CoT-VLA: Visual Chain-of-Thought Reasoning for Vision-Language-Action Models

Qingqing Zhao[1,2,*] Yao Lu[1] Moo Jin Kim[2] Zipeng Fu[2]
Zhuoyang Zhang[3] Yecheng Wu[1,3] Zhaoshuo Li[1] Qianli Ma[1] Song Han[1,3] Chelsea Finn[2]
Ankur Handa[1] Ming-Yu Liu Donglai Xiang[1†] Gordon Wetzstein[2†] Tsung-Yi Lin[1†]

[1]NVIDIA [2]Stanford University [3]MIT

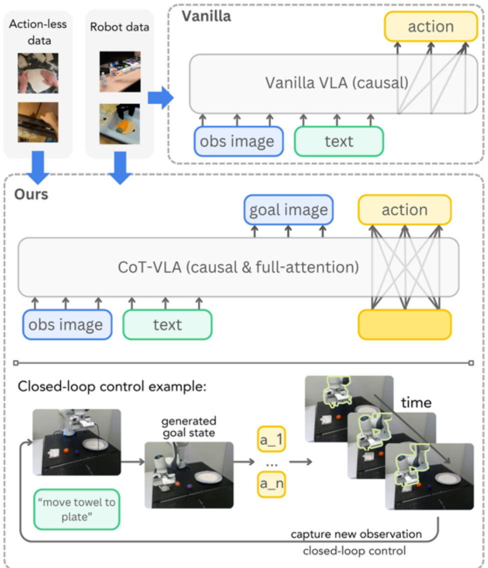

Figure 1. **Comparison between vanilla VLA and CoT-VLA frameworks**. Prior VLA models (top) directly predict robot actions from task inputs without explicit reasoning steps and only use action-annotated robot demonstration data for training. Unlike vanilla VLAs, CoT-VLA (bottom) can also leverage action-less datasets like EPIC-KITCHEN-100 [27] to enhance subgoal image generation ability, unlocking the potential of using abundant unlabeled video data to improve VLA's visual reasoning capability. CoT-VLA first generates a subgoal image as an intermediate reasoning step, and then generate a short action sequence to achieve the subgoal. We outline the robot arm for better visualization.

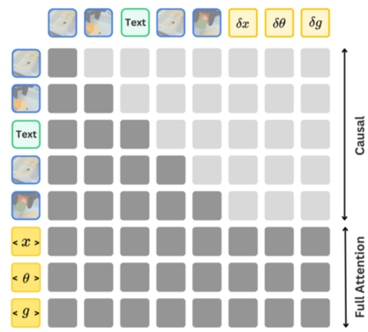

Figure 3. **Hybrid attention mechanism in CoT-VLA**. We use causal attention for image or text generation and full attention for action generation. $[x]$, $[\theta]$ and $[g]$ are special tokens for parallel decoding of actions.

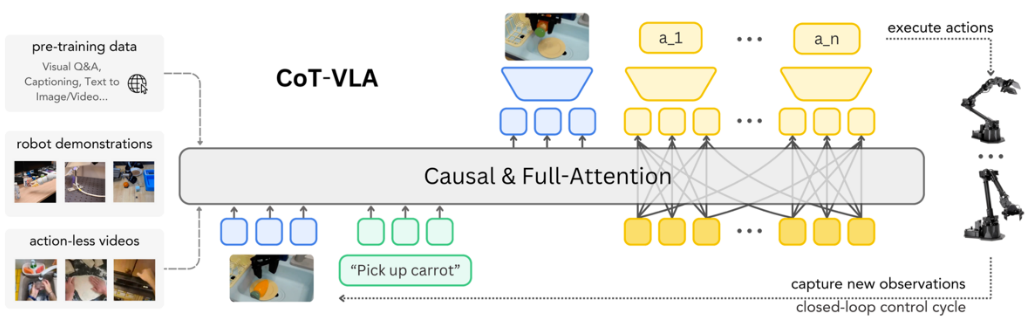

Figure 2. **Overview of CoT-VLA framework**. We build our model on VILA-U [67], a generative multimodal model pretrained on interleaved text–image data. The base model then trains on robot demonstrations [48] and action-less videos [20, 27]. During deployment, given a visual observation and a text instruction, the model performs visual chain-of-thought reasoning by generating a subgoal image (upper blue) with causal attention. It then generates a short action sequence with full attention ($\mathbf{a}_1 \cdots \mathbf{a}_n$) for robot execution. The system operates in a closed-loop control manner by capturing new observations after executing predicted action sequences.