Adaptive CoT

<think_off> ⚡

<think_on> 💡
<think>
I'm in the hallway, and I can see some paintings ...
I decide to ...
</think>
<summary>
I am in a bathroom containing a white toilet in the center...
</summary>

Trivial scene

Complex scene

Large Language Model

Instruction token

T Temporal-aware token

Visual token

Linguistic memory token

◊ <think_off>/<think_on>

CoT token

Action token

The latest step for which you generated a Chain-of-Thought is: 16. The memory of the entire thinking process is: 1ᵗʰ : The current environment contains a large white wardrobe straight ahead with two vertical handles...

T_0 T_1 T_2 ...... T_{t-1} T_t

Search for the brown chair

Track the man in blue top and dark pants

Seek for the object in the <image_goal>

Multimodal Projector

Dynamic Sampling & Pooling

Vision Encoder

Action Model

Visual-Assisted Linguistic Memory

Multimodal Instruction

Input Video Stream

time