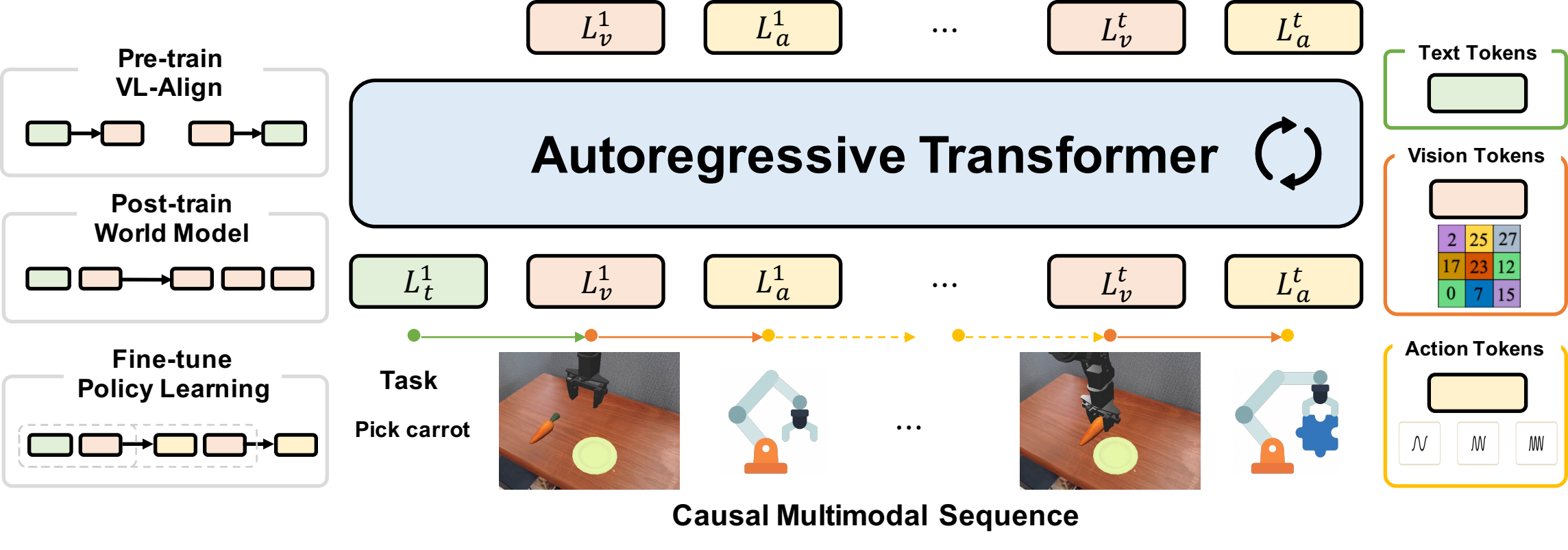
Pre-train
VL-Align
Post-train
World Model
Fine-tune
Policy Learning
L_v^1
L_a^1
...
L_v^t
L_a^t
Autoregressive Transformer
L_t^1
L_v^1
L_a^1
...
L_v^t
L_a^t
Task
Pick carrot
...
Causal Multimodal Sequence
Text Tokens
Vision Tokens
2 25 27
17 23 12
0 7 15
Action Tokens