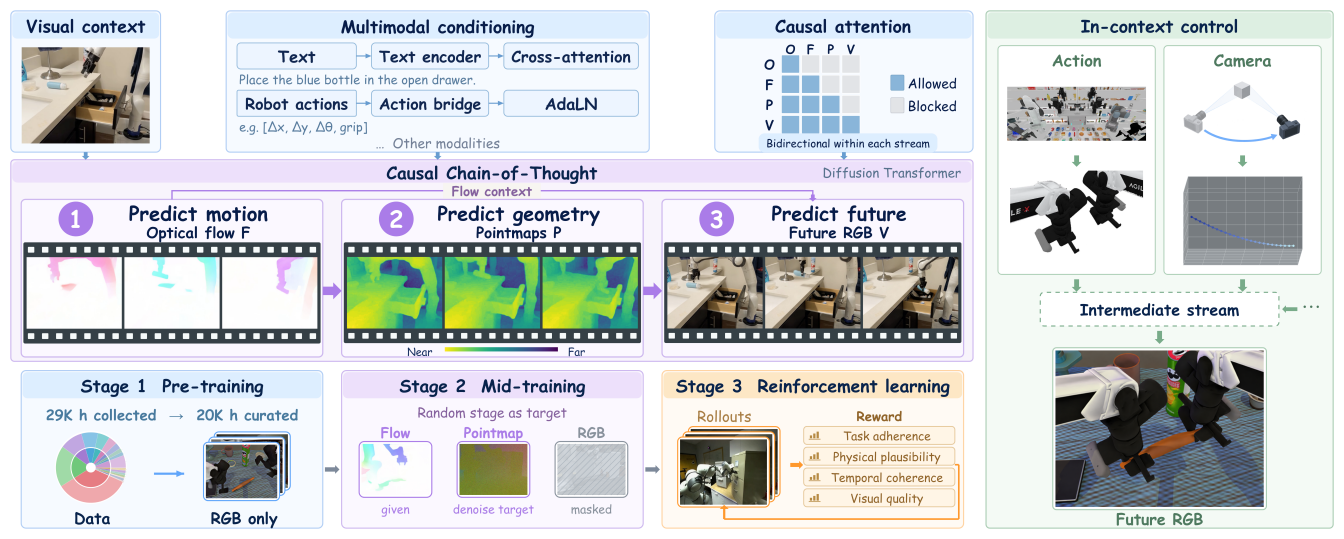
Visual context
Multimodal conditioning
Text
Text encoder
Cross-attention
Place the blue bottle in the open drawer.
Robot actions
Action bridge
AdaLN
e.g. [Δx, Δy, Δθ, grip]
... Other modalities
Causal attention
O F P V
Allowed
Blocked
Bidirectional within each stream
In-context control
Action
Camera
Causal Chain-of-Thought
Diffusion Transformer
Flow context
1 Predict motion
Optical flow F
2 Predict geometry
Pointmaps P
Near
Far
3 Predict future
Future RGB V
Intermediate stream
Future RGB
Stage 1 Pre-training
29K h collected → 20K h curated
Data
RGB only
Stage 2 Mid-training
Random stage as target
Flow
Pointmap
RGB
given
denoise target
masked
Stage 3 Reinforcement learning
Rollouts
Reward
Task adherence
Physical plausibility
Temporal coherence
Visual quality