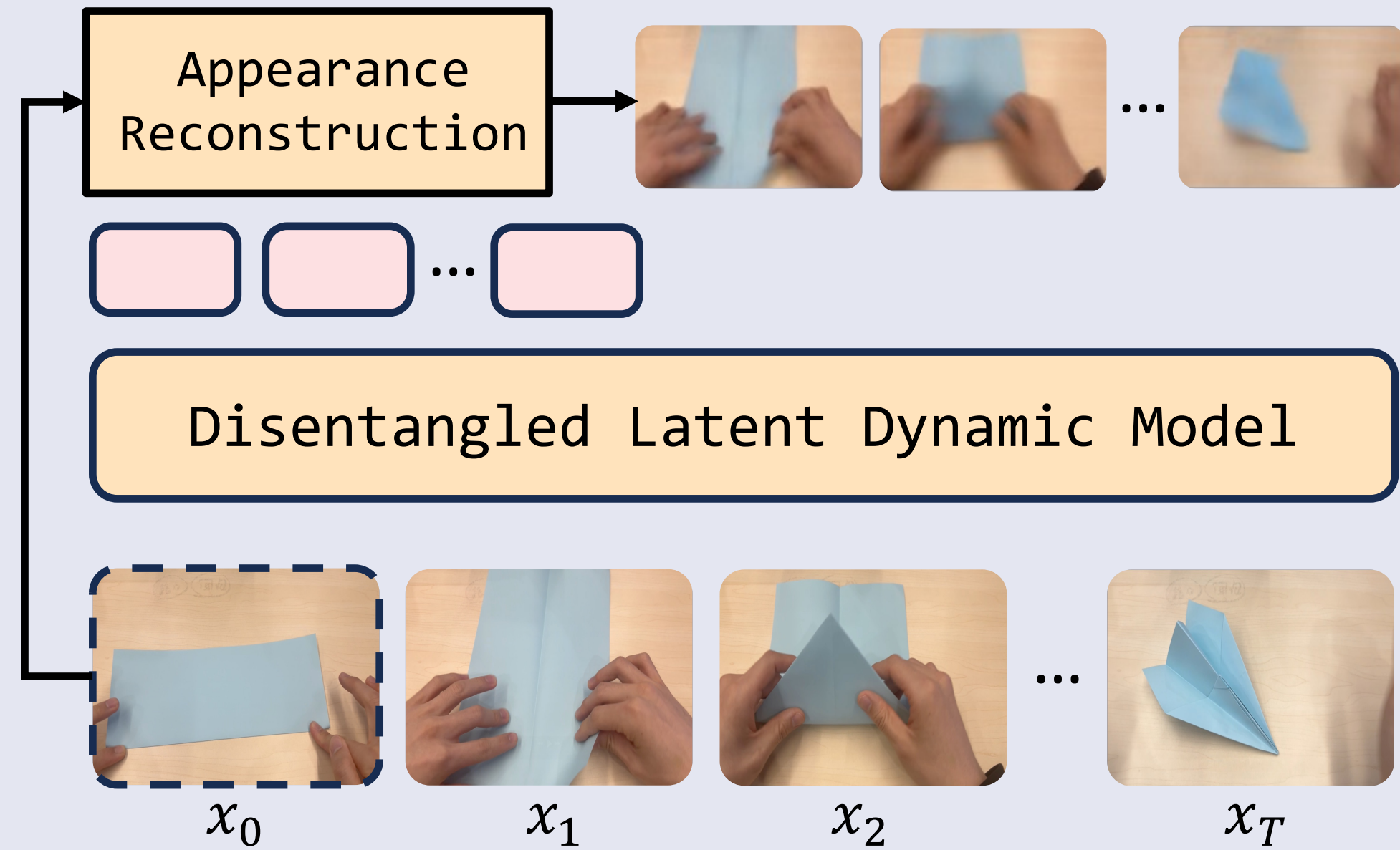
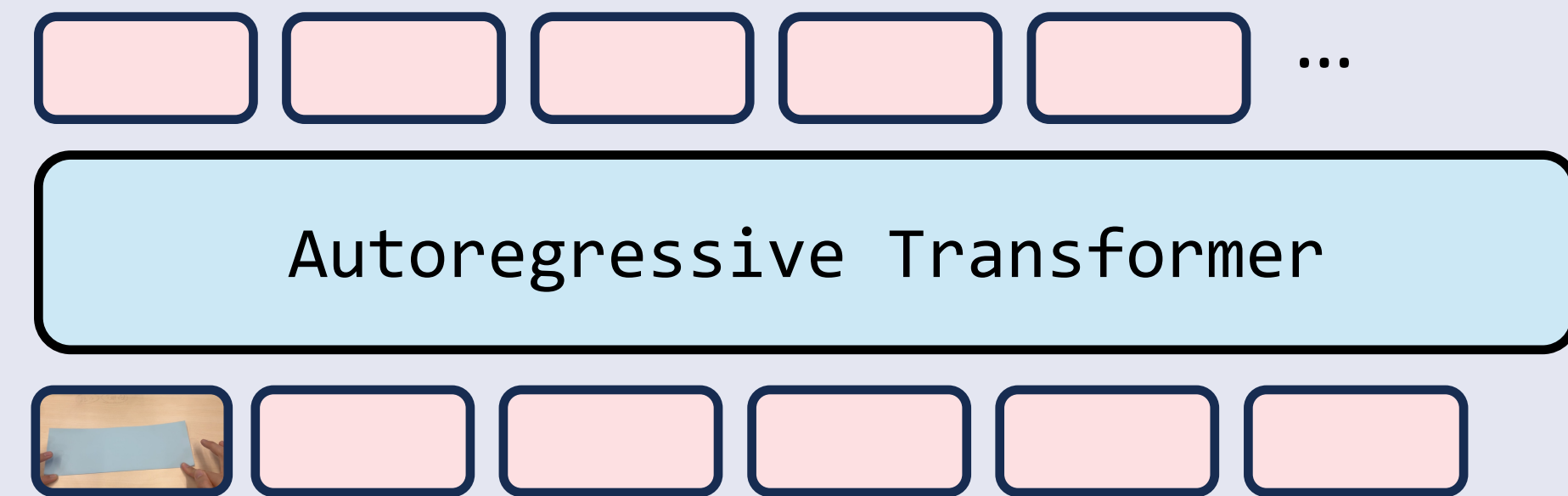


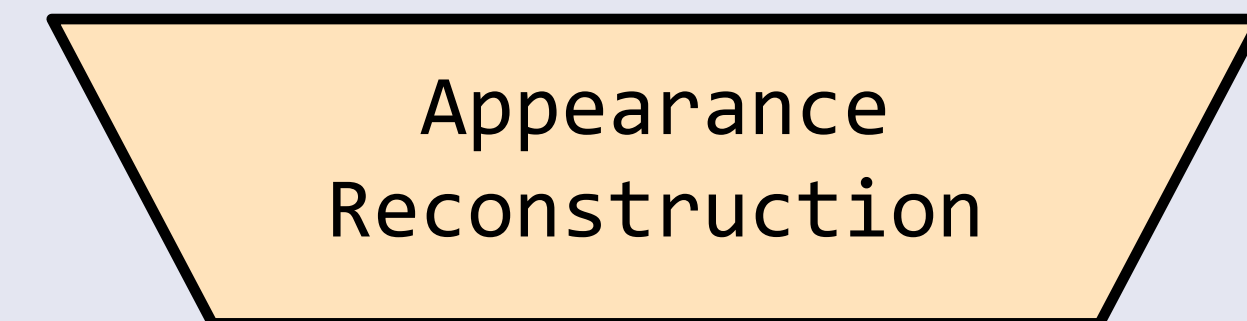
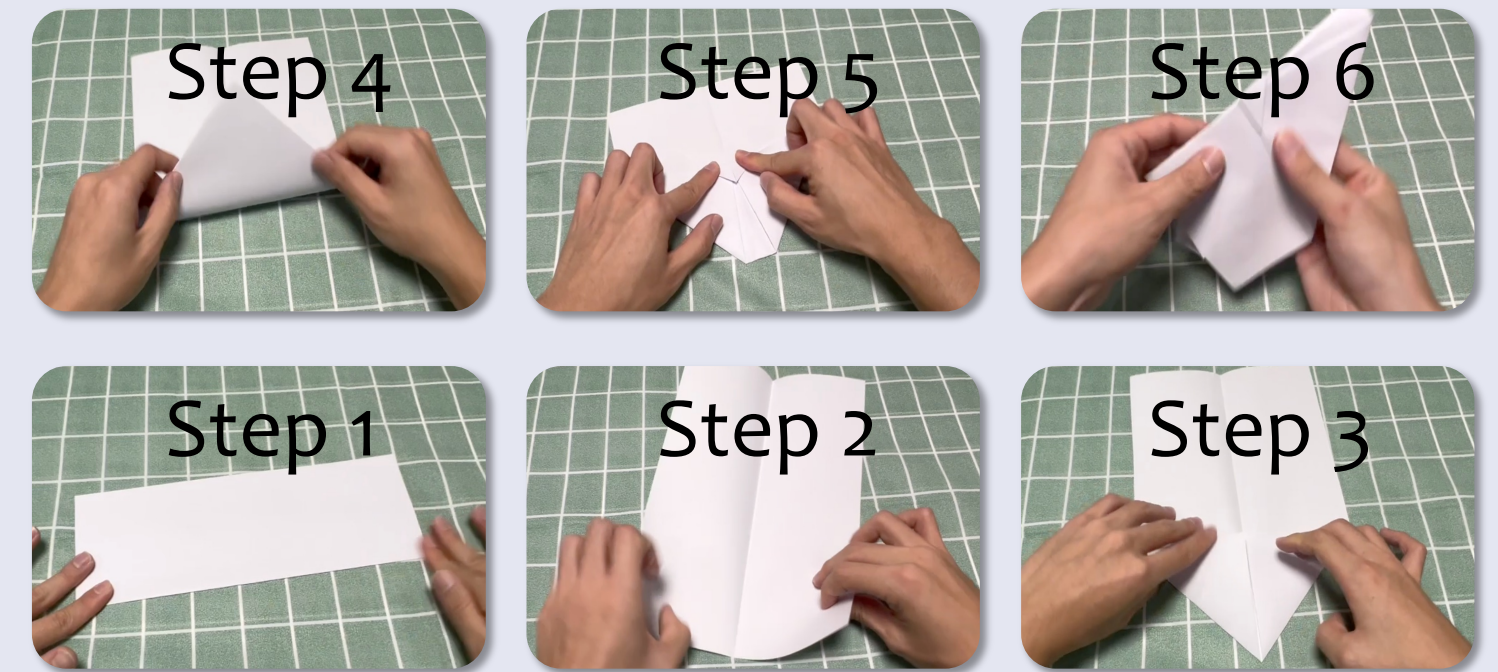
Stage 1: Visual Dynamics Compression



Stage 2: Next Token Prediction

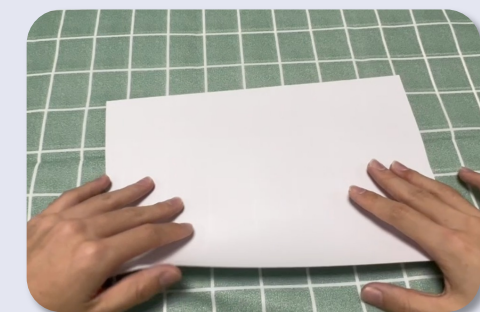


Inference Stage: Task completion by Codes Generation



Autoregressive Transformer

USER: Please fold
A Paper Airplane



The captured Dynamics

