PRIMUS: Pretraining IMU Encoders with Multimodal Self-Supervision

Arnav M. Das, Chi Ian Tang, Fahim Kawsar, Mohammad Malekzadeh

University of Washington, Nokia Bell Labs, University of Glasgow

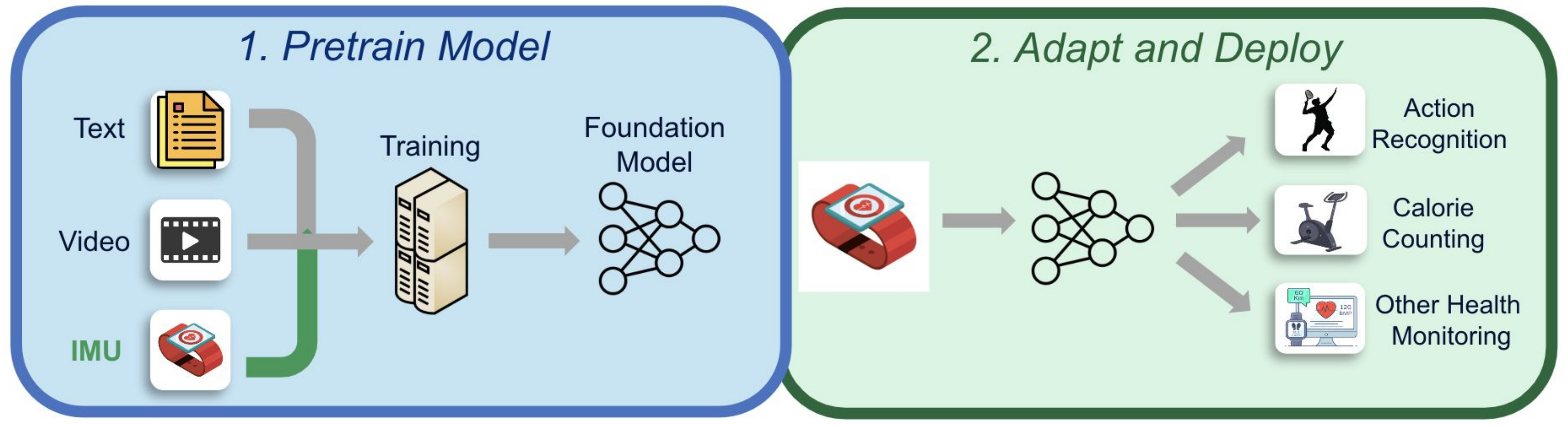
Paper

Background

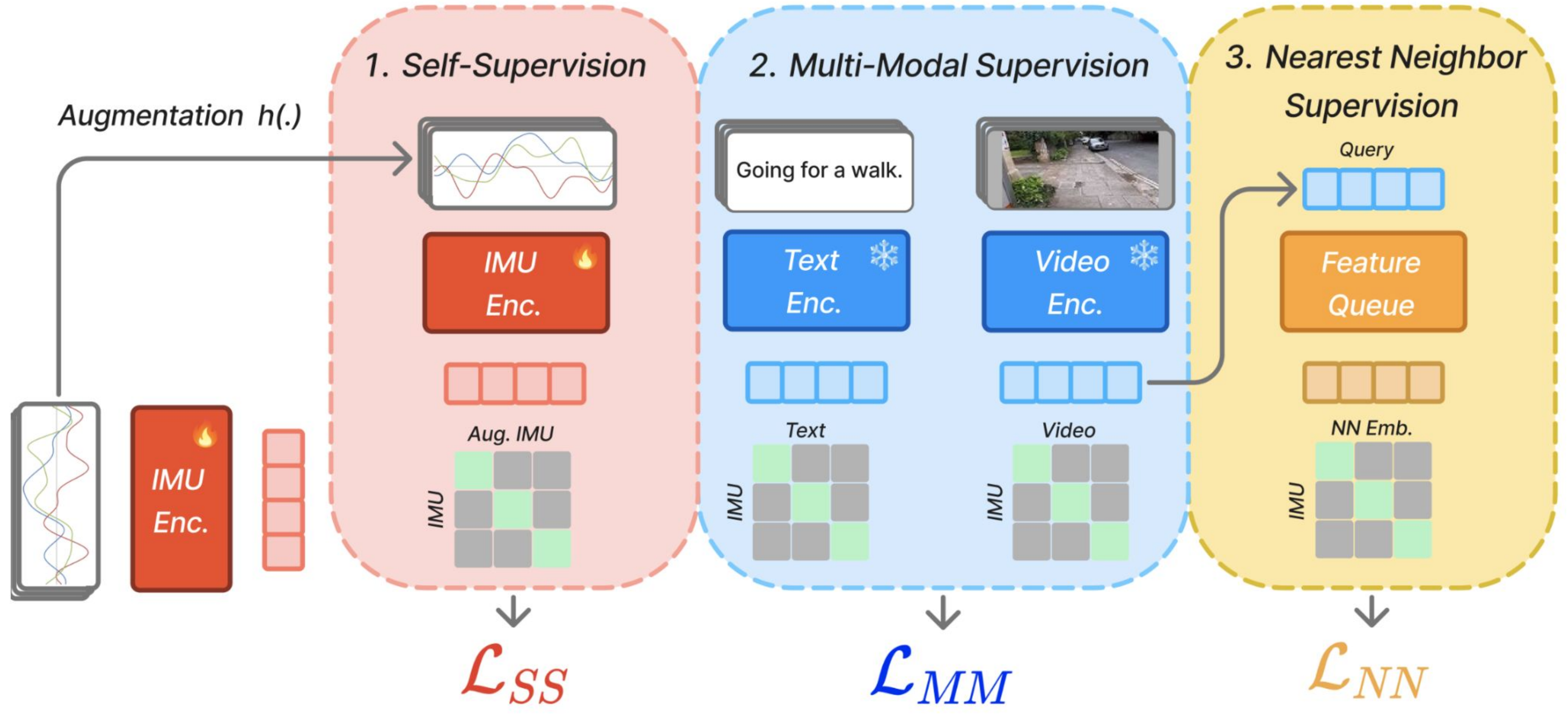
- Wearable devices contain Inertial Measurement Units (IMU), which produce rich information about human physical behavior
- Wild, uncurated IMU data is widely available!
- Labeled data is hard to acquire, since sensor data is inherently uninterpretable
- ***How do we extract meaningful information from IMU signals with limited labeled data?***

Pretrain and Adapt

- Modern ML pipelines in other domains use the pretrain and adapt framework
- EgoExo4D was recently released which contains large-scale IMU data from head-placed sensors, aligned with video and text
- Can we pretrain on this dataset to train a model with ***transferable representations?***

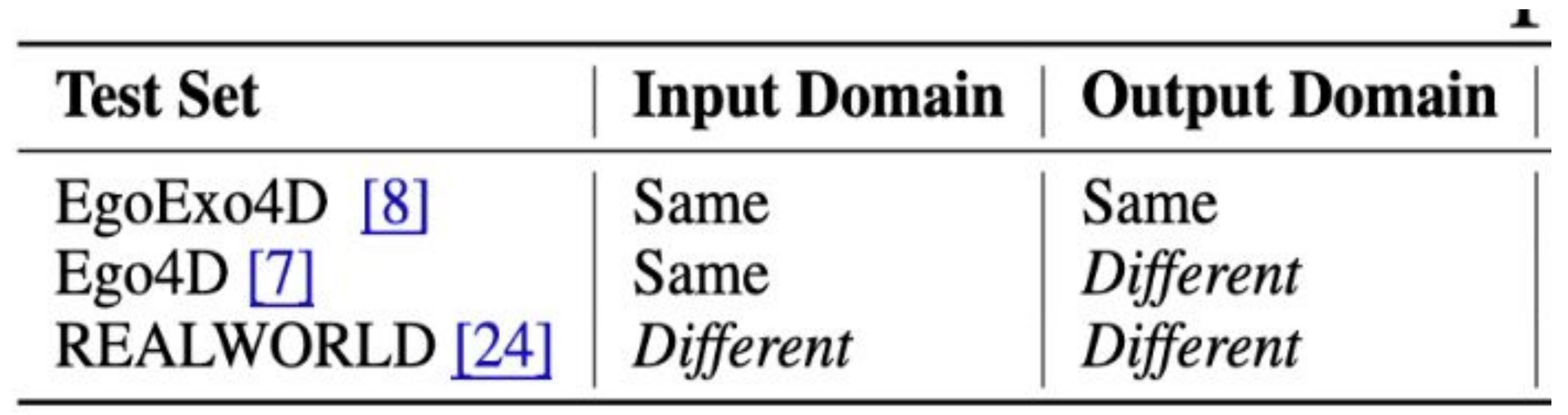


Training Objective



- $\mathcal{L}_{SS}$ encourages representations to be invariant to augmentations such as random scaling
  - ○ These types of invariances help cross-device generalization!
- $\mathcal{L}_{MM}$ allows us to ***distill*** rich semantic information from foundation models in other modalities into IMU representations
- $\mathcal{L}_{NN}$ increases the number of positive samples we have for each IMU frame
  - ○ Since we have reliable features for other modalities, we can use these to search for nearest-neighbors!

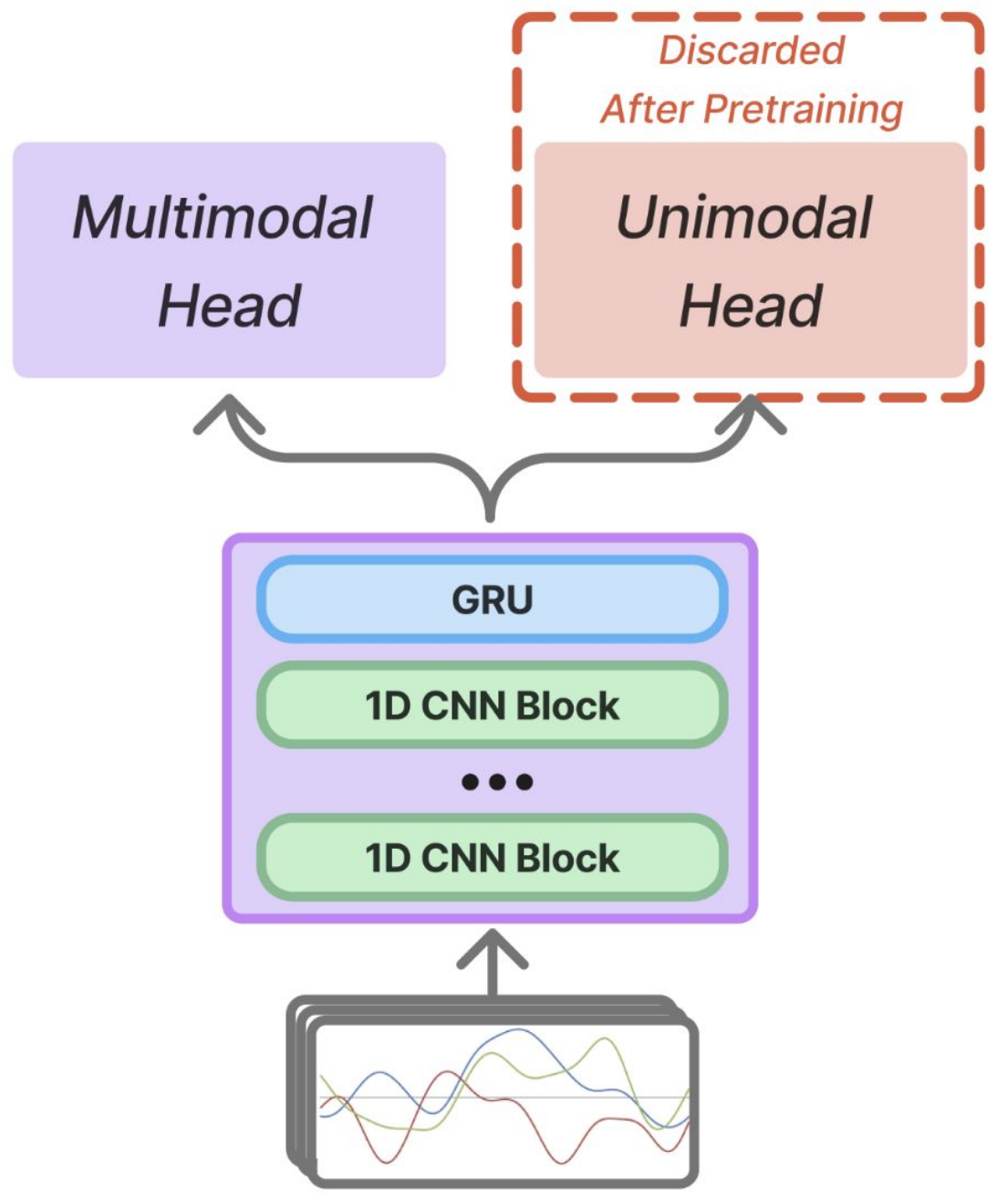
Evaluation Datasets



Test Set	Input Domain	Output Domain
EgoExo4D [8]	Same	Same
Ego4D [7]	Same	*Different*
REALWORLD [24]	*Different*	*Different*

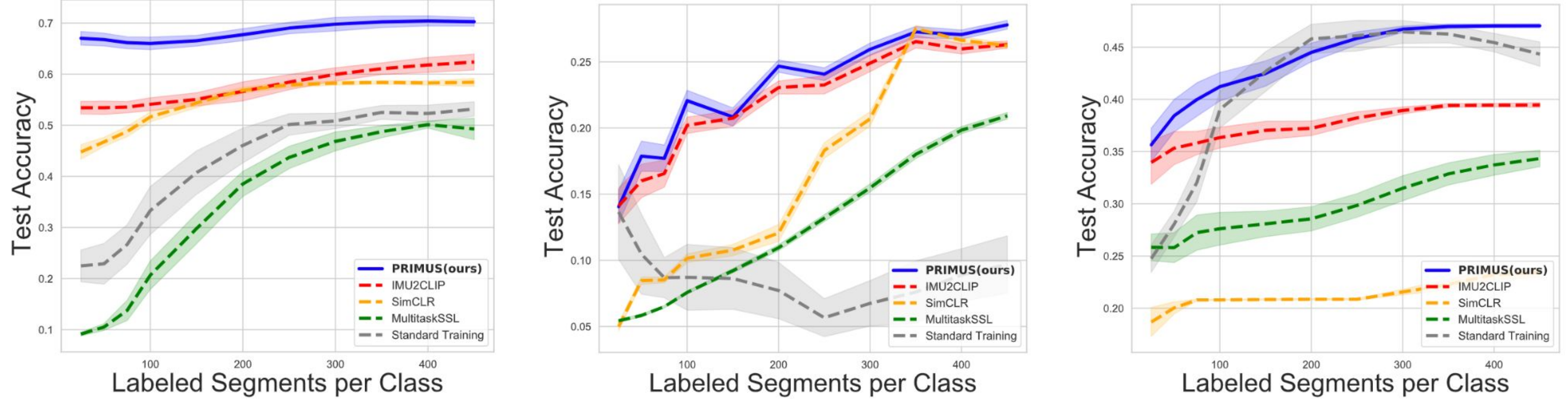
We evaluate on datasets that are OOD!

Architecture



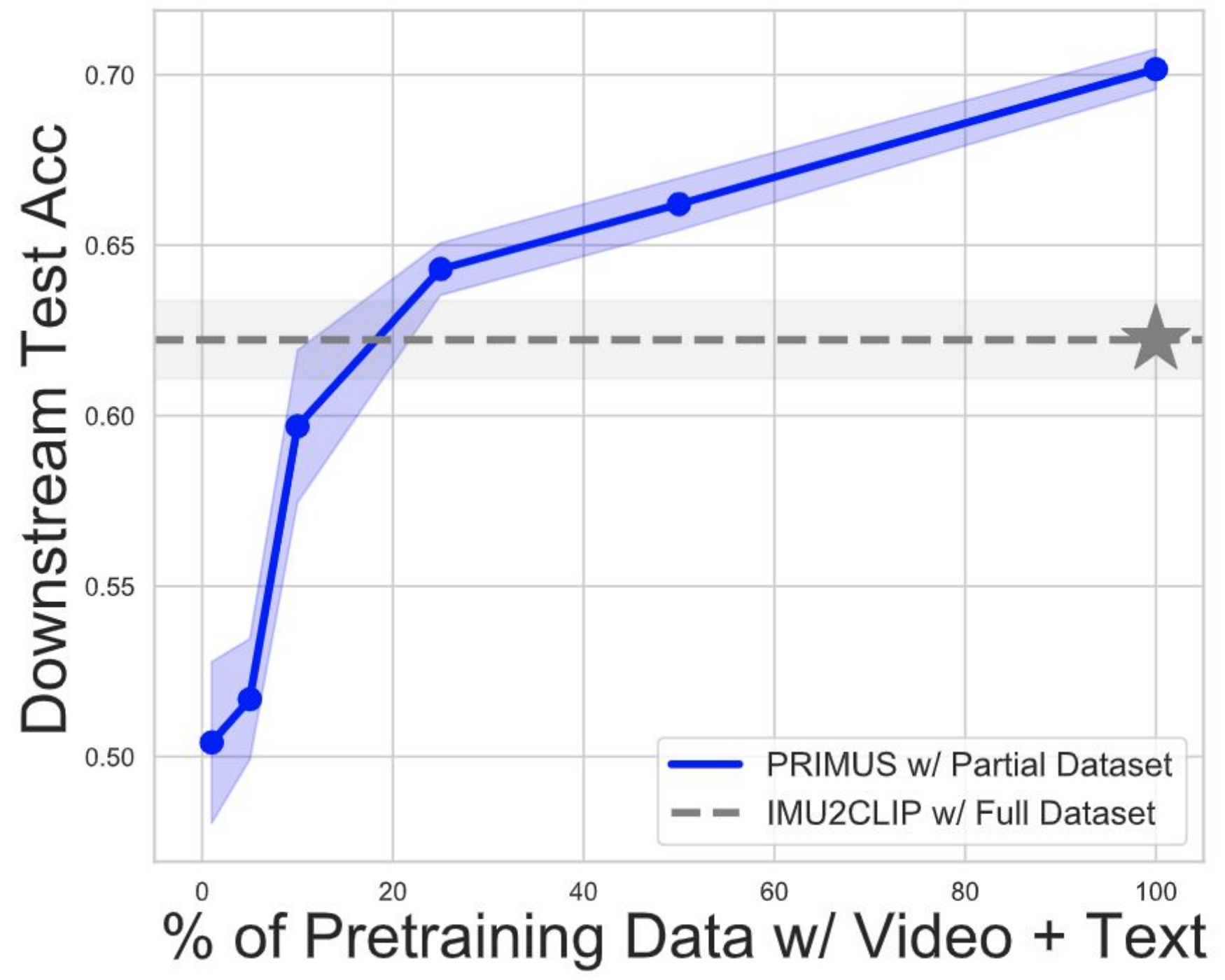
We use a lightweight CNN + GRU based architecture which only has 1.4M parameters

Results



(a) EgoExo4D Results (b) Ego4D Results (c) REALWORLD Results

PRIMUS provides consistent improves in few-shot transfer!



PRIMUS requires less multiview data to achieve the same performance as other approaches