Language-Guided Manipulation

Group 11:

Bryan Choo, Aayush Dulal, Jessi King, Zhiyu Yang

Motivation for Robot Cleanup

- Household and office cleanup behaviors (e.g., object sorting, placing items) are crucial for future general-purpose robots.

- Many everyday tasks involve:
- High within-class variation
- Partial occlusions
- Millimeter-level placement precision

- Goal: Achieve sample-efficient learning using a compact VLA model (SmolVLA) fine-tuned on a small real-world dataset.

Group Task

Our task:
Teach a robot to pick and place colored blocks (red/green/yellow) into a bin using imitation learning.

High-level theme:
This task is a micro-version of real household cleanup behaviors — sorting objects, organizing clutter, and placing items where they belong.

System Overview:

- Hardware: SO-101 6-DOF research arm
- Cameras: wrist-mounted first-person camera
- Policy: SmolVLA (vision–language–action)
- Training: Behavioral Cloning on small curated dataset
- Deployment: Real-time control via continuous action output

Model Training

Training Overview:

Goal: Fine-tune SmolVLA to perform a single robot manipulation task

Approach: Offline imitation learning on expert demonstrations

Model: SmolVLA (Vision-Language-Action transformer)

Training type: Fine-tuning using supervised learning on trajectories

Framework: LeRobot + Hugging Face

Model Training

Dataset Summary:

Link: https://huggingface.co/datasets/HenryZhang/Group11_data_1763075740.884942

- 90 episodes
- 45,000+ frames
- FPS: 30
- 1 camera view (Top-down) — 640×480
- Robot: so101_follower
- Actions: 6-dim joint positions
- Obs: images + 6 joint states
- Language instruction

Model Training

Training Data Inputs and Outputs:

Inputs:

- Front camera RGB (640×480)
- Robot state (6 joints)
- Task instruction: Language instruction → "Grasp a lego block and put it in the bin."

Outputs (labels):

- 6-D robot action: [shoulder_pan, shoulder_lift, elbow_flex, wrist_flex, wrist_roll, gripper]

Model Training

Training Method:

Objective: Predict the expert action given image + robot state + language instruction

Loss: Mean Squared Error (MSE) between predicted and expert actions

Batching: Sequence chunks (≈1000 frames per chunk)

Hardware: (A100 on colab)

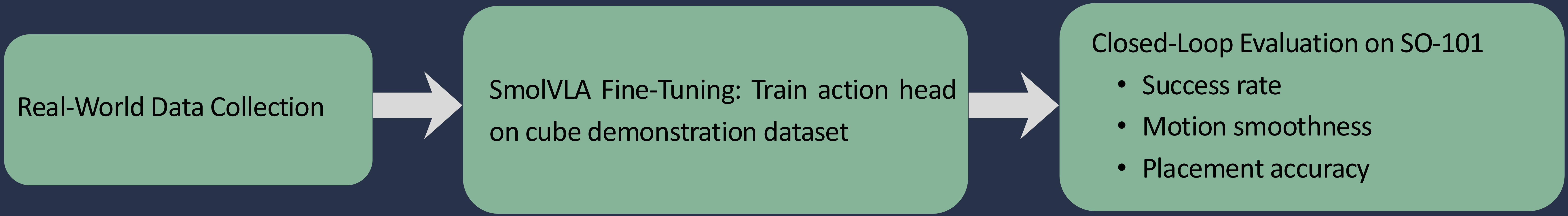
Duration: 70 minutes

Model Training

Config & Run:

```
!python src/lerobot/scripts/lerobot_train.py \
--policy.type=smolvla \
--policy.pretrained_path={CONFIG['policy_path']} \
--policy.repo_id=smolvla_finetuned \
--dataset.repo_id={CONFIG['dataset_repo_id']} \
--batch_size=4 \
--steps=20000 \
--optimizer.lr=5e-5 \
--save_freq=5000 \
--eval_freq=5000
```

Workflow:



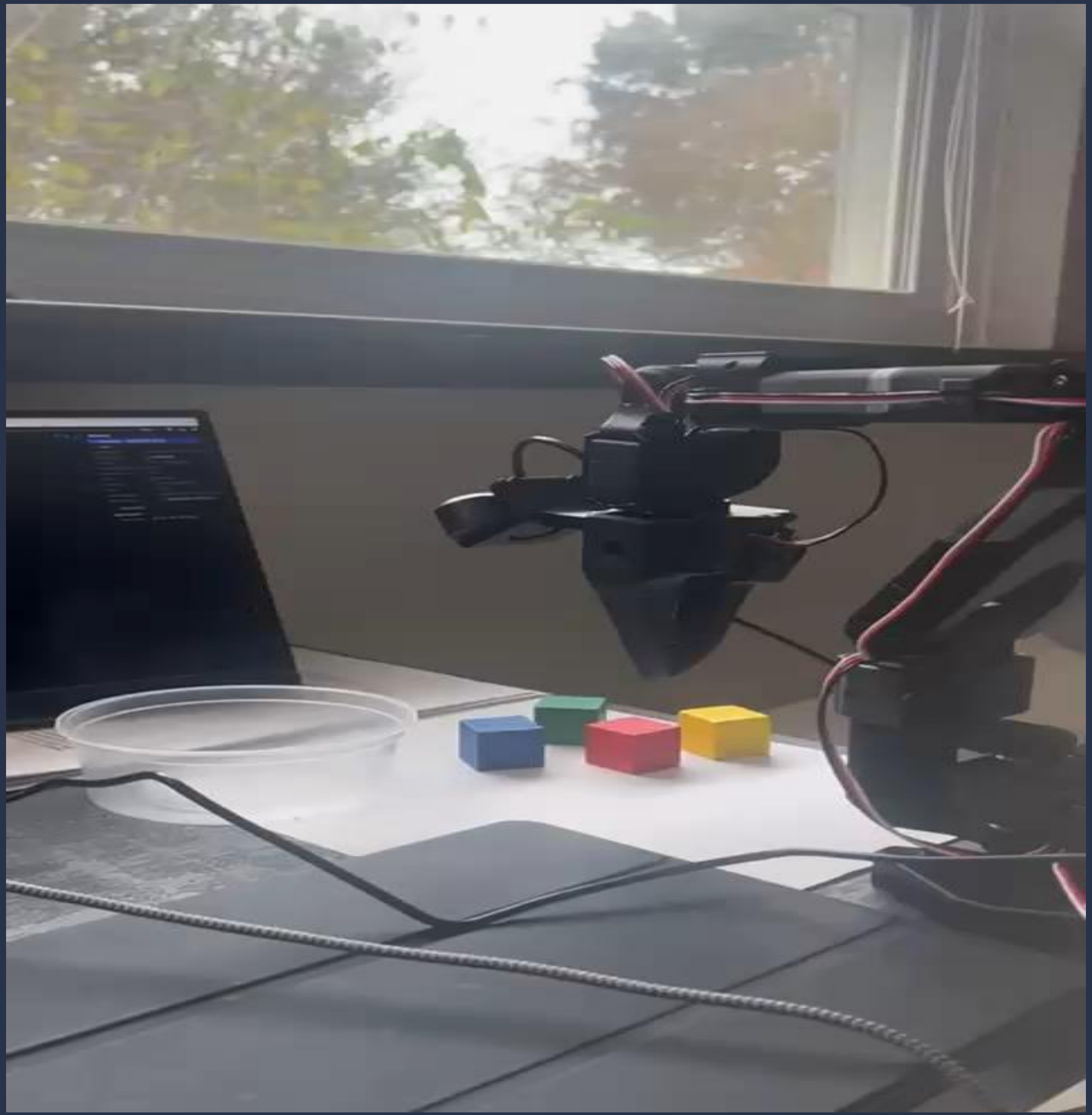
Real-World Data Collection

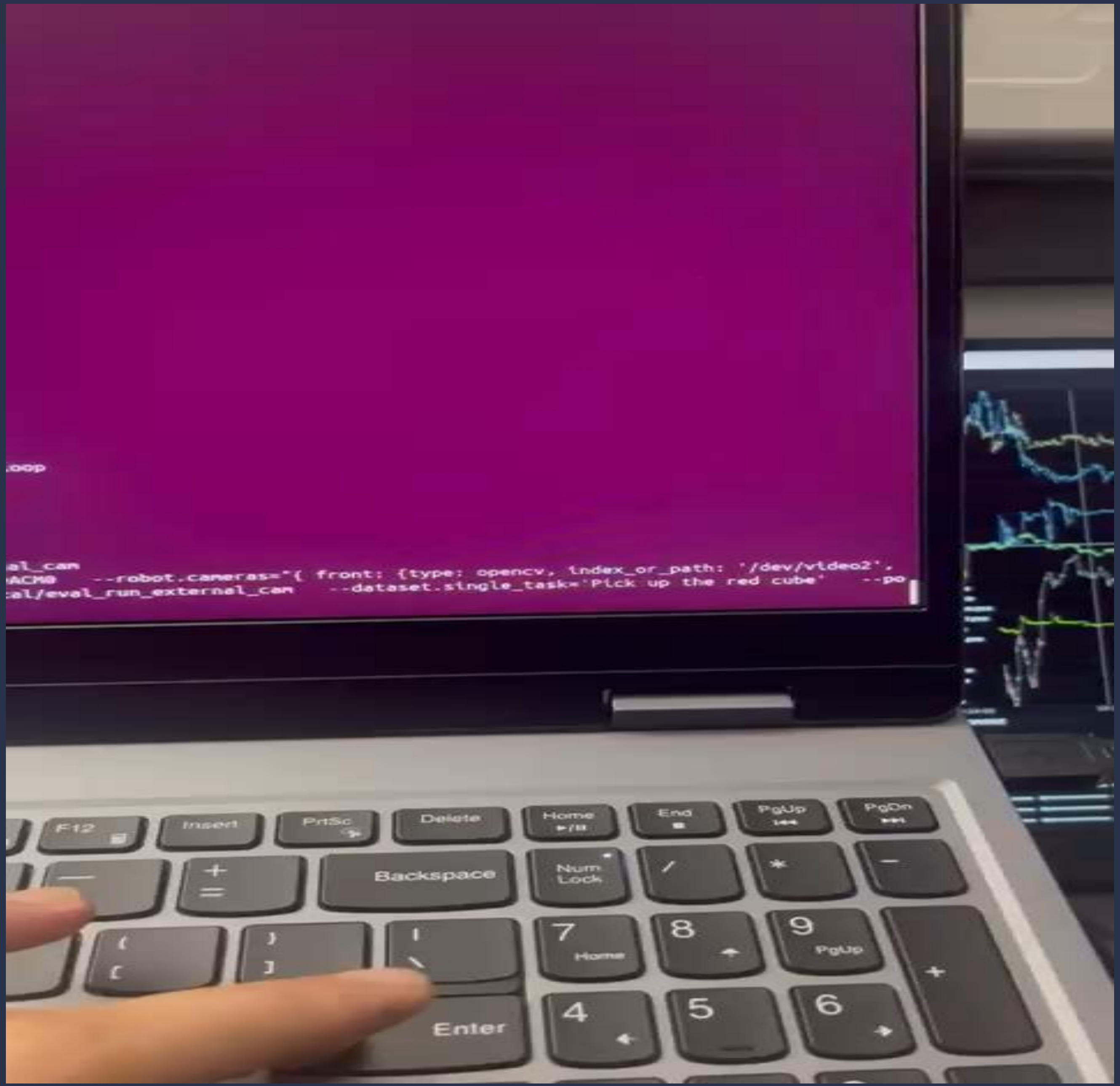
SmolVLA Fine-Tuning: Train action head on cube demonstration dataset

Closed-Loop Evaluation on SO-101

- Success rate
- Motion smoothness
- Placement accuracy

Evaluation





```
.oop

al_cam
ACM0    --robot.cameras="{ front: {type: opencv, index_or_path: '/dev/video2',
cal/eval_run_external_cam    --dataset.single_task='Pick up the red cube'    --po
```

- Even when other objects are present the model picks up the red cube based on our instruction.

- We trained the model on videos that were preprocessed to ensure the cube would be in frame.
- For objects the model has not seen, the S0-101 will ignore.

Challenges

The evaluation setup currently depends heavily on the first-person camera attached to the gripper.

Without a stable overhead view:

- The cube may leave the field of view
- Visual drift and occlusions degrade model predictions

This creates instability in:

- Grasp detection
- Placement alignment

Additional challenges:

- Partial occlusions and lighting variation
- Action jitter from VLA outputs
- Limited real-world demonstrations (<100)

Citations

[1] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, and S. Levine, “Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation,” 2018. [Online]. Available: https://arxiv.org/abs/1806.10293

[2] S. Levine, P. Pastor, A. Krizhevsky, and D. Quillen, “Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection,” 2016. [Online]. Available: https://arxiv.org/abs/1603.02199

[3] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, “Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics,” 2017. [Online]. Available: https://arxiv.org/abs/1703.09312

[4] L. Pinto and A. Gupta, “Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours,” 2015. [Online]. Available: https://arxiv.org/abs/1509.06825

[5] M. Shukor, D. Aubakirova, F. Capuano, P. Kooijmans, S. Palma, A. Zouitine, M. Aractingi, C. Pascal, M. Russi, A. Marafioti, S. Alibert, M. Cord, T.
Wolf, and R. Cad` ene, “Smolvla: A vision-language-action model for affordable and efficient

Thank you!