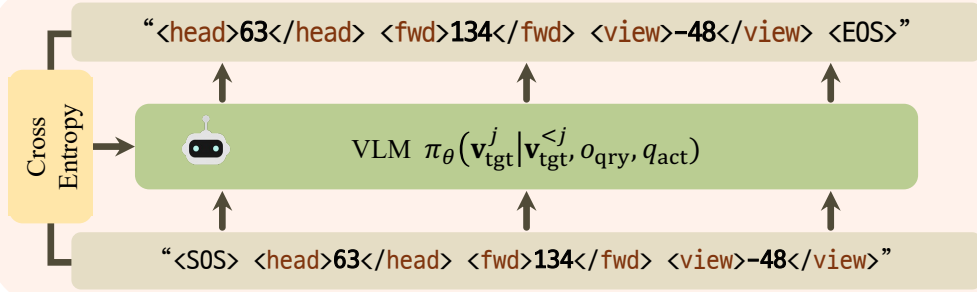
Input

q_{act} ="Given an image and a question about the scene, **predict the action parameters** to better answer the question: (1) heading rotation, (2) forward translation, and (3) view rotation.

DO NOT answer the question; ONLY predict the action parameters.

Question q: Is a book present on the dresser?"

SFT



RL