Week 3

Shriyadita Medavaram



Project Overview

Efficient and Robust VLA Models for Real Time Robotic Manipulation

- Problem: Current open source VLA models are
  - Slow
  - Computationally Expensive
- Goal:
  - Smaller faster VLA model for robots
  - Can run easily on smaller computers
  - Can deal with interference

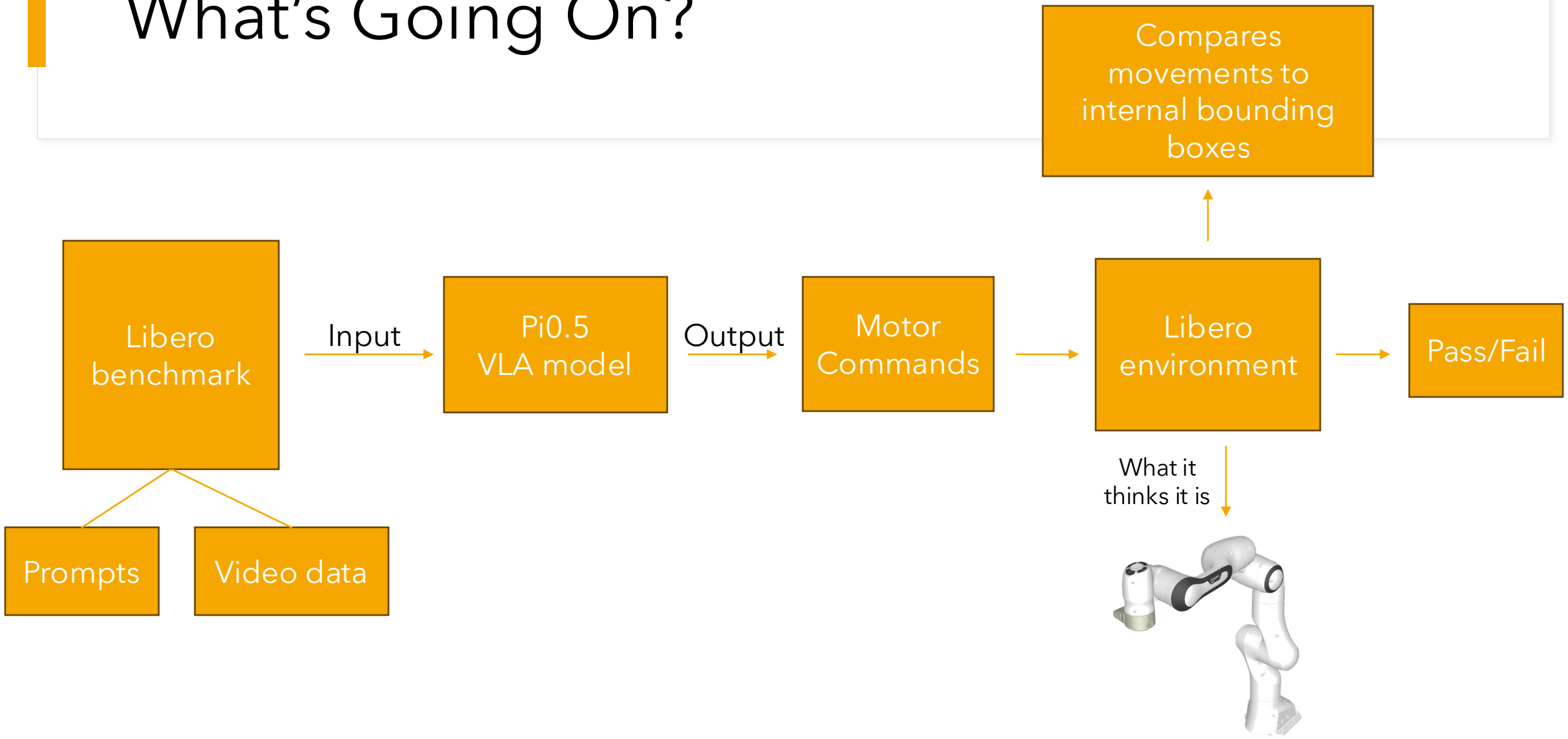
Schedule

- Week 1: Load Pi-0.5 and Libero on clusters and evaluate
- **Week 2-3: Look into Pi-0.5 model and find bottleneck (Run profiles)**
- Week 4-6: Run compression experiments
- Week 7: Action-Head Distillation, Full Libero evaluation
- Week 8: Make poster and release code

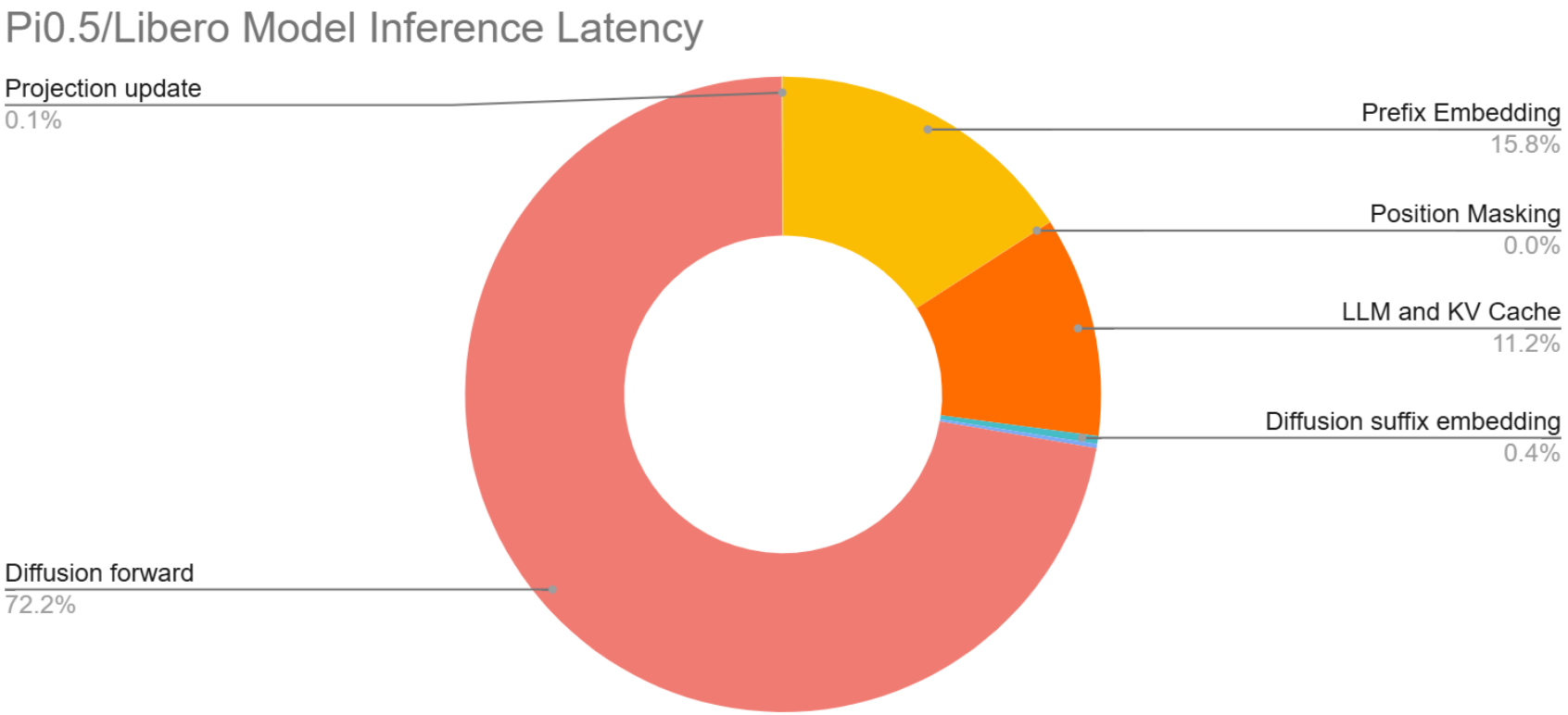
Running Inference with Libero

```
NFO:root:
Task: pick up the butter and place it in the basket
INFO:root:Starting episode 1...
INFO:root:Success: True
INFO:root:# episodes completed so far: 31
INFO:root:# successes: 31 (100.0%)

NFO:root:
Task: pick up the butter and place it in the basket
INFO:root:Starting episode 2...
INFO:root:Success: False
INFO:root:# episodes completed so far: 32
INFO:root:# successes: 31 (96.9%)

NFO:root:
Task: pick up the butter and place it in the basket
INFO:root:Starting episode 3...
INFO:root:Success: True
INFO:root:# episodes completed so far: 33
INFO:root:# successes: 32 (97.0%)
```

What's Going On?



Latency with Libero



Latency with Libero – Table Legend

Sections	Meaning
Preprocess Observations	Observation normalization (Not pictured)
Noise Initializations	Initializes noise for diffusion (Not Pictured)
Prefix Embedding	Encodes image observations and prompt into tokens
Position Masking	Builds attention masks and token positions for the prefix
LLM and KV Cache	Runs prompt and image through LLM and builds KV cache
Diffusion suffix embedding	Embeds the current noisy action chunk and diffusion timestep during the action refinement loop
Diffusion position masking	Builds masks/positions for the diffusion/action tokens (Not Pictured)
Diffusion forward	Repeated LLM/action-expert forward pass inside the diffusion loop. Model repeatedly refines noisy actions into the final action chunk.
Projection update	Projects model outputs back into action space and updates the action trajectory

Real-Time Manipulation

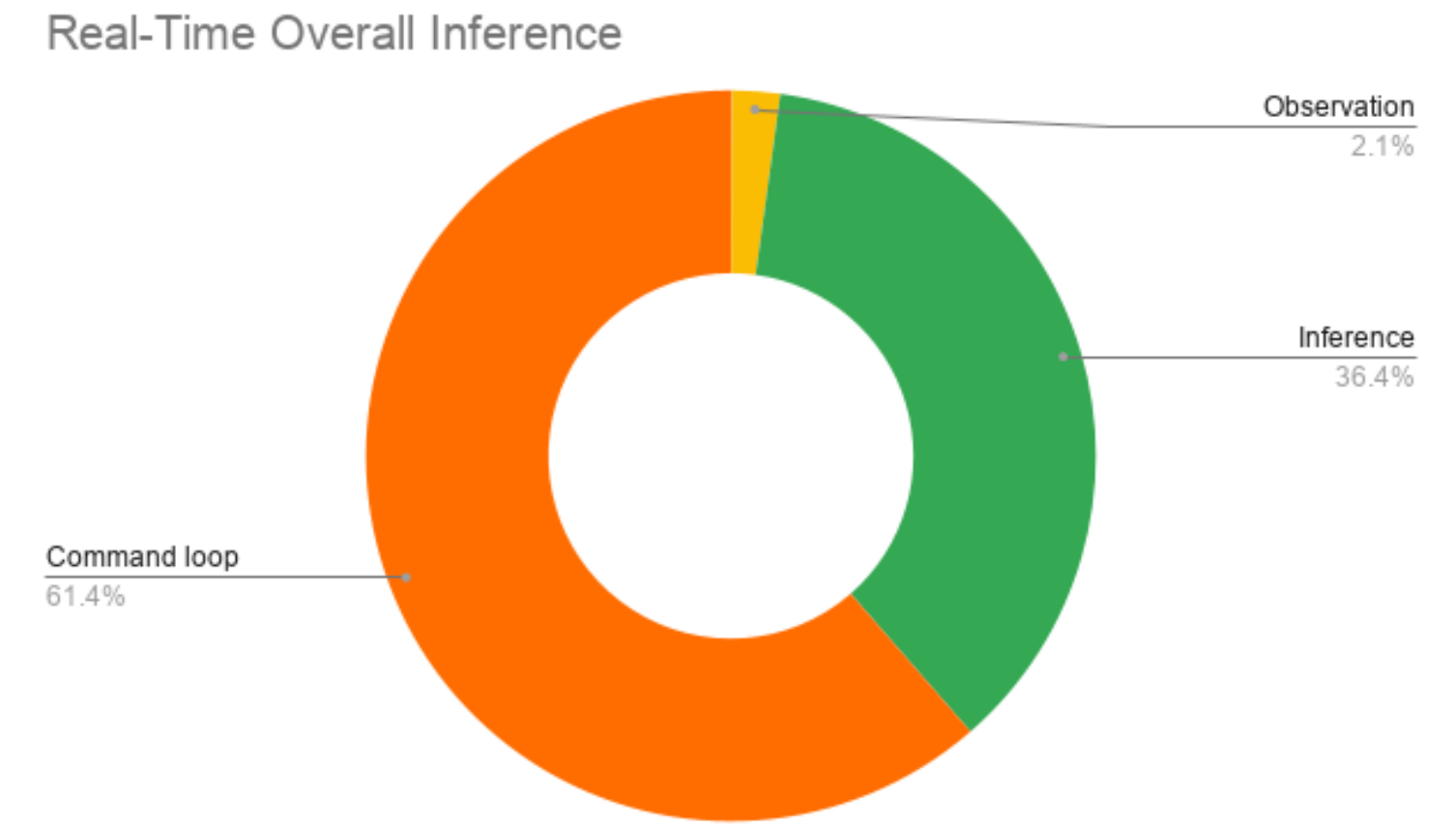


Why is it like that?

- Very small allowance to move
  - 0.04 radians of movement per step
- Bad feedback on joints 2 and 3
  - Strange values
  - Unsafe for arm
  - Disabled



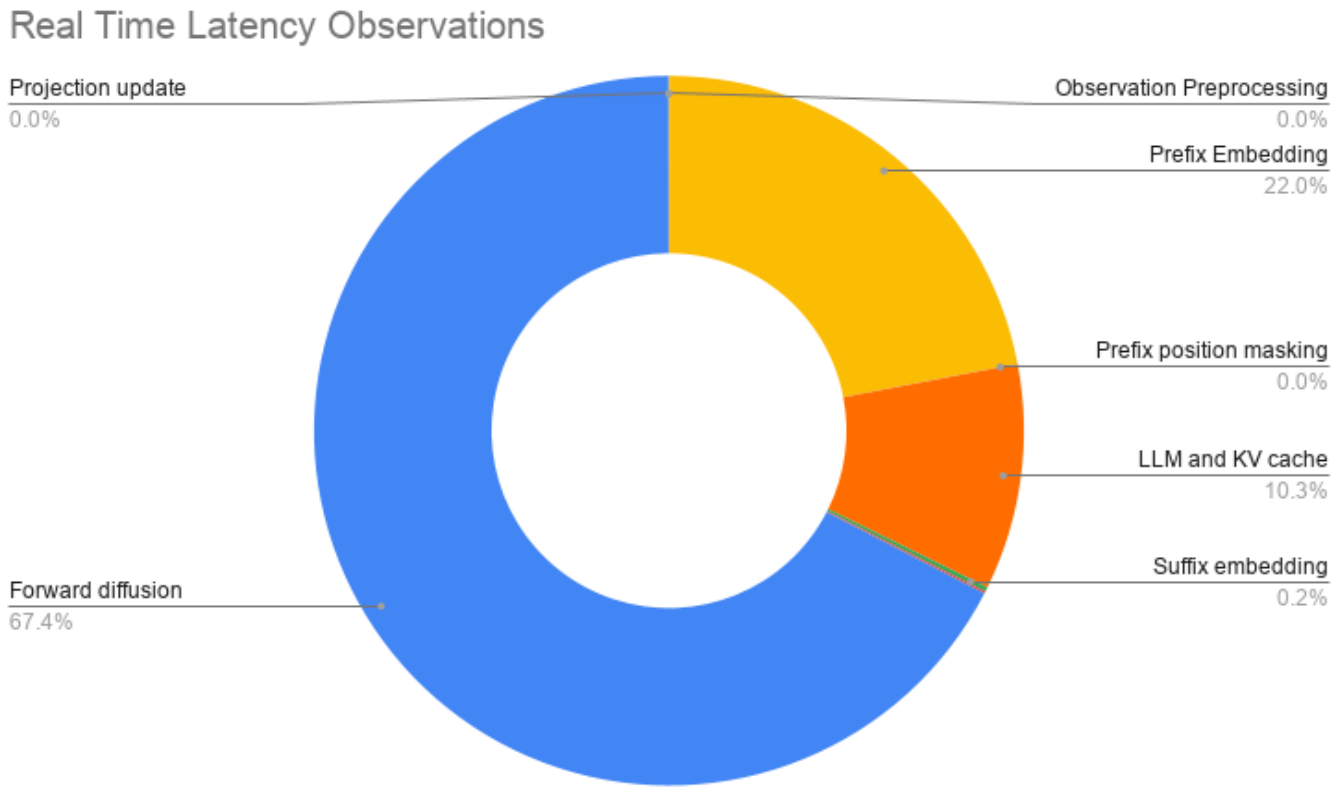
Overall Real Time Latency



Overall Real Time Latency – Table Legend

Sections	Meaning
State	State of arm (joint data)
Camera	Image capture time
Observation	Sizing and repackaging camera data
Inference	Policy inference (Pi0.5 running)
Command Loop	Sends commands to motors

Model Real Time Latency



Model Real Time Latency – Table Legend

Sections	Meaning
Preprocess Observations	Observation normalization
Noise Initializations	Initializes noise for diffusion (Not Pictured)
Prefix Embedding	Encodes image observations and prompt into tokens
Position Masking	Builds attention masks and token positions for the prefix
LLM and KV Cache	Runs prompt and image through LLM and builds KV cache
Diffusion suffix embedding	Embeds the current noisy action chunk and diffusion timestep during the action refinement loop
Diffusion position masking	Builds masks/positions for the diffusion/action tokens (Not Pictured)
Diffusion forward	Repeated LLM/action-expert forward pass inside the diffusion loop. Model repeatedly refines noisy actions into the final action chunk.
Projection update	Projects model outputs back into action space and updates the action trajectory

Before next week



JOINTS 2 AND 3



MEMORY PROFILES