Stage 1: Visual Dynamics Compression

Appearance Reconstruction

Disentangled Latent Dynamic Model

x_0 x_1 x_2 x_T

Stage 2: Next Token Prediction

Autoregressive Transformer

Inference Stage: Task completion by Codes Generation

Step 4 Step 5 Step 6

Step 1 Step 2 Step 3

Appearance Reconstruction

Autoregressive Transformer

USER: Please fold A Paper Airplane

The captured Dynamics

Fold in Half

Move top-right

Move top-left

Move bottom-right