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

Supplementary Material

6. Implementation Details

6.1. Data Details

We select part of the Open X-Embodiment dataset [48] as our robot demonstration pre-training data, and Something2Something [20], and EPIC-KITCHEN-100 [27] as our action-less video data. The u_l and u_u is upper bound and lower bound for predicted subgoal horizon. We manually set those number for each dataset.

Dataset	**Weight**	u_l	u_u
Bridge [16, 60]	24.14%	5	10
RT-1 [3]	6.90%	5	10
TOTO [81]	10.34%	20	24
VIOLA [83]	10.34%	15	20
RoboTurk [42]	10.34%	1	2
Jaco Play [43, 51]	10.34%	10	15
Berkeley Autolab UR5 [8]	10.34%	5	10
Berkeley Fanuc Manipulation [82]	10.34%	10	15
Something2Something [20]	3.45%	5	7
EPIC-KITCHEN-100 [27]	3.45%	5	7

Table 4. Dataset Weights and Hyperparameters

6.2. Hyperparameters

In this section, we list the important hyperparameters for our model pre-training and pose-training stage.

Hyperparameter	**Pre-training**
Learning Rate	1e-4
LR Scheduler	Cosine decay
Global Batch Size	2048
Image Resolution	256×256
Action Token Size	10
Epoch	10

Table 5. Hyperparameters for pre-training

For fine-tuning on LIBERO [37] and Franka-Tabletop [29] experiments, we fine-tune the model (LLM backbone, projector, depth transformer) with constant learning rate 1e-5 for 150 epochs.

6.3. Training

We perform training on 12 A100 GPU nodes with 8 GPUs each. The pre-training with data mixture in 6.1 takes 11K A100 GPU hours in total. The training cost for LIBERO and Franka-Tabletop fine-tuning is done on a single A100 GPU node for 10-24 hours depends on the dataset size.

7. Example Rollouts

We refer user to our project website (zipped in supplementary material) for more example rollouts.