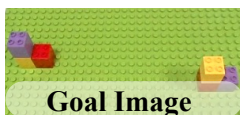
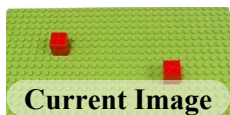
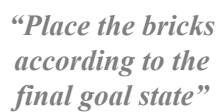


(a) ManualVLA



Text Tokenizer



Vision Tokenizer

Planning Expert

QKV

Cross-task Shared Attention

FFN



Text Detokenizer

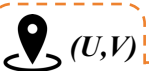
Vision Detokenizer

Implicit CoT

Manual

*“Yellow blocks
Assemble”*

Text



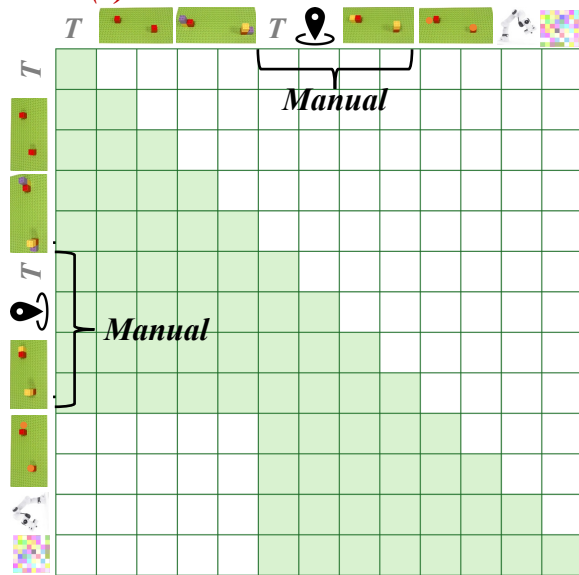
Position



Subgoal Image

Explicit CoT

(b) Cross-task Attention Mask



(c) Training stage

Stage 1:  Planning Expert  Action Expert

Stage 2:  **Planning Expert**  **Action Expert**

Stage 3: ⚡ Planning Expert ⚡ Action Expert

❄️ : Frozen 🔥 : Pretrain ⚡ : Fine-tune

Actions

 a_t a_{t+1}

■ ■ ■ ■ ■

 a_{t+H} 

Execution Examples

