Where am I and where should I go? Grounding positional and directional labels in a disoriented human balancing task

Sheikh Mannan and Nikhil Krishnaswamy

(Dis)embodiment Conference, September 15-16, 2022, Gothenburg, Sweden

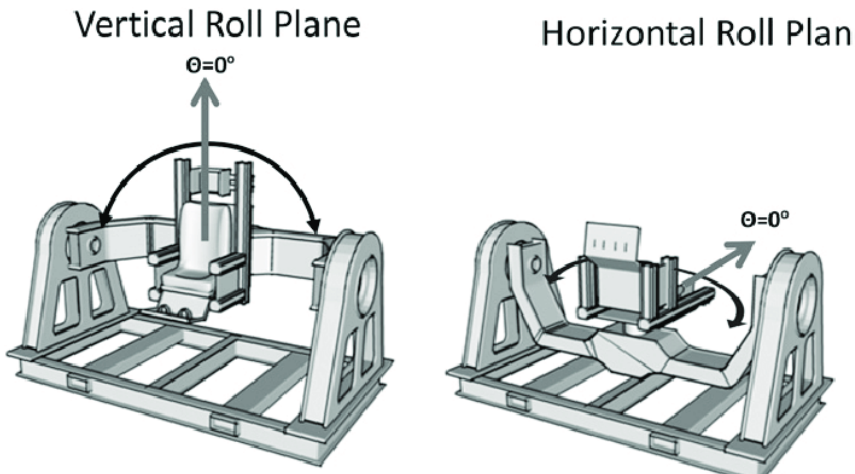
Colorado State University

Outline

- Introduction
- MARS balancing task
- Data
- Model architecture
- Evaluation
- Results
- Discussion
- Conclusions and future work

Introduction

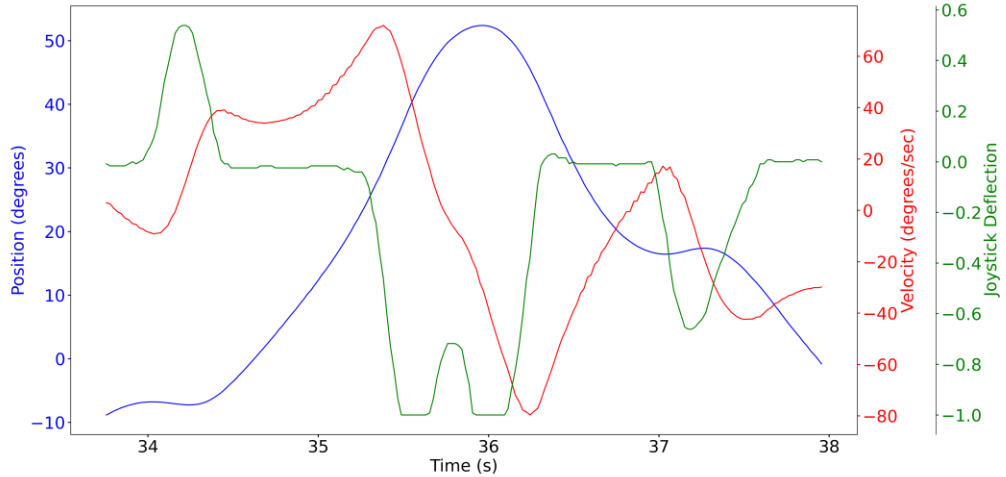
- Meteoric rise of large language models (LLMs) facilitate coherent, grammatical text generation using high-dimensional representations of language
- LLMs still fail at understanding the current situational context that comes from non-textual (or non-visual) context
- Consider human spatial disorientation, where even expert humans subject to gravitational transitions where gravitational cues sensed by the vestibular system are absent, lead to fatal accidents (Shelhamer, 2015; Cowings et al., 2018)
- Numerical AI models with access to quantitative information about position and movement can potentially determine when humans may lose control and intervene by telling humans what to do
- Can embeddings from such numerical models grounded with BERT (Devlin et al., 2019) embeddings representing thought vectors for position embody a spaceflight-analog balancing task and act as a countermeasure for spatial disorientation?



MARS balancing task

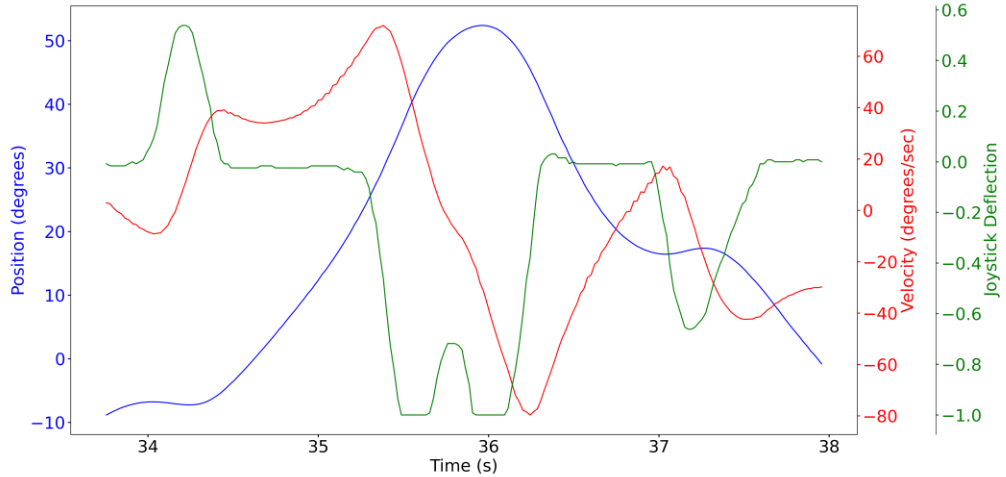
- Multi-Axis Rotation (MARS) device
- MARS programmed with inverted pendulum dynamics
- Crash limits set to +/- 60º from the Direction of Balance (DOB)
- Subjects balance themselves about the supine axis with blindfolds and noise cancelling headphones

Data



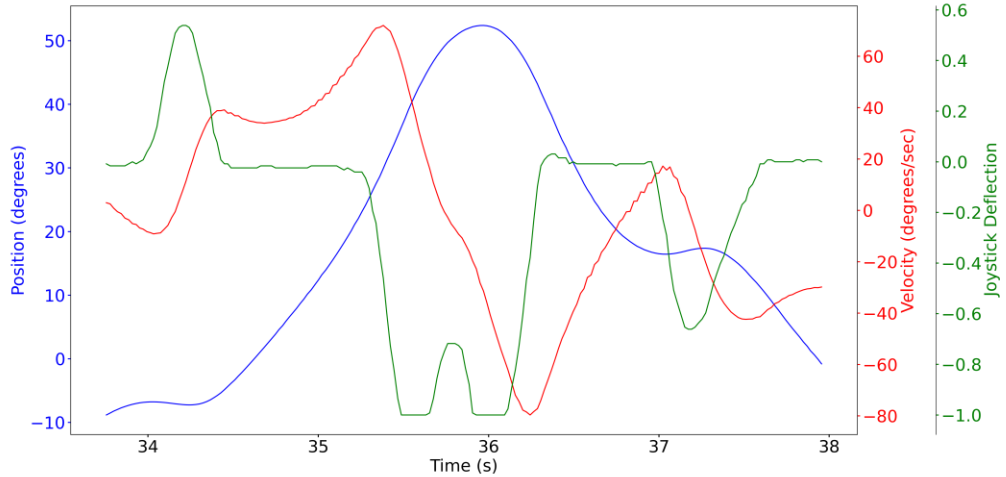
- MARS data from Vimal et al. (2020)
  - 34 participants with performing 40 trials over 2 days, each trial 100 seconds long
  - primary data points collected: ***angular position, angular velocity & joystick deflections***
- Proficiency labels from Vimal et al. (2020)
  - participants clustered based on their balancing performance using various engineered features, such as ***Crash frequency, Anticipatory joystick deflections, Destabilizing joystick deflections,*** etc.
  - ***Proficient*** (or "Good"), ***Somewhat Proficient*** (or "Medium"), and ***Not Proficient*** (or "Bad")
- Positional and Direction labels (our addition)
  - for grounding the situated numerical features from the MARS to a linguistic representation
  - representations would be possible answers to the questions ***"where am I?"*** and ***"where should I go?"***
  - for position relative to the DOB a human may think "I have drifted more towards the right"; generated by third-party annotators for each of the three regions ***left, right,*** and ***center***
  - for direction, the human would have 3 choices to deflect the joystick ***left, right,*** or ***center***

Data

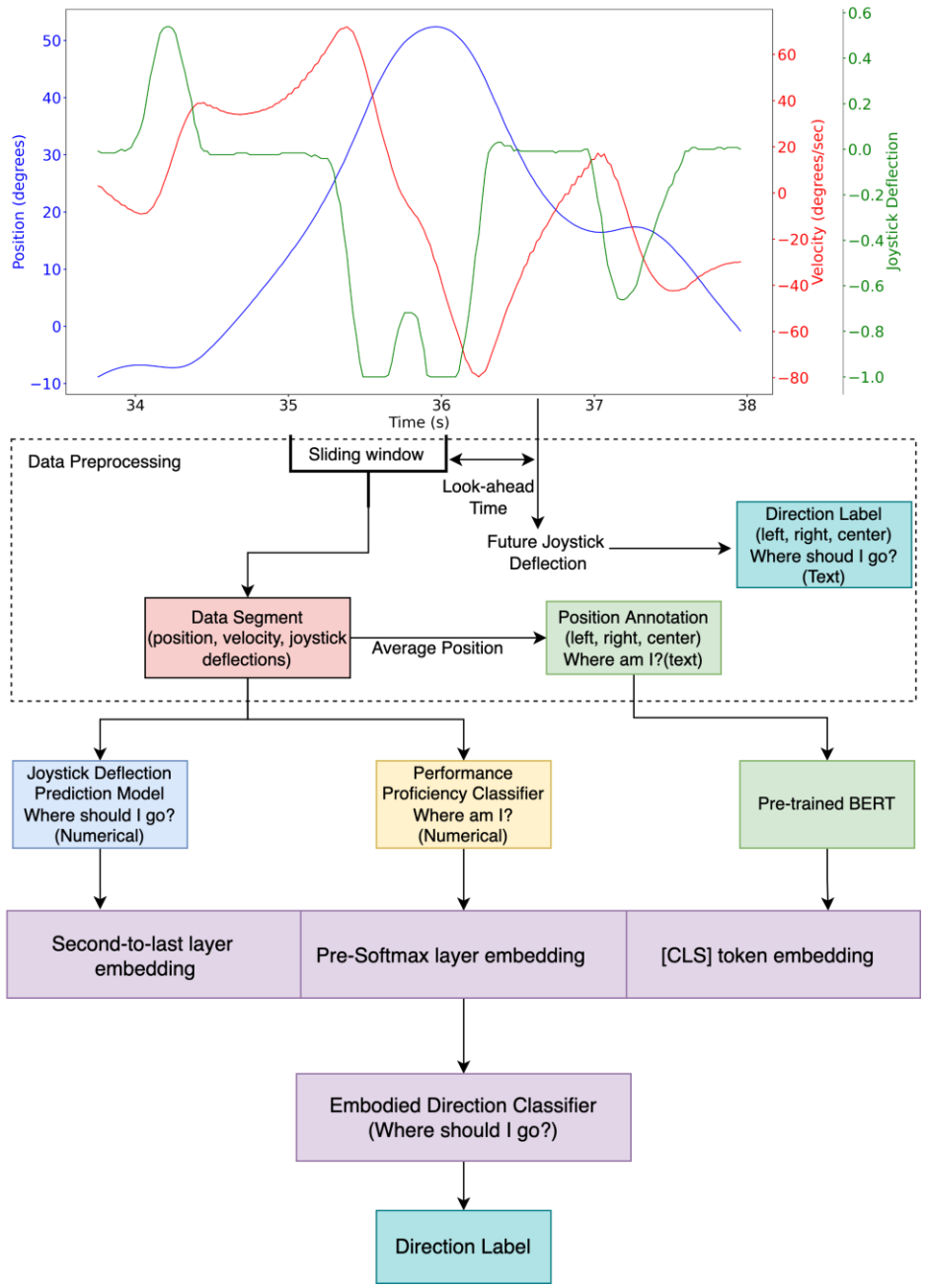


- MARS data from Vimal et al. (2020)
  - 34 participants with performing 40 trials over 2 days, each trial 100 seconds long
  - primary data points collected: ***angular position, angular velocity & joystick deflections***
- Proficiency labels from Vimal et al. (2020)
  - participants clustered based on their balancing performance using various engineered features, such as ***Crash frequency, Destabilizing joystick deflections,*** etc.
  - ***Proficient*** (or "Good"), ***Somewhat Proficient*** (or "Medium"), and ***Not Proficient*** (or "Bad")
- Positional and Direction labels (our addition)
  - for grounding the situated numerical features from the MARS to a linguistic representation
  - representations would be possible answers to the questions ***"where am I?"*** and ***"where should I go?"***
  - for position relative to the DOB a human may think "I have drifted more towards the right"; generated by third-party annotators for each of the three regions ***left, right,*** and ***center***
  - for direction, the human would have 3 choices to deflect the joystick ***left, right,*** or ***center***

Data



- MARS data from Vimal et al. (2020)
  - 34 participants with performing 40 trials over 2 days, each trial 100 seconds long
  - primary data points collected: ***angular position, angular velocity & joystick deflections***
- Proficiency labels from Vimal et al. (2020)
  - participants clustered based on their balancing performance using various engineered features, such as ***Crash frequency, Anticipatory joystick deflections, Destabilizing joystick deflections,*** etc.
  - ***Proficient*** (or "Good"), ***Somewhat Proficient*** (or "Medium"), and ***Not Proficient*** (or "Bad")
- Positional and Direction labels (our addition)
  - for grounding the situated numerical features from the MARS to a linguistic representation
  - representations would be possible answers to the questions ***"where am I?"*** and ***"where should I go?"***
  - for position relative to the DOB a human may think "I have drifted more towards the right"; generated by third-party annotators for each of the three regions ***left, right,*** and ***center***
  - for direction, the human would have 3 choices to deflect the joystick ***left, right,*** or ***center***

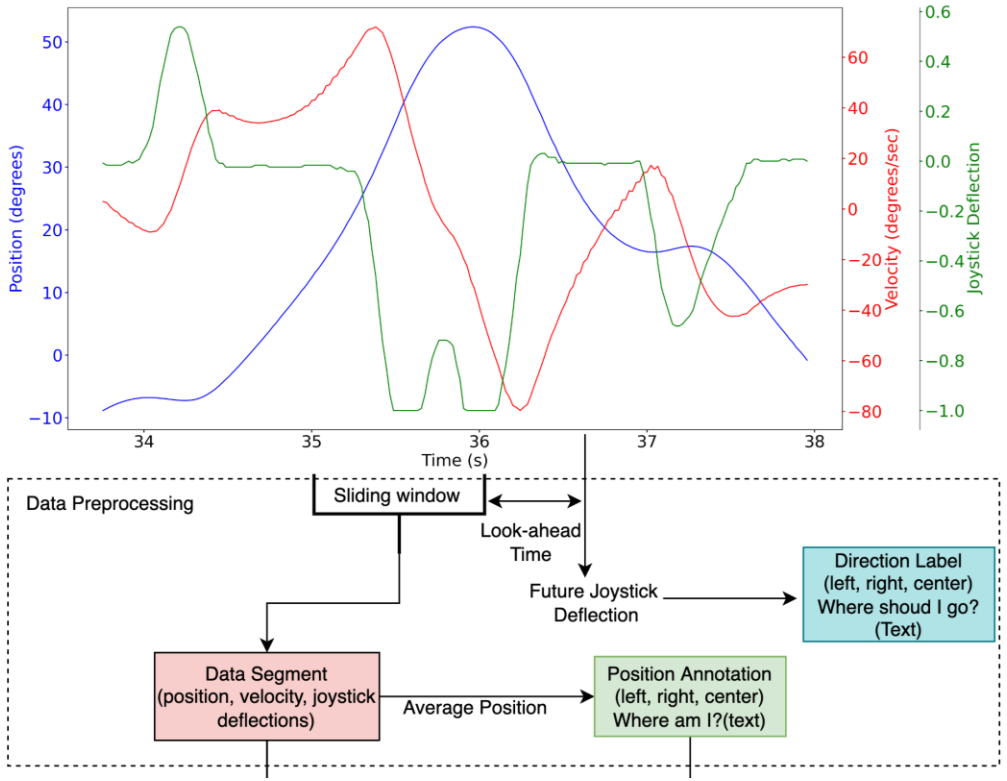


Model Architecture

1. Data Preprocessing
2. Joystick-Deflection Prediction Model
3. Performance Proficiency Classifier
4. BERT Sentence Embeddings
5. Embodied Direction Classifier (EDC)

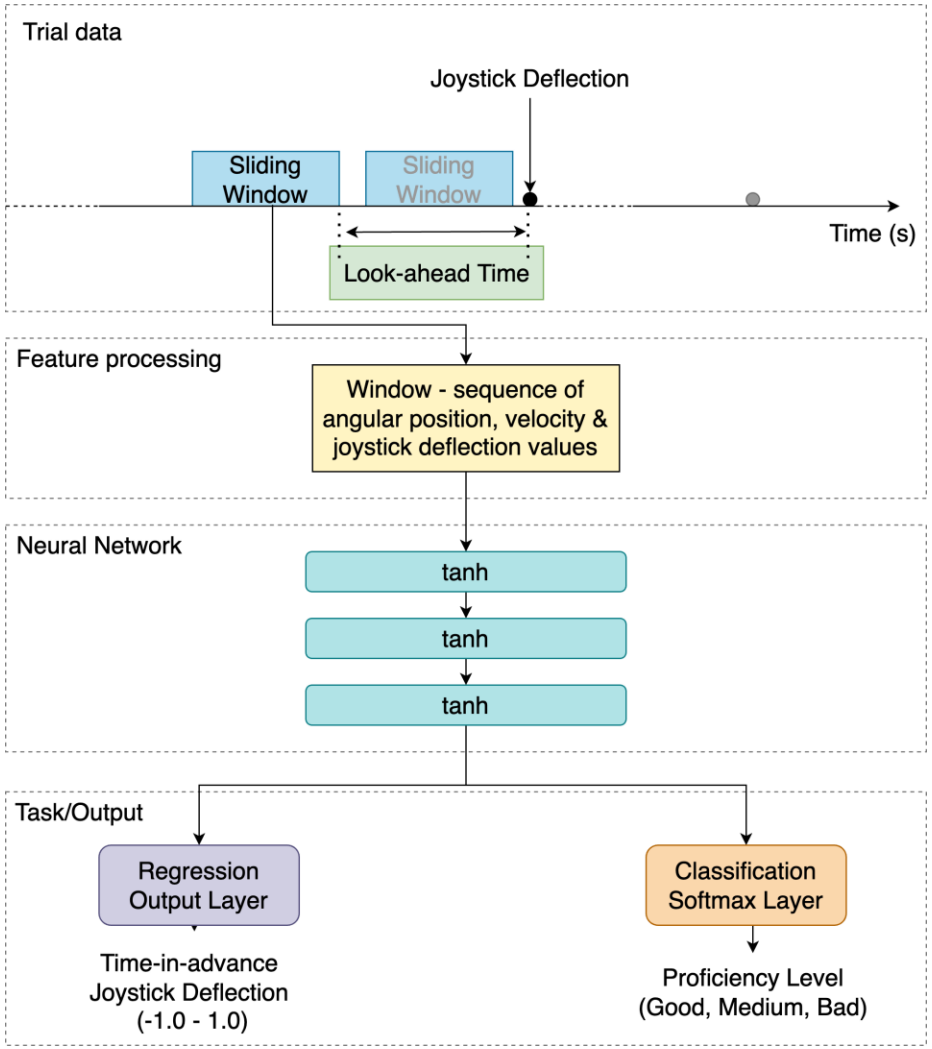
Model Architecture: Data Preprocessing

- Fixed sliding window technique to extract segments of joystick deflections, angular velocity and positions
- User in control and no crashes occurred while looking ahead ***y*** seconds in the future
- For each viable window, random sentence annotation assigned for region corresponding to the user's average position in the window, e.g., "I think I am somewhere in the center" or "I have drifted more towards the right."
- Direction label assigned to joystick deflection made by user ***y*** seconds in the future indicating ground truth label of "***where should I go?***"



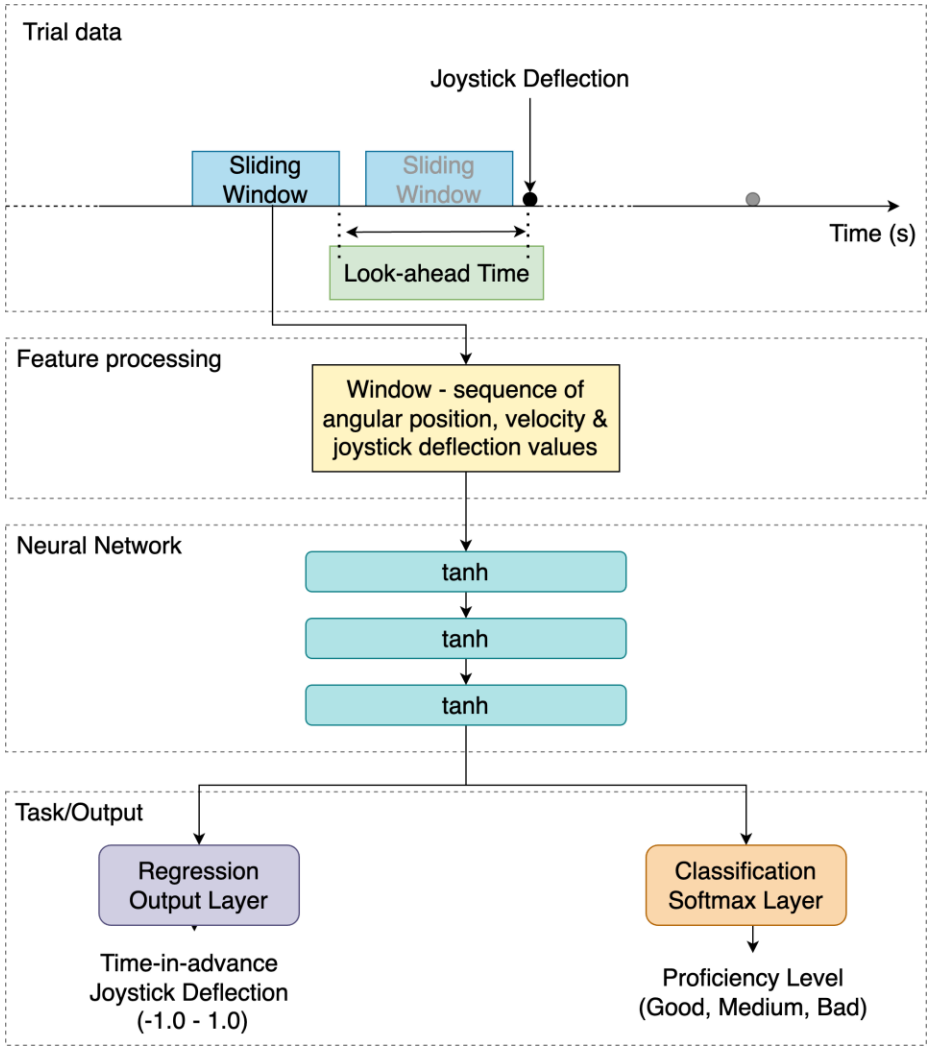
Model Architecture: Joystick-Deflection Prediction Model

- "Where should I go?", numerical model
- Inputs are 1000ms segments of joystick deflections, positions and velocities, and target values are joystick deflections made ***y*** seconds in the future.
- Model should tell how a user should deflect their joystick to balance themselves



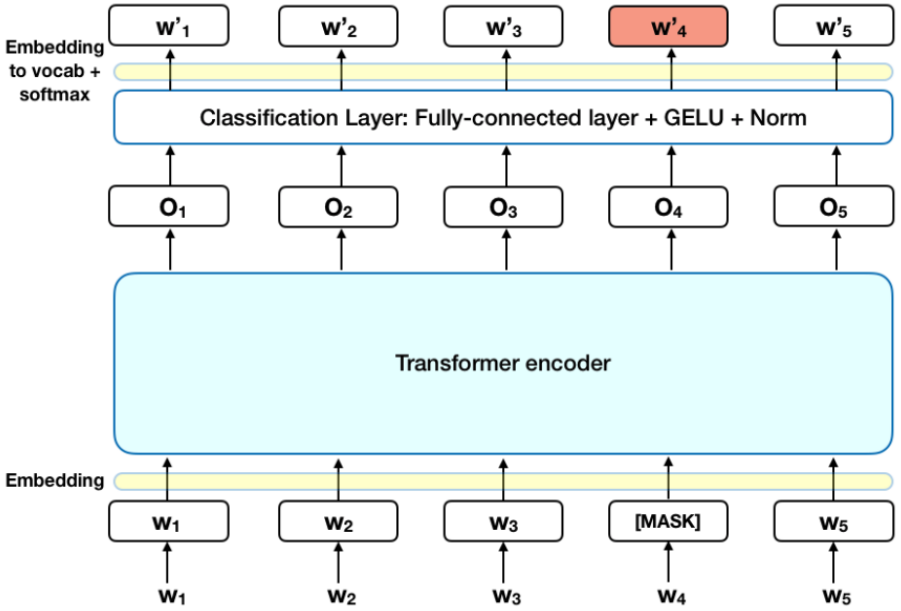
Model Architecture: Performance Proficiency Classifier

- "Where am I?", numerical model
- Need to account how well user is performing the balancing task
- Same inputs as Joystick-Deflection Prediction Model, target labels are discrete proficiency labels of the participant for each sample; ***Proficient, Somewhat Proficient,*** and ***Not Proficient***



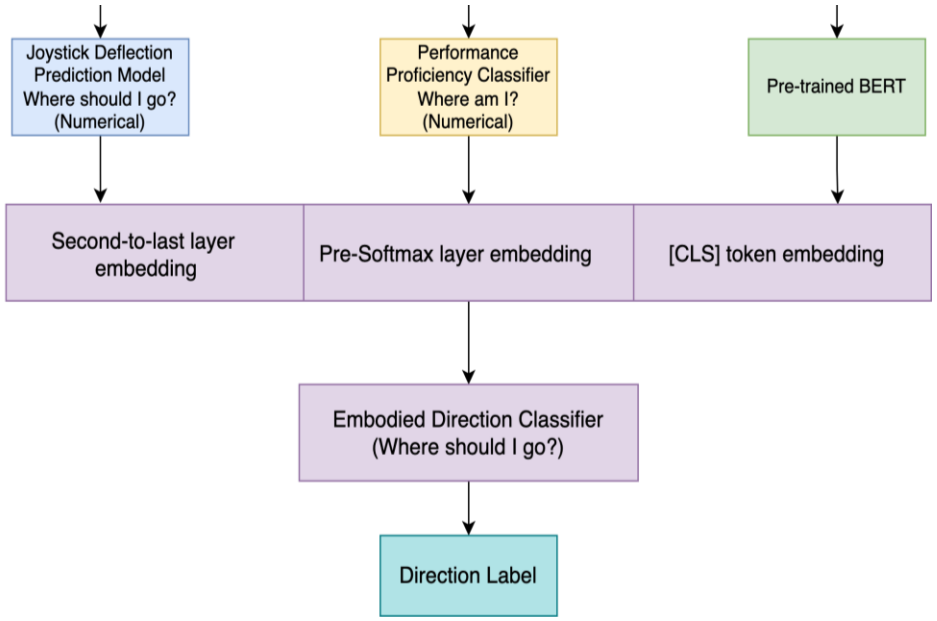
Model Architecture: BERT Sentence Embeddings

- "Where am I?", linguistic representations of position
- Pretrained BERT to extract pooled sentence embedding (the embedding of the [CLS] token) of size 768
- Embeddings extracted for position annotations of each window e.g., "I think I am somewhere in the center" or "I have drifted more towards the right."
- Natural language representation serves as literal "thought vector," representing the "where am I?" grounded positional label input to final classifier



Model Architecture: Embodied Direction Classifier (EDC)

- Final task to ground linguistic representation from BERT embeddings to situated embeddings defined by numerical data models
- Classification model would essentially embody the operational physics of the disorienting balancing task through human performance data, and has grounding annotations of positional language (“where am I?”)
- Input to EDC is three-fold
  - ***Joystick-Deflection Embeddings*** extracted from penultimate layer of the Joystick-Deflection Prediction Model representing what magnitude and direction user should deflect joystick to maintain balance
  - ***Performance Embeddings*** extracted from pre-softmax layer of the Performance Proficiency Classifier representing how well user can gauge their position and direction
  - ***BERT Sentence Embeddings*** for positional thought vectors (“where am I?”) are extracted
- Model would predict the grounded directional label (left, right or center), “where should I go?” for better balance
- EDC would give cues to guide a human participant through linguistic instruction to either deflect the joystick to the left, right, or do nothing (center)



Evaluation

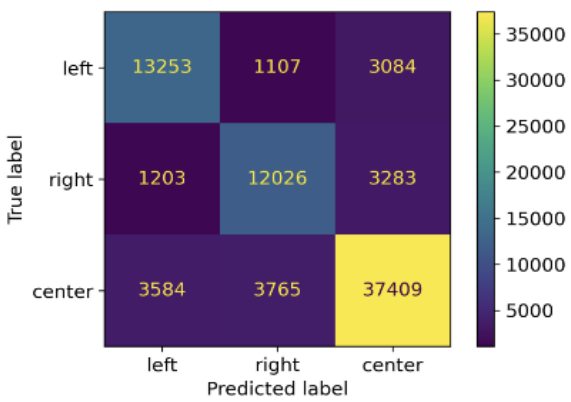
- Randomly selected 12 participants from the trial ~ 4 from each Proficiency group
- Out of 40 trials for each participant; 38 used for train set & 2 for test set
- Sliding window size: 1000ms, look-ahead time: 400ms
- After data processing, ended up with about 1.7 million training samples and 80,000 testing samples; ~95:5 train-test split
- All neural networks have 3 layers (100 units each, tanh activation), trained with Adam optimization for 50,000 epochs
- Joystick-Deflection Prediction Model trained with MSE Loss and both Performance Proficiency Classifier and EDC trained with Cross Entropy Loss and final softmax layer
- Embedding size of 100 for Joystick-Deflection model and Performance Proficiency Classifier each, BERT embedding size is 768

Results

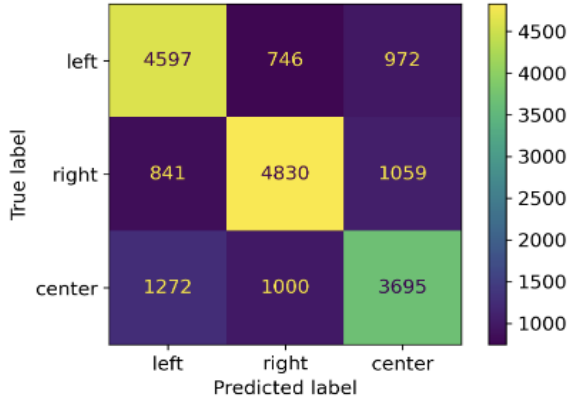
- "Correct" answer: both human and model make/predict the same deflection choice
- EDC's precision, recall, and F1 for target labels, i.e., left, right, and center for each proficiency group
- Bad proficiency group; they think they are in the center, but model thinks otherwise. However, they have better understanding of being in the left or right problem space.
- Better proficiency groups (Medium & Good) have better understandings of where they are, especially in the center.

		Overall	*Bad*	*Medium*	*Good*
Prec.	LEFT	73	69	76	77
	RIGHT	71	73	67	74
	CENTER	85	65	84	91
Rec.	LEFT	76	73	76	80
	RIGHT	73	72	74	73
	CENTER	84	62	81	91
F1	LEFT	75	71	76	78
	RIGHT	72	73	70	73
	CENTER	85	63	82	91
Acc.		80	69	78	87

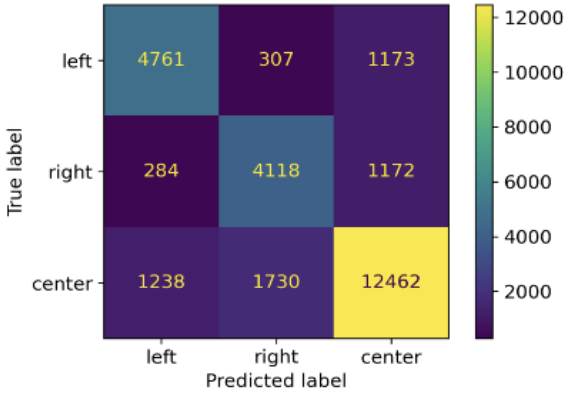
Table 1: EDC performance as %.



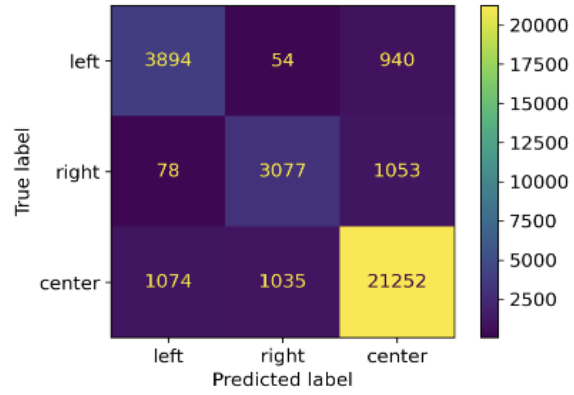
(a) ***Overall***



(b) ***Bad***



(c) ***Medium***



(d) ***Good***

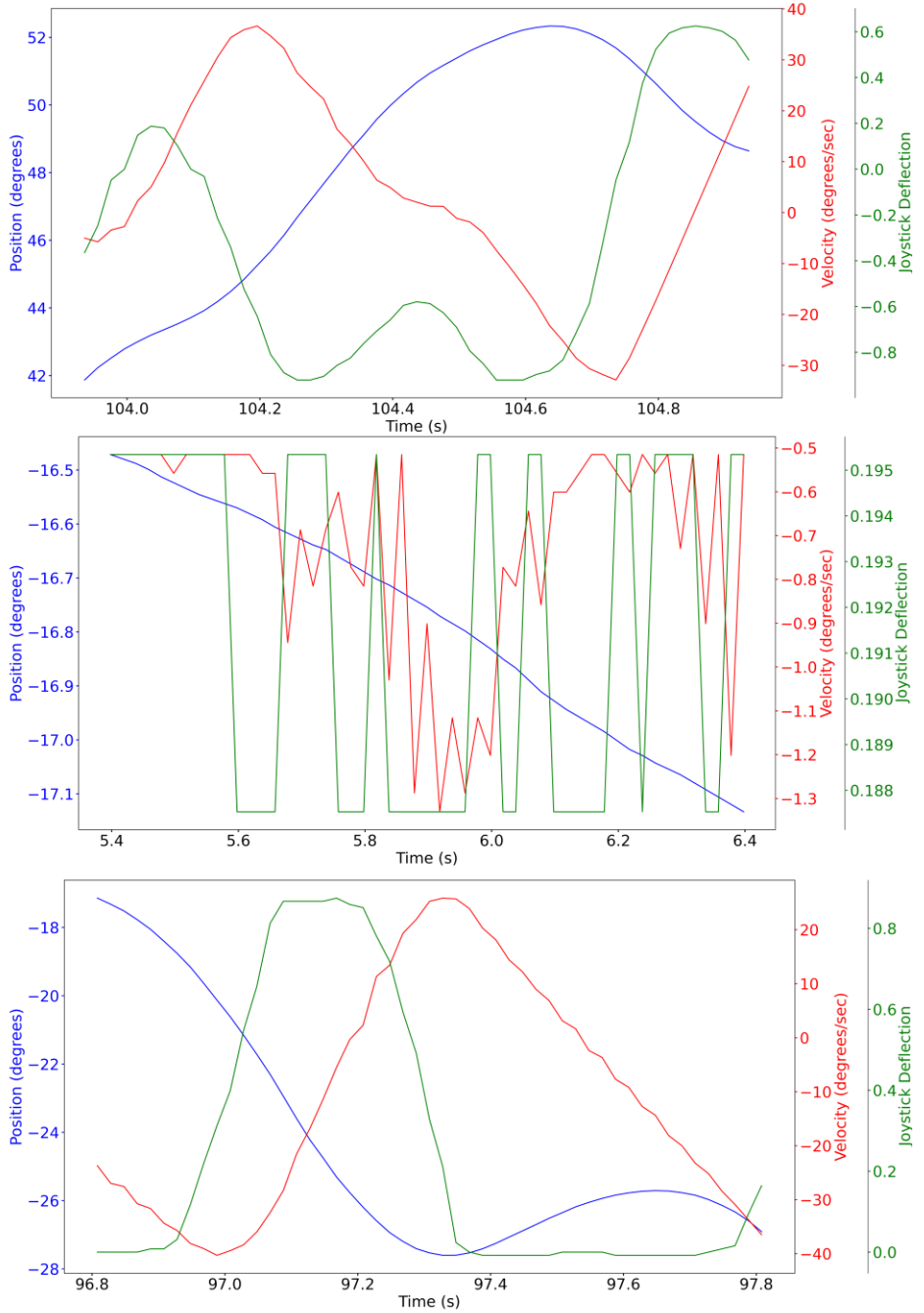
Figure 4: (a) represents the confusion matrix for the full test set of the EDC. (b), (c), and (d) are broken down by proficiency group over the same test set.

Discussion

- Confusion matrices further validate performance of model for each proficiency groups
- Amount of time spent moving left/right (for correctly classified samples); Bad participants ~72%, Medium and Good participants spend an average of 42% and 25% respectively
- EDC model trained on data from all proficiency groups, makes decisions that align, in aggregate, with a Somewhat Proficient participant.

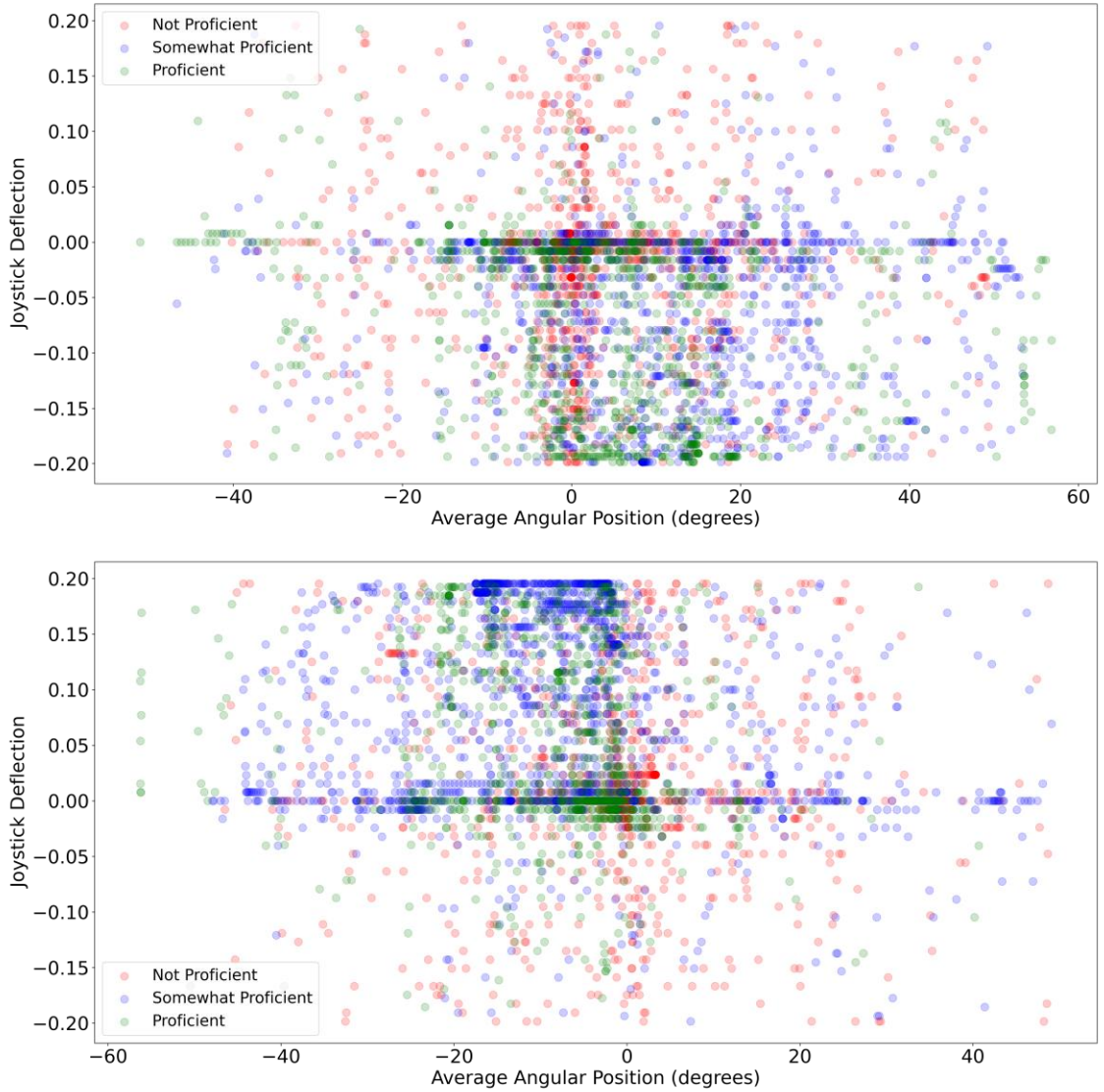
Discussion cont'd

- Fig. 5 - misclassified samples from each proficiency group with truth label center but EDC predicts left/right label; Bad(top) user closer to right crash boundary, Medium(middle) & Good(bottom) users either drifting or in left region
- Model predictions seem more objectively true as it learns better intuitive representations from combination of embodied data and language data from better participants
- Shows that EDC learns a better model of both disoriented balancing task performance and can perform as an in-the-moment guidance tool through language by learning from multiple participants



Discussion cont'd

- Fig. 6 - shows samples labeled center where the human does not move the joystick, but the classifier predicted an optimal movement to the left (top plot) or right (bottom plot).
- Proficient and Somewhat Proficient samples - mostly in center region making slight deflections - model predicts best move is a stronger deflection
- Not Proficient participants - much wider spread of average positions. EDC disagrees with them, demonstrating the ability of the EDC to make objectively "good" decisions in the context of this task.



Conclusions and future work

- Ultimate goal – train AI to provide humans real-time guidance during an embodied task such as the MARS balancing or similar
- Model's apparent mislabels may be more "objectively" correct
- Future work
  - Are there better cues to guide humans other than linguistic cues?
  - Improve situated embodiment with the speed/velocity of MARS i.e., thought vectors representing statements like "too fast" or "in control".
  - Ablation studies to quantify the effect of each type of embedding, especially the role of language
  - Adapt the virtual inverted pendulum environment of Vimal et al. (2020) to facilitate additional high throughput studies with language e.g., subjects call out their perceived direction in real-time, etc.
  - Improve intermediate models using techniques like LSTMs and GRUs to pick up on time-series patterns
  - Train models to provide cues/directions greater than 400ms in future to account for different human reaction times

Thank you!

{sheikh.mannan, nkrishna}@colostate.edu

References

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for
  Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics
- Mark Shelhamer. 2015. Trends in sensorimotor research and countermeasures for exploration-class space flights. Frontiers in Systems Neuroscience, 9:115.
- Patricia S. Cowings, William B. Toscano, Millard F. Reschke, and Addis Tsehay. 2018. Psychophysiological assessment and correction of spatial disorientation during simulated Orion spacecraft re-entry International Journal of Psychophysiology, 131:102–112.
- Vivekanand Pandey Vimal, Han Zheng, Pengyu Hong, Lila N Fakharzadeh, James R Lackner, and Paul DiZio. 2020. Characterizing individual differences in a dynamic stabilization task using machine learning. Aerospace medicine and human performance, 91(6):479–488.
- https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270