Toward Aligning Human and Robot Actions via Multi-Modal Demonstration Learning

Azizul Zahid[1], **Jie Fan**[2], **Farong Wang**[1], **Ashton Dy**[1], **Sai Swaminathan**[1*], **Fei Liu**[1*]

[1] University of Tennessee Knoxville, USA [2] Istituto Italiano di Tecnologia, Italy

IEEE INTERNATIONAL CONFERENCE ON ROBOTICS AND AUTOMATION

Introduction

Motivation:

- 2D Human demos (videos / images) & 3D robots perception (RGB-D/point clouds) create a spatial mismatch.
- Task-specific LfD methods limit scalability and generalization.
- Lack of clear alignment between human intent and robot actions hinders effective imitation.

Contributions:

- **An unified cross-modal framework** aligning 2D human actions with 3D voxelized RGB-D robot inputs.
- **A dual-branch model** for sequential human intent (ResNet18+LSTM) and robot action (Perceiver Transformer).
- Achieves over 71% accuracy on RH20T (8 class tasks) for both intent and action prediction.

Problem Formulation

Objective:

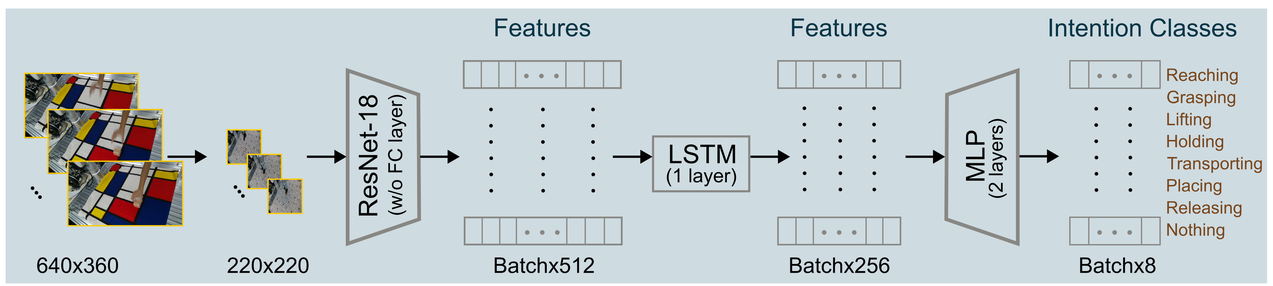
- Introduce a generalized framework to align human-robot actions across visual and geometric modalities.
- Quantify semantic correspondence by defining an alignment score :

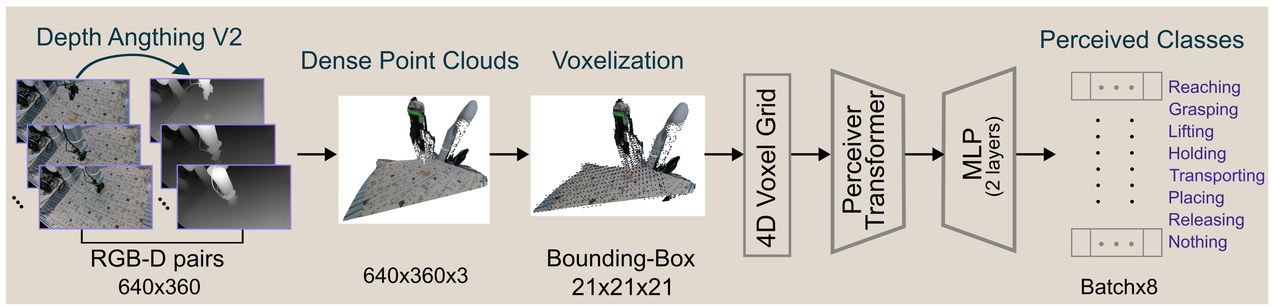
$$S(H,R) = \frac{1}{T}\sum_{t=1}^{T} \delta(i_t, j_t)\, P(i_t|H)\, P(j_t|R)$$

- Maximizing this alignment score with respect to model parameters θ :

$$\theta^* = \arg\max_{\theta} S(H, R; \theta).$$

Methodology





Arxiv

Website

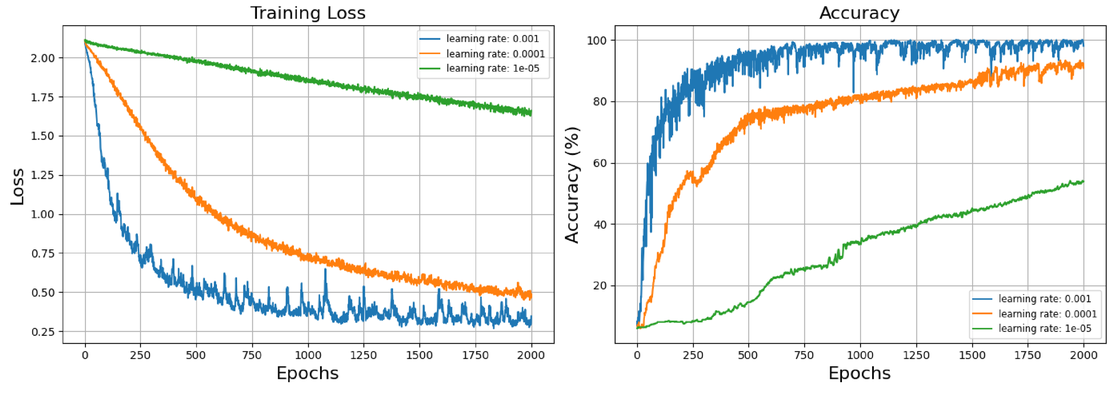
Github

Experimental Results

Trained independently on RH20T (**"PICK" task: 5 users, 10 scenes, 8 classes**) using weighted cross-entropy.

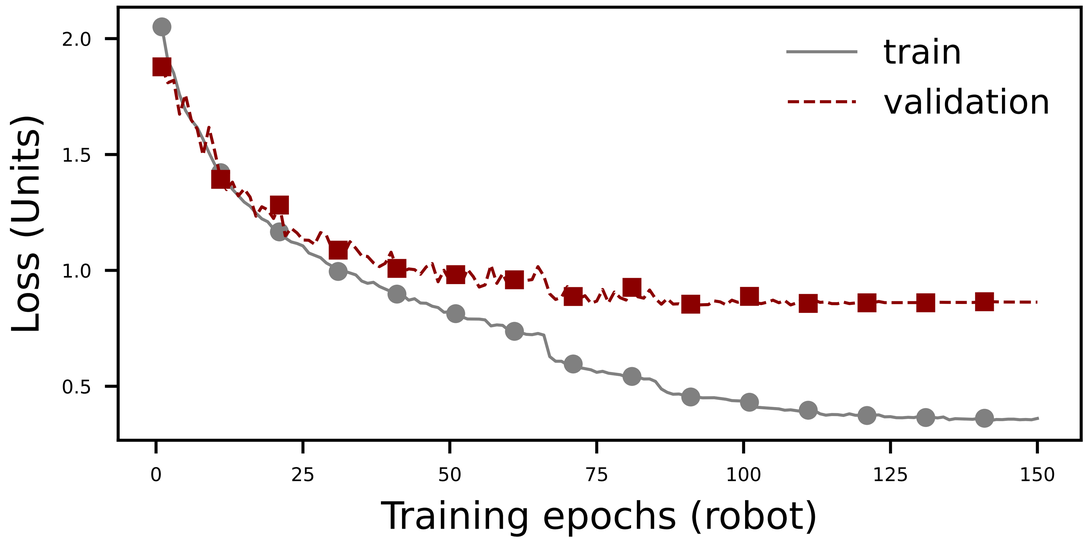
A. Human Branch:

- Validation: **71.7%**



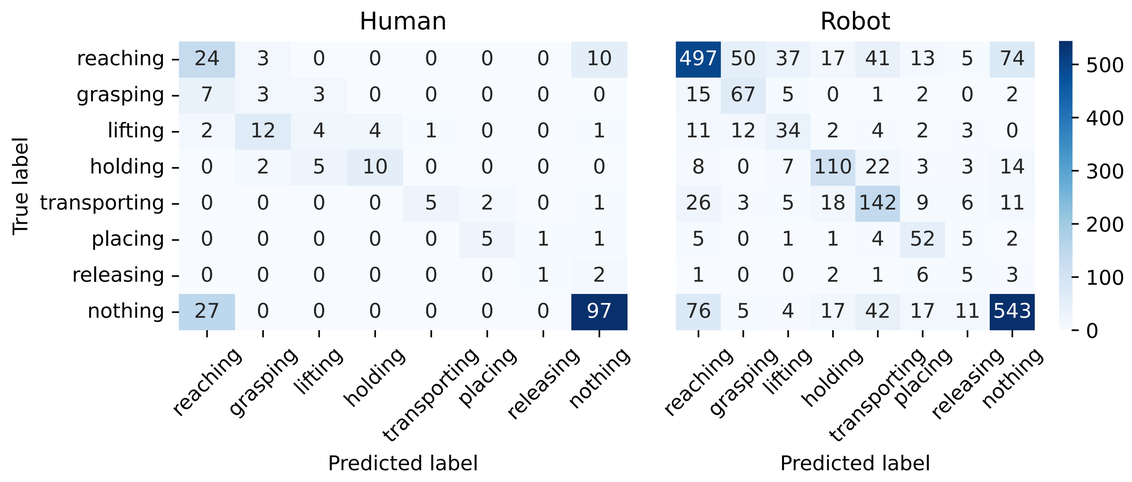
B. Robot Branch:

- Perceiver: **76.0%**
- Prediction: **83.0%**



C. Comparison:

- **Higher confidence:** nothing, holding
- **Lower confidence:** grasping, placing



* Correspondence: {fliu33,sai}@utk.edu