Cross-modality Force and Language Embeddings for Natural Human-Robot Communication

Ravi Tejwani, Karl Velazquez, John Payne, Paolo Bonato and Harry Asada

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

HARVARD MEDICAL SCHOOL

SPAULDING REHABILITATION NETWORK

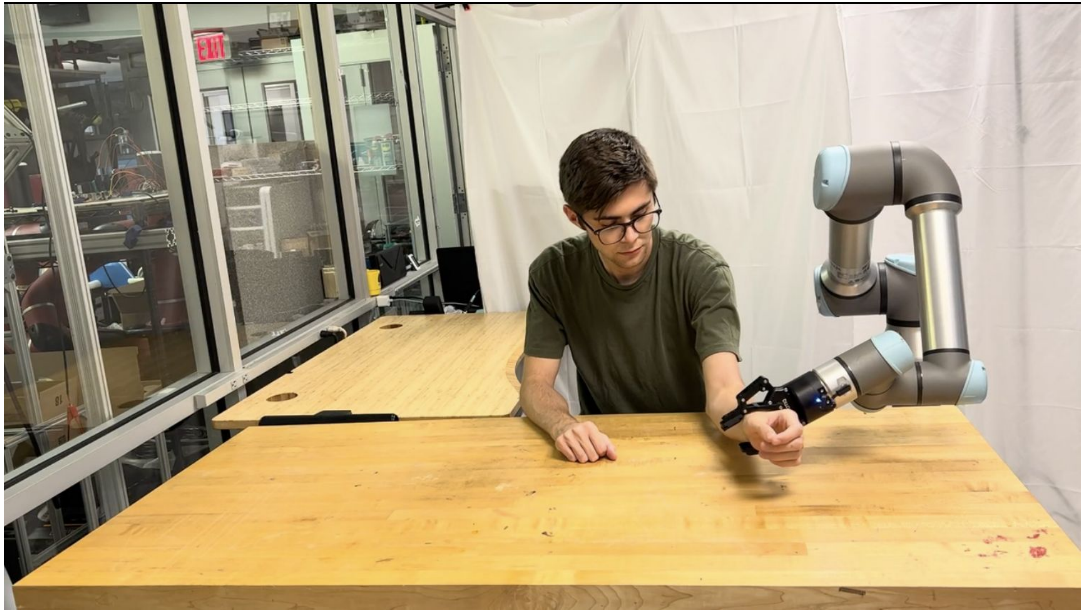
Motivation

Physical therapist guides a patient to complete a shoulder-flexion therapy exercise

Motivation

The physical therapist from the observational study demonstrated how humans naturally coordinate verbal instructions (*"gently forward"*) with precise physical forces (7N forward force)

Motivation

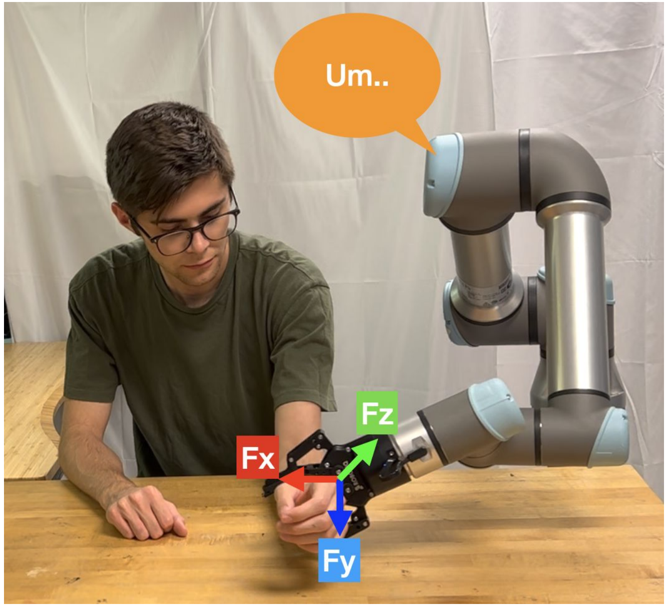


Commonly in HRI, robots interact with humans using only force

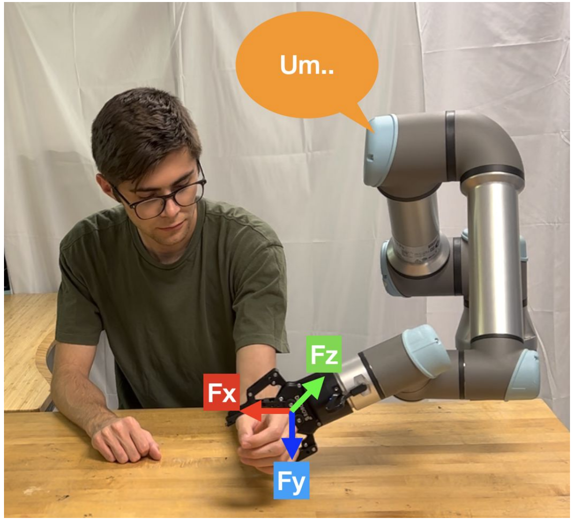
Motivation

Combination of ***force and language*** in HRI is advantageous for two reasons:

1) Forceful demonstration alone misses critical understanding and intent behind physical interactions
2) Verbal descriptions alone may not be able to fully articulate certain physical interactions



Motivation



If we want robots to use both forces and language, we need a method for ***translation*** between them to enable adaptable HRI

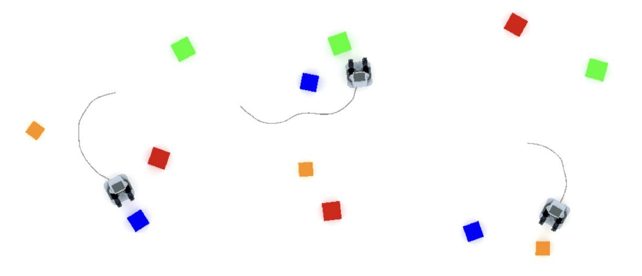
Research Question

How can we bridge the gap between continuous force signals and discrete language to create a ***unified representation*** for more intuitive human-robot interactions?

“gently right and above”

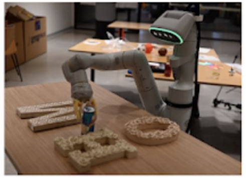
Related Work

A natural language planner interface for mobile manipulators

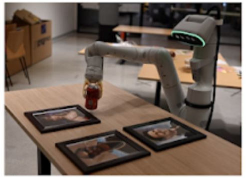


(a) “go to the blue box” (b) “move towards the green object” (c) “travel to the orange object”

RT-1

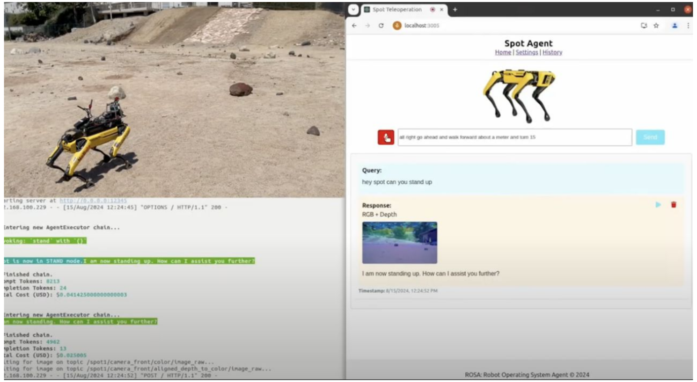


move Red Bull can to H



move coke can to Taylor Swift

Robot Operating System Agent



Related Work

Limitations of Existing Approaches

1) Focus on command-to-robot-action mapping, lacking the ability to process real-time continuous human reactive forces

2) Do not consider the nuanced relationship between force application direction, magnitude, and duration with corresponding natural language descriptions

Related Work

Our Work

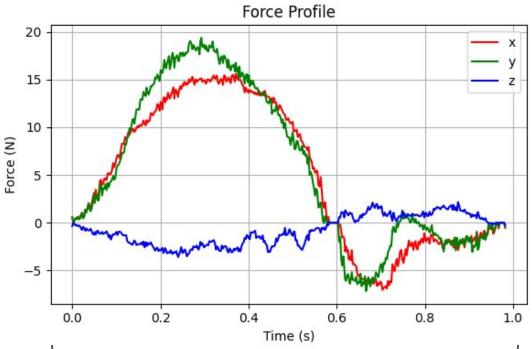
We aim to develop a method for learning a ***shared representation*** of natural language words to real time force profiles from human demonstrations

Preliminaries

Force

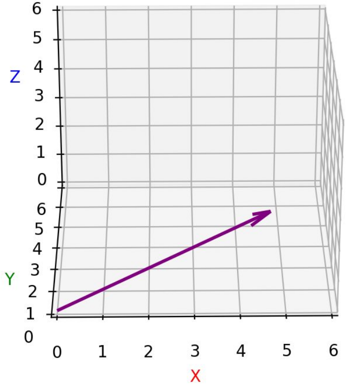
$$\text{Force Profile} = \overbrace{\begin{bmatrix} t_0 & t_1 & \dots & t_{\mathcal{N}-1} \\ F_x(t_0) & F_x(t_1) & \dots & F_x(t_{\mathcal{N}-1}) \\ F_y(t_0) & F_y(t_1) & \dots & F_y(t_{\mathcal{N}-1}) \\ F_z(t_0) & F_z(t_1) & \dots & F_z(t_{\mathcal{N}-1}) \end{bmatrix}}^{\mathcal{N}} \left.\vphantom{\begin{bmatrix} t_0 \\ F_x \\ F_y \\ F_z \end{bmatrix}}\right\}4$$

Intuitively, humans apply force in an overall ***direction, magnitude, and duration***



$\hat{\bar{F}} \approx$

$T \approx 1.0$



Can begin to infer it is in +x (*"right"*) and +y (*"forward"*) directions

Preliminaries

Language

18 direction words that describe overall force application ***direction***

12 modifier words that describe overall force application ***magnitude*** and ***duration***

A ***phrase*** is defined as a direction word plus an optional modifier word

Ex: *"down"*, *"sharply up and forward"*

Minimal Viable Vocabulary

Direction	Modifier
backward	slightly
backward-down	greatly
backward-left	smoothly
backward-right	sharply
backward-up	slowly
down	quickly
down-forward	lightly
down-left	significantly
down-right	softly
forward	harshly
forward-left	gradually
forward-right	immediately
forward-up	
left	
left-up	
right	
right-up	
up	

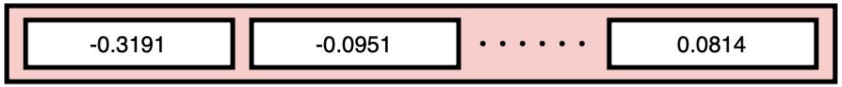
Preliminaries

Language

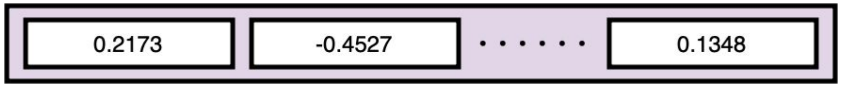
SBERT Vector Representation

We leverage SBERT (Sentence-BERT), a contextual large language model, to produce ***semantically meaningful*** continuous 768D vector representations of entire phrases

Ex: *“right”*



Ex: *“gently down right”*



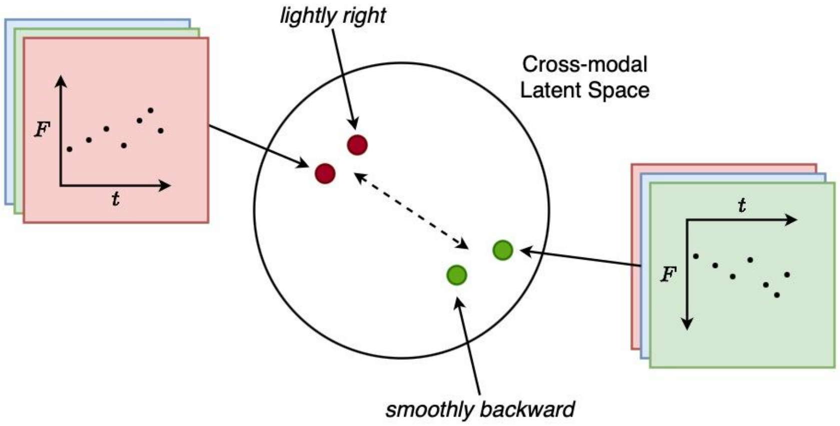
Fixed-length continuous vector embeddings enable ***mathematical operations*** such as cosine similarity to judge how aligned two different phrases are

Preliminaries

Cross-Modality Embedding

We aim to develop a framework that can model the shared representation of force and language as a ***shared latent space***, $\vec{z} \in \mathbb{R}^{16}$, to align force profiles and phrases

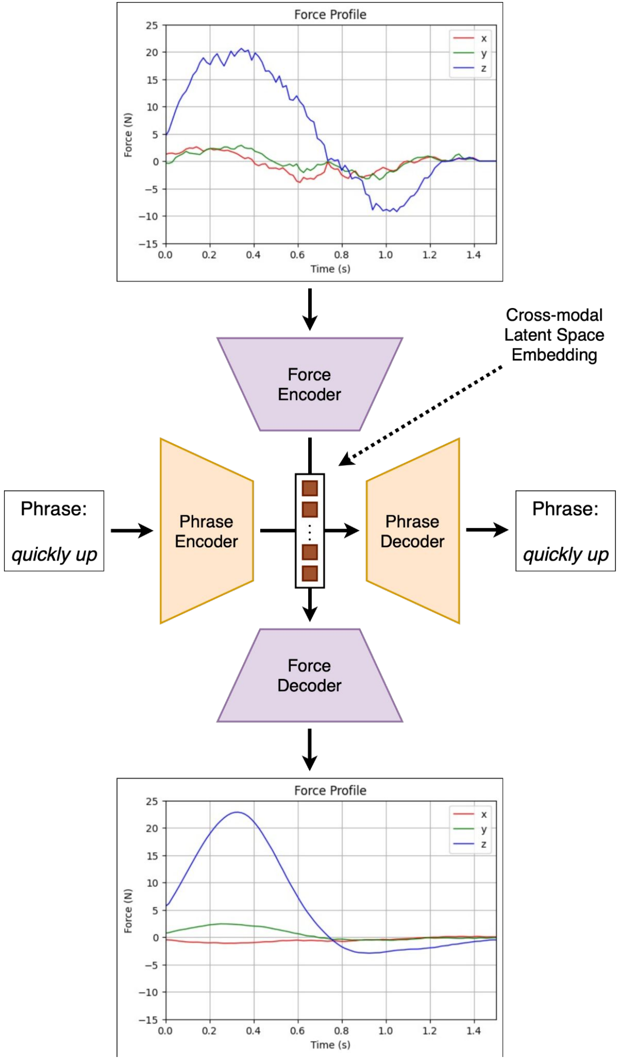
Enables translation between modalities via encoding/decoding instances to and from latent space



Corresponding forces and language are mapped distance-wise closer than noncorresponding instances

Dual Autoencoder Model

There is an autoencoder for each modality that is responsible for ***encoding and decoding*** instances to and from the shared latent space



Multitask Learning

$$\mathcal{L} = k_r \mathcal{L}_r + k_z \mathcal{L}_z + k_t \mathcal{L}_t$$

The model was trained to minimize 3 cost functions: *reconstruction loss* (r), *contrastive loss* (z), and *translation loss* (t)

Forcing the model to perform well in multiple tasks instead of a single one encourages it to learn a more robust cross-modality space representation and prevents overfitting

Hyperparameters k_r, k_z, k_t control relative importance of each loss function $\mathcal{L}_r, \mathcal{L}_z, \mathcal{L}_t$ respectively

Multitask Learning

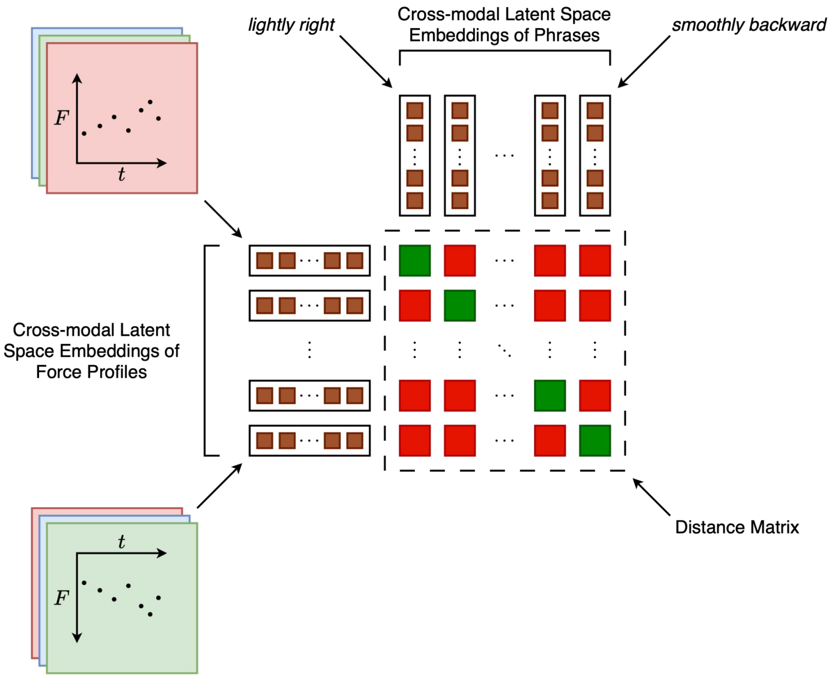
Contrastive Loss $\mathcal{L}_z$

Aligns corresponding force profile and phrase embeddings similarly in the shared latent space by bringing them distance-wise closer and pushing away non-corresponding instances:

The contrastive loss for a batch of n force-phrase pairs (z_f^i, z_p^i) in the shared latent space is:

$$\mathcal{L}_c = \sum_{i=1}^{n} \|z_f^i - z_p^i\|^2 - \lambda \sum_{i=1}^{n} \sum_{j \neq i}^{n} \max(0, m - \|z_f^i - z_p^j\|^2)$$

where λ controls the negative pair weighting and m is the margin parameter.



On-diagonal elements minimized, while off-diagonal elements maximized

Data Collection

10 volunteers completed a total of 840 trials involving human demonstrations of force and language translation

backward

to the front and downward

to the back quickly

smoothly right

quickly rightward and downward

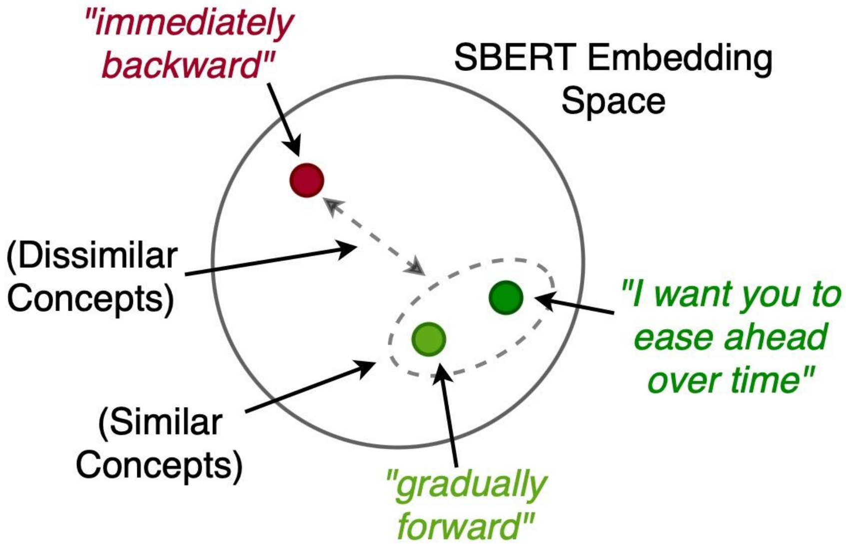
to the front and left slightly

below significantly

Arbitrary Language Input

Leverage SBERT to map arbitrary language inputs into phrases defined by our Minimal Viable Vocabulary

Find the Minimal Viable Vocabulary phrase that most closely matches the arbitrary language by using the cosine similarity of their respective SBERT embedding vectors



Evaluation

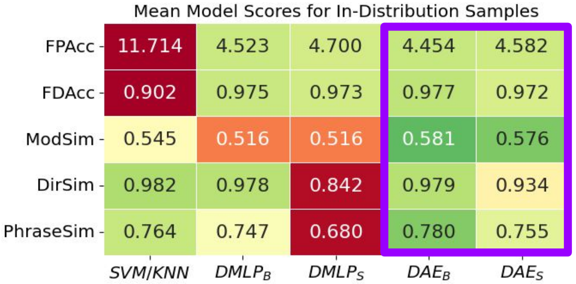
Baselines

- Support Vector Machine & K Nearest Neighbors: ***Nonparametric*** approach
- Multilayer Perception: Basic neural network approach, ***no shared latent space***
- Dual Autoencoder: ***Our method***

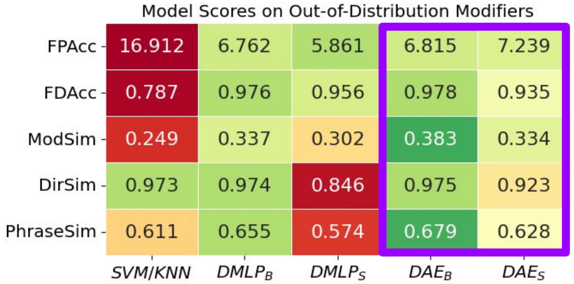
Evaluation

Results

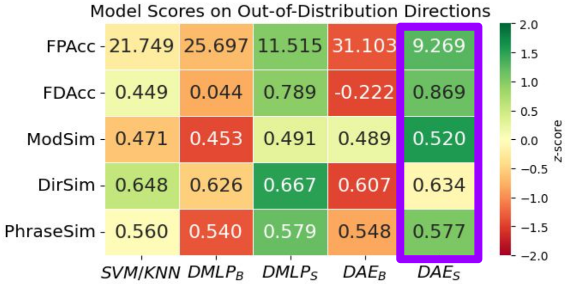
Random Train-Test Split Experiment



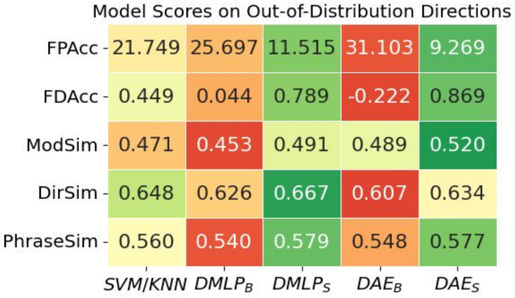
Unseen Modifiers Experiment



Unseen Directions Experiment



Our method is capable of ***translating*** between force profiles and phrases

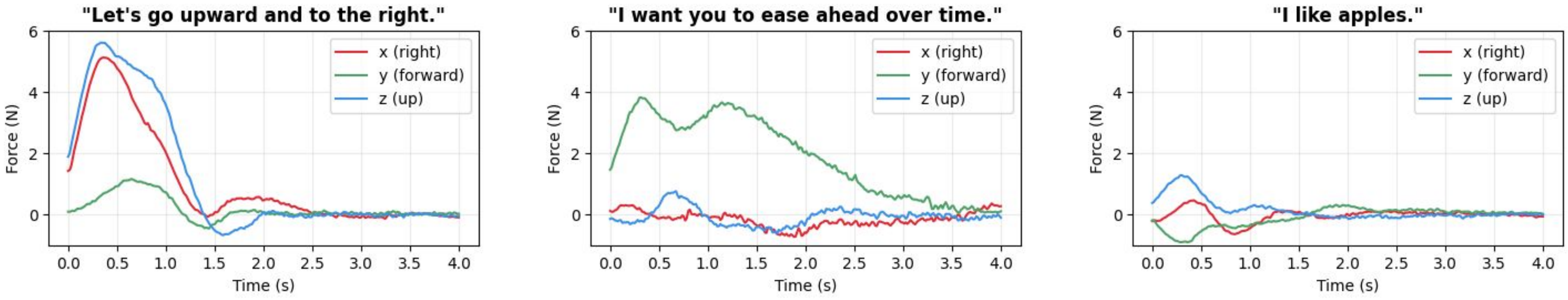


Using SBERT embeddings enables even ***greater*** generalization capability

Evaluation

Results

Inference examples of going from arbitrary phrases to force profiles…

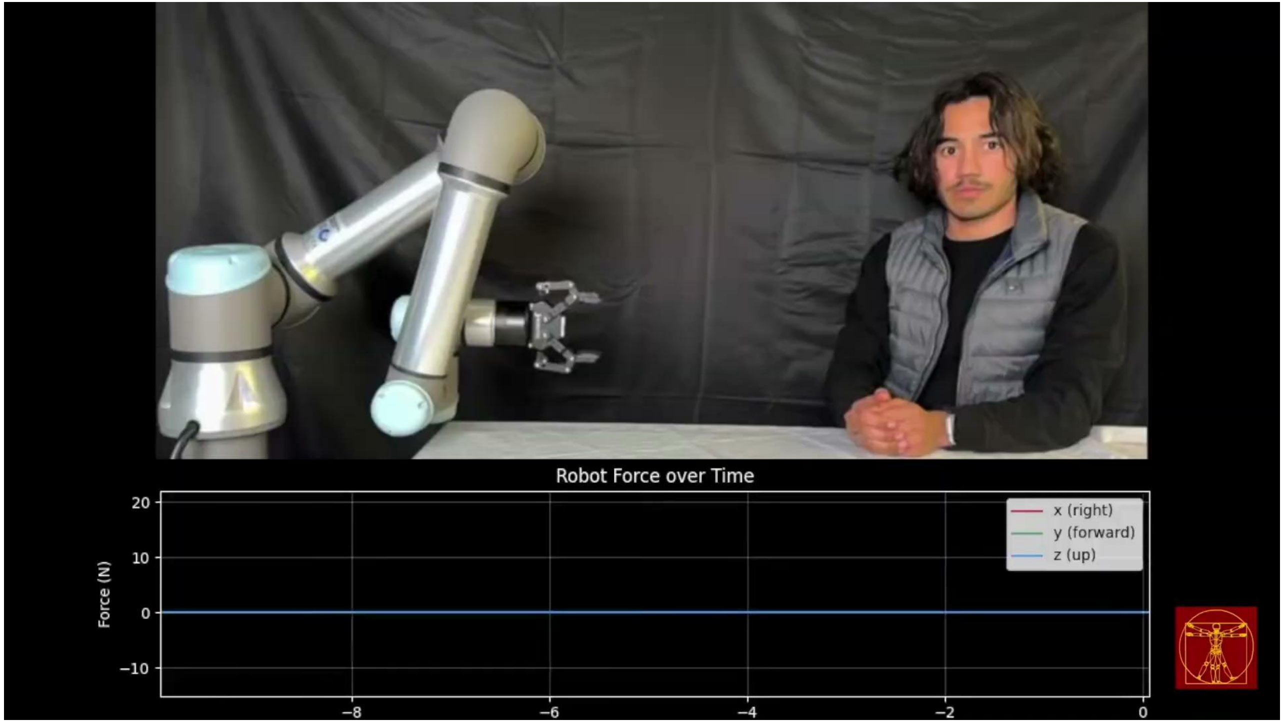


…using our dual autoencoder with SBERT phrase representation

Demo

The robot is given arbitrary verbal commands and translates them into force profiles

Demo


Robot Force over Time
Force (N)
20
10
0
−10
−8
−6
−4
−2
0
x (right)
y (forward)
z (up)

More details on our website