Experimental Setup and Analysis

Kazybek Askarbek, Ruslan Rakhimov

December 14, 2024

Goal of the Experiment

- Develop and evaluate a Diffusion Transformer policy for multi-modal Vision-Language-Action (VLA) robotic models.
- Integrate diverse modalities (text, images and 3D depth data) into a unified representation for robust robotic control.
- Show the generalization and adaptability of policies across diverse tasks and environments.

Data Description and General Statistics

Datasets:

1. **Open X Embodiment Dataset:** Collected from 22 robots across 21 institutions, with 527 skills and 160,266 tasks.
2. **RLBench Dataset:** Multi-modal robotic tasks designed for imitation and reinforcement learning.
3. **ManiSkill2 Benchmark Dataset:** Large-scale multi-task robotic manipulation dataset.

Statistics:

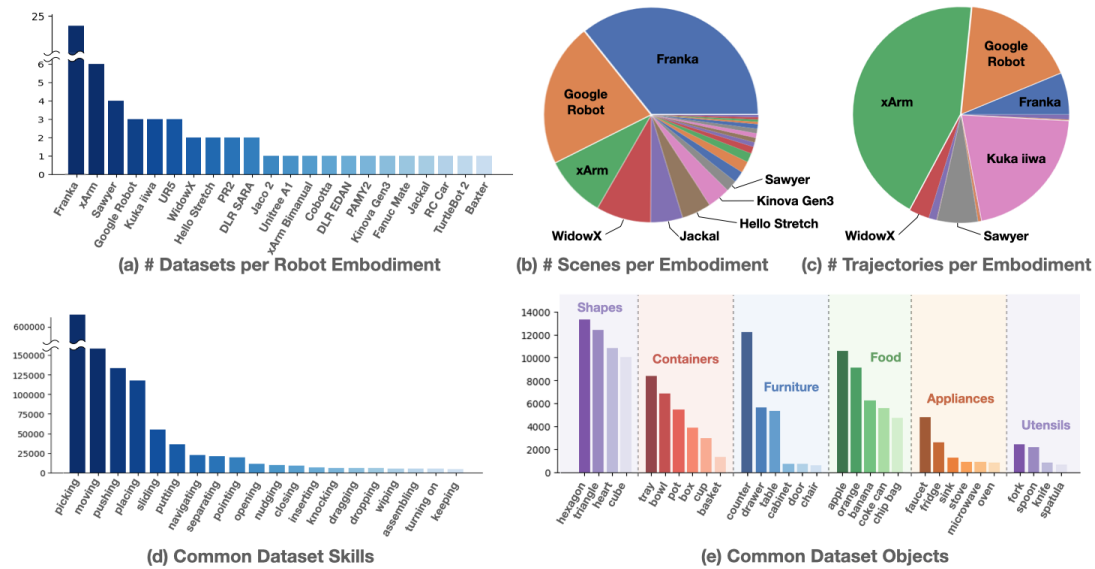

Figure: Open X-Embodiment Dataset Data Analysis

Error Analysis

- **Primary Challenges:**
  - Modality alignment: Difficulty in synchronizing temporal features from video and action data.
  - Generalization: Variability in task performance across unseen environments.
- **Common Errors:**
  - Misinterpretation of ambiguous instructions.
  - Incomplete task execution in multi-step tasks.
- **Expected Improvements:**
  - Enhanced fine-tuning protocols for low-data environments.
  - Better pre-training with larger, diverse datasets.

Expected Plots

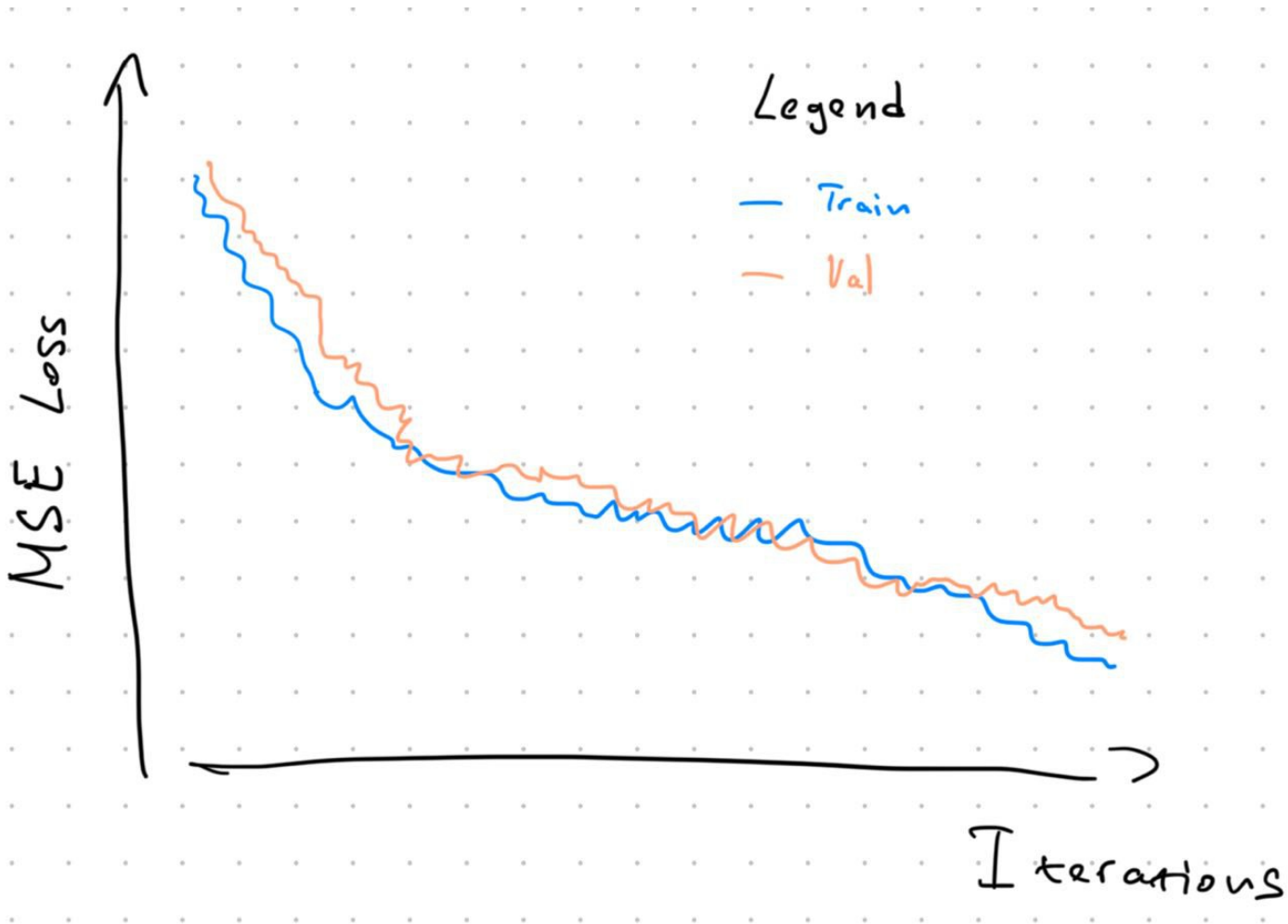

Figure: Training Loss Curve: Illustrating convergence across iterations.

Expected Plots

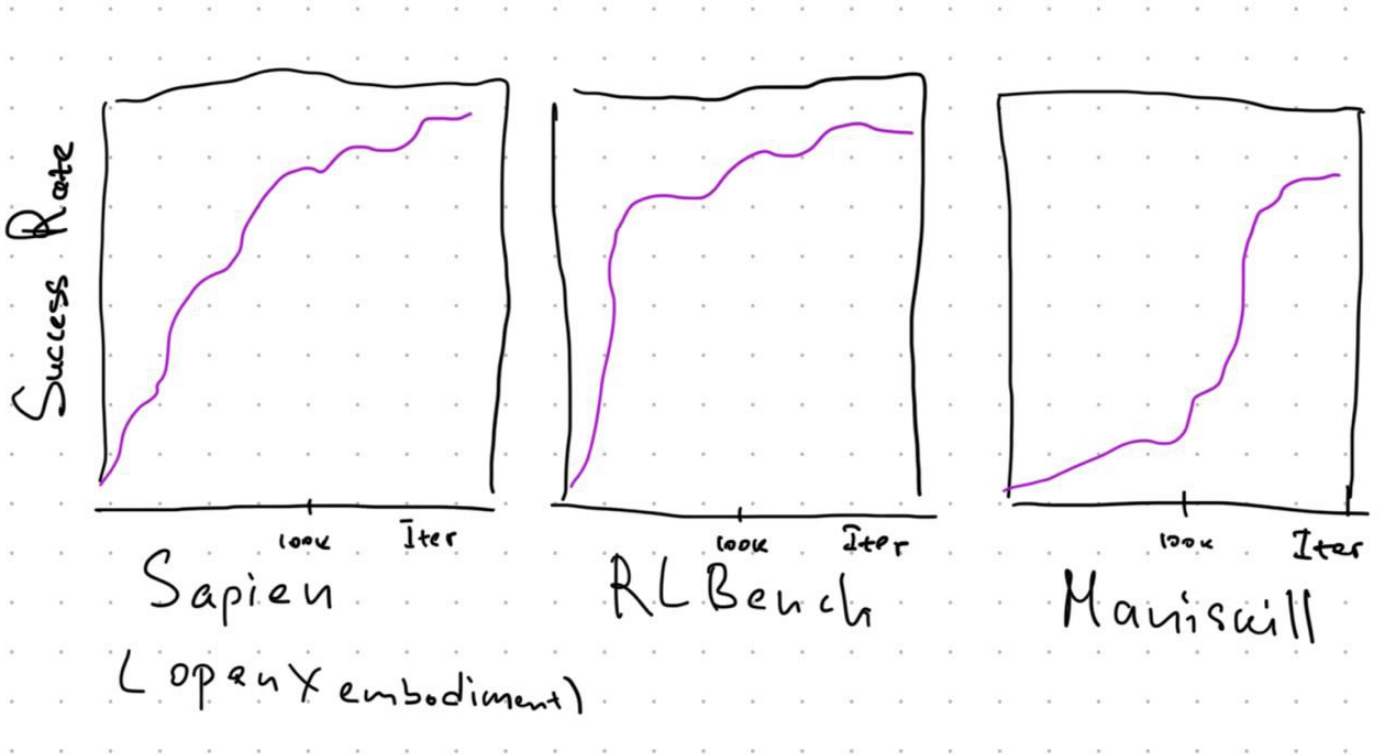

Figure: Success Rate through iterations for different benchmarks

Expected Plots

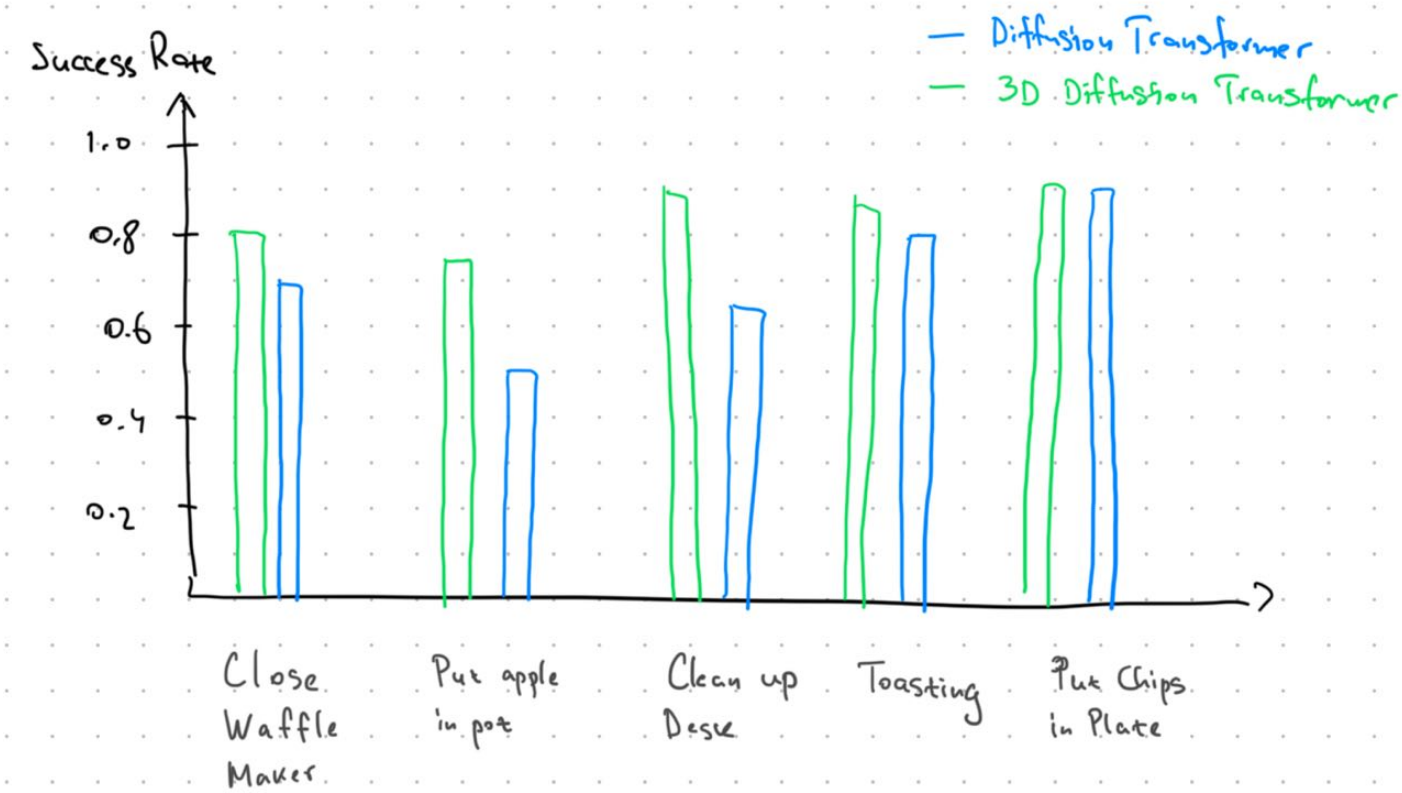

Figure: Comparing baseline vs proposed model across datasets.