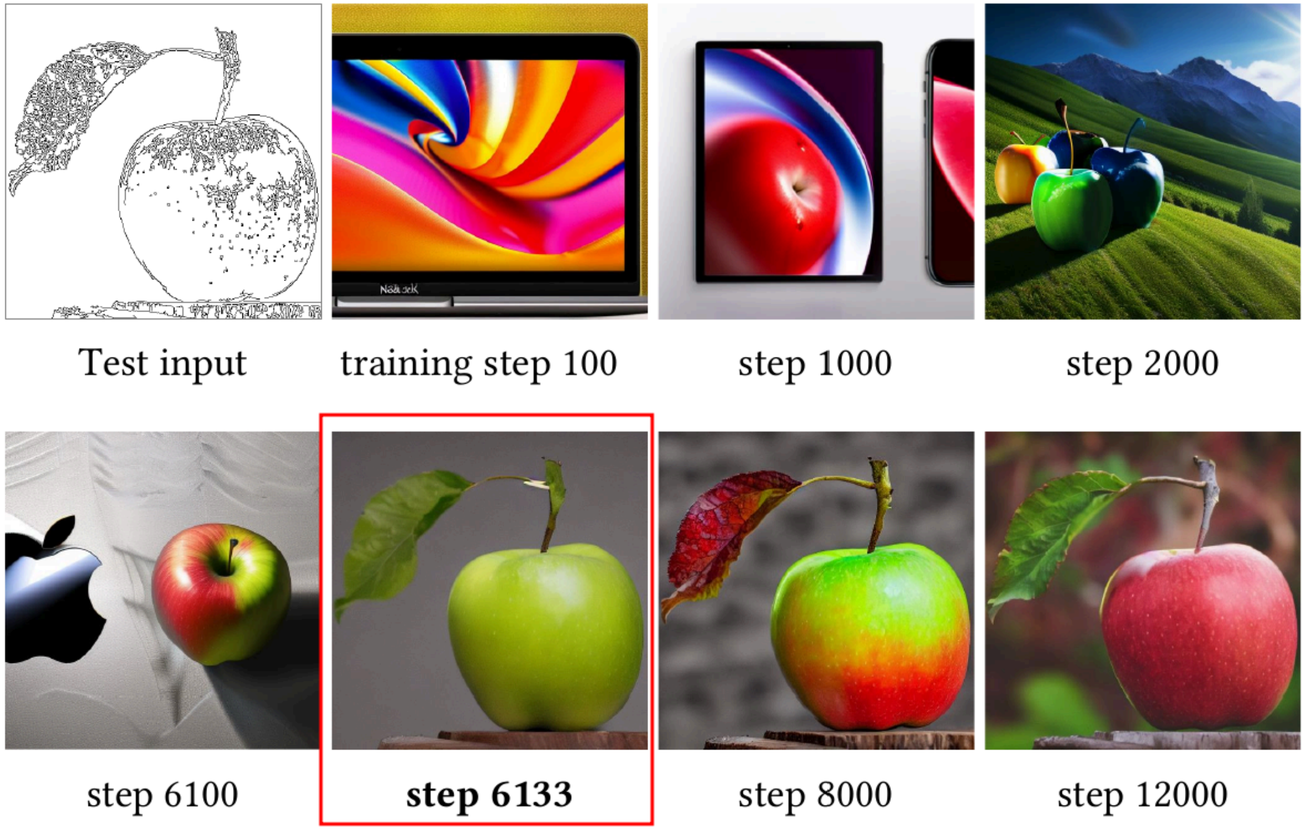
[1] "Adding Conditional Control to Text-to-Image Diffusion Models" (Zhang et al, ICCV 2023)
[2] "Control-A-Video: Controllable Text-to-Video Diffusion Models with Motion Prior and Reward Feedback Learning" (Chen et al, 2023)

Overall Architecture



(a) Stable Diffusion (b) ControlNet

Sudden Convergence Phenomenon



Test input

training step 100

step 1000

step 2000

step 6100

step 6133

step 8000

step 12000

The model does **NOT gradually learn the control conditions** but abruptly succeeds in following the input conditioning image; usually in less than 10K optimization steps.

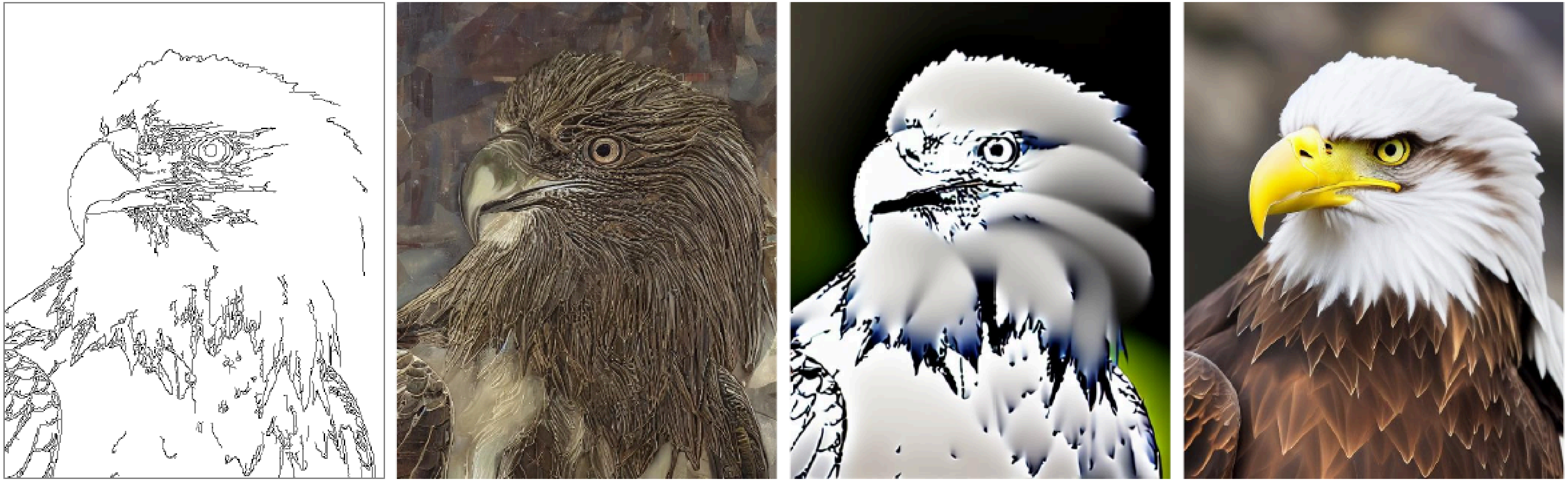
Classifier-Free Guidance

$$\epsilon_{\mathrm{prd}} = \epsilon_{\mathrm{uc}} + \beta_{\mathrm{cfg}}(\epsilon_{\mathrm{c}} - \epsilon_{\mathrm{uc}})$$

where

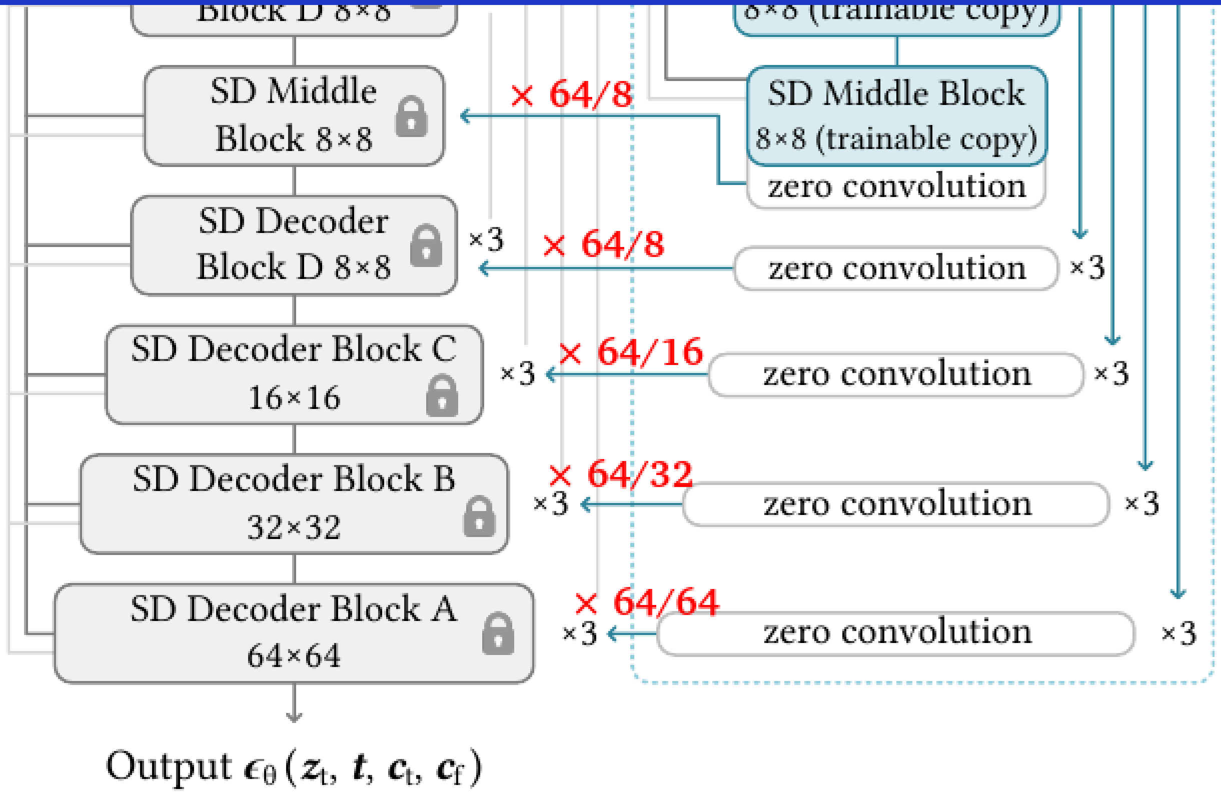
- ϵ_{prd} — the model's final output
- β_{cfg} — the weight of guidance
- $\epsilon_{\mathrm{uc}} = \mathtt{UnetWithoutControlNet}(\mathbf{z}_t, t, \mathtt{""};\ \theta)$ — unconditional output
- $\epsilon_{\mathrm{c}} = \mathtt{UnetWithConrolNet}(\mathbf{z}_t, t, \mathbf{c}_t, \mathbf{c}_f; \theta, \phi)$ — conditional output
- $\mathbf{z}_t$ — noisy latent image
- $\mathbf{c}_t$ — text prompt
- $\mathbf{c}_f$ — condition image
- θ — pre-trained diffusion model's weights
- ϕ — control-net's weight

Classifier-Free Guidance Resolution Weighting



(a) Input Canny map (b) W/o CFG (c) W/o CFG-RW (d) Full (w/o prompt)

Classifier-Free Guidance Resolution Weighting

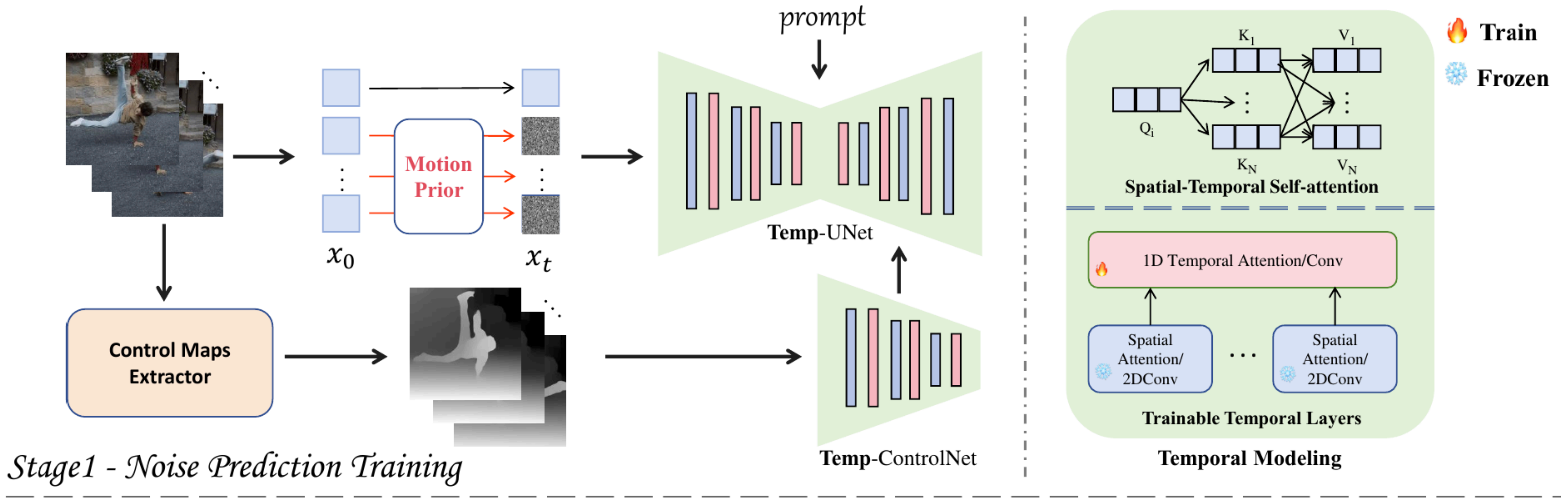


(a) Stable Diffusion

(b) ControlNet

By multiplying a weight w_i to each connection between Stable Diffusion and ControlNet according to the resolution of each block $w_i = 64/h_i$, where h_i is the size of ith block, *e.g.*, $h_1 = 8, h_2 = 16, \ldots, h_{13} = 64$, we can achieve the better generation result.

ControlNet for VideoGeneration



prompt
Motion Prior
x_0
x_t
Temp-UNet
Control Maps Extractor
Temp-ControlNet
Stage1 - Noise Prediction Training
K_1
V_1
Q_i
K_N
V_N
Spatial-Temporal Self-attention
1D Temporal Attention/Conv
Spatial Attention/ 2DConv
Spatial Attention/ 2DConv
Trainable Temporal Layers
Temporal Modeling
🔥 Train
❄ Frozen

"Control-A-Video: Controllable Text-to-Video Diffusion Models with Motion Prior and Reward Feedback Learning" (Chen et al, 2023)