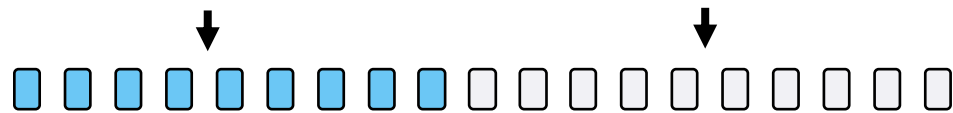
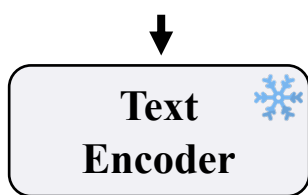
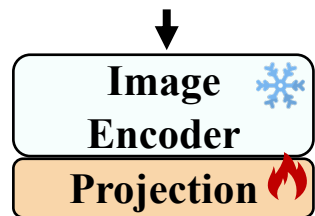


Give a brief description of the given **observation** / **subsequent keyframes** / **short demonstrations**.



Multi-modal LLM



<Scene Description> :

On the table, there are two bigger green blocks, one smaller red bowls on the left bottom corner ...

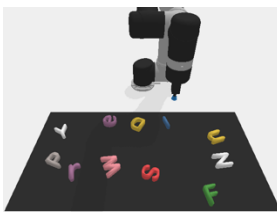
<Action Recognition>:

Move the smaller blue block into the black bowl.

<Video Understanding>:

The given demonstration is to first put smaller blue block onto the smaller red block and then ...

Observation



Task instruction

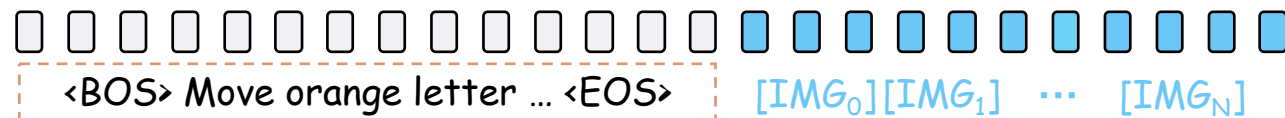
Spell a word using letters on the table for the name of top CS conference.

Prompt

Please give a brief stepwise instruction to do next and the corresponding goal image.



Multi-modal LLM



Output Projection



L_{Reason}

Move orange letter N to the bottom left area

Diffusion Model



$L_{Imagine}$



Multi-modal LLM

$L_{Alignment}$

Action Recognition