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

Sheikh Mannan
Department of Computer Science
Colorado State University
Fort Collins, CO USA
sheikh.mannan@colostate.edu

Nikhil Krishnaswamy
Department of Computer Science
Colorado State University
Fort Collins, CO USA
nkrishna@colostate.edu

Abstract

In this paper, we present an approach toward grounding linguistic positional and directional labels directly to human motions in a disoriented balancing task in a multi-axis rotational device. We use deep neural models to predict human subjects' joystick motions and proficiency in the task. We combine these with BERT embeddings for annotated positional and directional labels into an *embodied direction classifier*. Combining contextualized BERT embeddings with embeddings representing human motion and proficiency can successfully predict the direction a hypothetical human participant should move to achieve better balance. Our accuracy is comparable to a moderately-proficient human subject, and we find that our combined embodied model may actually make objectively better decisions than some humans.

1 Introduction

Much of the recent success in AI can be attributed to the meteoric rise of large language models (LLMs), such as BERT (Devlin et al., 2019) and the GPT family (Radford et al., 2019). These language models facilitate coherent, grammatical text generation using high-dimensional representations of words, sentences, and more, that preserve similarity relations across dimensions. Although pretrained on a enormous amount of text, there are many ways in which they fail to demonstrate "understanding" as commonly defined. As argued by, e.g., Bender and Koller (2020), these models lack knowledge of the current situational context, because that context comes from non-textual modalities. Certain multimodal language models, e.g., multimodal BART-large (Lewis et al., 2020) appear to perform better according to certain benchmarks (Moon et al., 2020; Kottur et al., 2021), but there remain many important domains which for the moment appear to be out of reach for state of the art AI.

Consider the problem of human spatial disorientation. During extreme conditions, such as piloting a spacecraft, even expert humans are subject to gravitational transitions where they may not be able to rely on gravitational cues sensed by the vestibular system, leading to fatal accidents (Shelhamer, 2015; Cowings et al., 2018). Even on Earth, the leading cause of fatal aircraft accidents in military pilots is spatial disorientation (Gibb et al., 2011).

Numerical AI models, however, with direct access to quantitative information about position and movement, can potentially determine when a human appears to be losing control and intervene, such as by telling the human what to do in order to right themselves. A successful AI partner that counteracts human disorientation to enhance task performance in real time would need to predict the intent of the human's motions, make decisions with incomplete information or under environmental uncertainty (Weber, 1987; Talamadupula et al., 2010) and, perhaps most importantly, foster trust in the human (Hengstler et al., 2016).

These are not requirements that even the impressive benchmark performance of modern LLMs can meet. Successful guidance of a human through language requires that the AI "embody" relations between linguistic terms and the human's situation.

In this paper we combine disambiguated and contextualized linguistic embeddings (Wiedemann et al., 2019) from BERT, with embeddings extracted from numerical AI models that are trained to predict control movements and human performance in a spaceflight-analog disoriented balancing task. Unlike the BERT embeddings, these latter embeddings are "situated," in that they come from models that are trained to *embody* a human participant's position in a phase space parameterized by angular position and velocity in the balancing task. This combined model is trained to predict the direction the human should move towards for

Proceedings of the 2022 CLASP Conference on (Dis)embodiment, pages 70–79
September 15–16, 2022. ©2022 Association for Computational Linguistics

better balance given BERT embeddings that represent "thought vectors" about position relative to the balance point, and performance and motion control features extracted from the numerical models. We show that predictions made by our model "agree" on average with those made by a human with a moderate level of proficiency in the balancing task, and a deeper dive into misclassifications suggest that the model may actually be performing better in this task than the raw numerical results indicate.

2 Related Work

This paper brings together research in two distinct and to date largely disjunct areas: multimodal language grounding through human-AI collaboration, and mitigating the effects of spatial disorientation. This section discusses relevant work in these two domains and our goals in synthesizing them.

The Collaborative Research Center's Situated Artificial Communicator project was a significant early attempt to model the integration of language and sensorimotor skills in a situated context (Rickheit and Wachsmuth, 2006). Recent work in multimodal conversational modeling has continued similar lines of research with multimodal Transformer architectures (Chen et al., 2020; Hu et al., 2020). Other relatively recent work attempts to integrate neurally-encoded robotic arm control with guidance and instruction through dialogue (She et al., 2014; She and Chai, 2017).

Alomari et al. (2017a) use unsupervised learning for concepts such as colors, names and activities by an autonomous robot. Alomari et al. (2017b) combine PCFG trees and visual feature clustering to ground video depictions of actions to linguistic labels. Ilinykh and Dobnik (2022) find that language models in a multimodal task setting learn different semantic information about objects and relations crossmodally and unimodally (text-only).

Importantly, though, these lines of research subsume all grounding and multimodality under combinations of language and *vision*, to the exclusion of other channels, and where AI and humans interact, the interaction focuses on humans guiding AI, not AI assisting humans. Our work brings in modal channels directly related to human motion in a situated environment, to train an AI that ultimately assists humans to mitigate spatial disorientation.

While there is a wide and varied body of research from the neuroscience and biomechanics communities on other modal information channels, such as human spatial awareness, AI has largely not been applied here.

Rupert (2000) presents a tactile stimulation system that provides intuitive orientation information to aircrew and operators of remote platforms and is compatible with a pilot's natural sensory system. Intelligent control of such a system could help provide pilots with appropriate cues in disorienting situations, but only if human proclivites in such situations are well-understood and modeled.

Vimal et al. (2016) use a multi-axis rotation system (MARS) device programmed with inverted pendulum dynamics to investigate learning in a dynamic balancing task about an unstable equilibrium point. Subjects attempt to remain balanced by applying joystick deflections to control the motion of the device, and the authors find that subjects improve their performance by making fewer destabilizing joystick movements, and more persistent short-term joystick movements intermittently. Later, they further investigate learning about different roll planes (vertical, horizontal) that disrupt the natural orientational capabilities of humans, combined with the role of gravitationally-dependent otolith and somatosensory cues in the learning of the balancing task (Vimal, 2017; Vimal et al., 2017, 2018, 2019, 2022). They find that absence of gravitationally-dependent otolith and somatosensory cues degrades balancing performance. However, their findings also indicate that balance control can be enhanced in situations lacking gravitationally dependent position cues as in weightlessness, when initial training occurs with such cues present. They also observe that some participants re-learn how to balance themselves in the disorienting condition, demonstrating learning, while others do not. Data from this line of research is used in this paper.

Recent work in this line of research has begun to use machine learning and AI techniques, providing a path forward to integrate the two aforementioned broad areas. Vimal et al. (2020) group subjects performing the balancing task in the horizontal roll plane (HRP), without any gravitational cues, into performance proficiency categories using a Bayesian Gaussian Mixture model. Wang et al. (2022) use the same data to train a stacked gated recurrent unit (GRU) model to predict the occurrence of crashes (where crash boundaries are set to ±60° from the balance point) 800ms in advance. Our work extends this line of research toward modeling human behavior in the balancing task so that AI can predict and counteract disorientation.

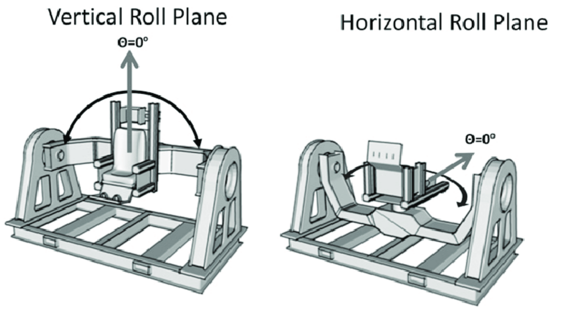


Figure 1: The multi-axis rotation system (MARS), programmed with inverted pendulum dynamics, in the vertical roll axis (left) and the horizontal roll axis (right). Straight grey arrows represent the direction of balance (DOB). Placing participants in the horizontal roll plane disrupts normal gravitational cues, making the balancing task disoriented (figure courtesy of Vimal et al. (2020)).

3 Dataset

We use data and performance proficiency labels from Vimal et al. (2020) which are further explained below. Additionally, we further annotate the data with grounded positional annotations and directional labels for training an embodied AI classifier that predicts optimal direction of movement.

3.1 MARS Data

The data is collected from 34 consenting healthy adult participants (18 females and 16 males, $\mu \approx$ 20.4 years old, $\sigma \approx$ 2.0 years) with no prior experience in the Multi-Axis Rotation System (MARS).

The MARS was programmed with inverted pendulum dynamics about a horizontal roll axis as shown in Fig. 1 and controlled by a joystick. MARS dynamics were governed by the equation, $\ddot{\theta} = k_P sin\theta$, where θ is the angular deviation from the direction of balance (DOB) in degrees, and k_P is the pendulum constant. Here, a pendulum constant of $600° \cdot s^{-2}$ ($\approx$ 0.52Hz) was used. Crash limits restricted the angular range of the MARS to $\pm$ 60° from the DOB. Angular velocity was limited to $\pm$ $300° \cdot s^{-1}$, and angular acceleration to $\pm$ $180° \cdot s^{-2}$. Every $\sim$ 0.02s, a velocity increment proportional to the joystick deflection was added to the MARS velocity and computed by a Runge-Kutta RK4 solver to calculate the new MARS angular position and velocity. The latency between a joystick deflection and a change in MARS angular velocity was 30ms over the observed range of MARS spectral power of 0 to $\sim$0.75Hz (further experimental details in Vimal et al. (2020)).

Fig. 2 shows a segment of trial data from a representative participant showing changes in angular position (blue), angular velocity (red) and joystick deflection (green). We can see that this participant was able to just barely avert a crash as the MARS angular position reached +50°, or 10° from the crash boundary.

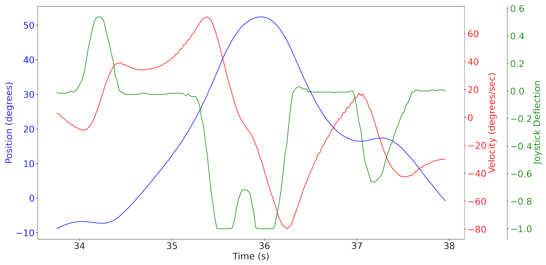


Figure 2: A segment of trial data from a medium proficiency participant showing angular position (blue), angular velocity (red) and joystick deflection (green). The participant just barely prevents a crash as the MARS angular increases to +50° from DOB.

3.2 Proficiency Labels

Vimal et al. (2020) clusters participants based on their balancing performance using various engineered features, such as:

- **Crash frequency** equals the number of crashes in a trial divided by the trial duration. Higher values correlated with poorer balancing performance. Proficient participants had a mean crash frequency of 0.002Hz and not proficient participants had a mean crash frequency of 0.11Hz.
- **Anticipatory joystick deflections** are those that removed energy from the MARS by decelerating it as it was moving toward the DOB. Anticipatory joystick deflections can help stabilize the MARS; they are often used when poor control leads to high velocities near the balance point. As participants learn to stabilize the MARS the percentage of anticipatory joystick deflections decreases. 0.2% of proficient participants' deflections were classified as anticipatory while not proficient participants used this strategy 14% of the time.
- **Destabilizing joystick deflections** accelerate the MARS away from the DOB. Proficient participants made destabilizing deflections on average 0.0005% of the time and non-proficient participants made them 4.8% of the time.

Vimal et al. (2020) trained a Bayesian Gaussian Mixture model using these features that clustered participants into three distinct groups ***Proficient*** (or "Good"), ***Somewhat Proficient*** (or "Medium"), and ***Not Proficient*** (or "Bad") based on their balancing performance. Participants were clustered based on their performance after 2 days of trials, meaning that some proficient participants demonstrated substantial learning in the task over the successive trials, and occasionally some non-proficient participants' performance actually became worse with repetition. These are the same per-participant proficiency labels we use here.

3.3 Positional & Direction Labels

To ground the situated numerical features from the MARS to a linguistic representation, we annotate the numerical features with sentences that represent position relative to the DOB, or simply put, with possible answers to the question "where am I?" given the numerical features. For example, if they are far off to the right of the DOB, a human may think "I have drifted more towards the right" or if they think they are balanced near the DOB the equivalent thought may be "I think I am somewhere in the center". These sentence annotations were generated by third-party annotators for each of the three regions; ***left*** ($< -20°$ from the DOB), ***right*** ($> +20°$ from the DOB), and ***center*** (within $\pm 20°$ of the DOB), within a total possible range of $\pm 60°$.

For the direction labels, representing the direction towards which the human should move the MARS (or deflect the joystick) for better balance about the DOB or "where should I go?", we again divide it into three categories; ***left***: deflect the joystick with such amplitude that it prompts the MARS to the left, ***right***: deflect the joystick with such amplitude that it prompts the MARS to the right, and ***center***: deflect the joystick with as little amplitude as possible such that there is little to no change in the position of the MARS. These are discrete, one-hot vectors depicting the "where I should be going" grounded label, and are assigned using the joystick deflection made after the look-ahead time. The direction labels are defined as *left*: < -0.2, *right*: $> +0.2$, and *center*: between -0.2 and $+0.2$. +1 and -1 represent full deflection.

4 Methodology

Our goal is to combine representations of motion and performance proficiency, which are learned from data directly capturing human embodiment during the MARS balancing task, with linguistic representations of the position and directional concepts involved. A successful model is one which can predict the label for the best direction of motion given the current circumstances by learning correlations between motion, proficiency, and linguistic representation.

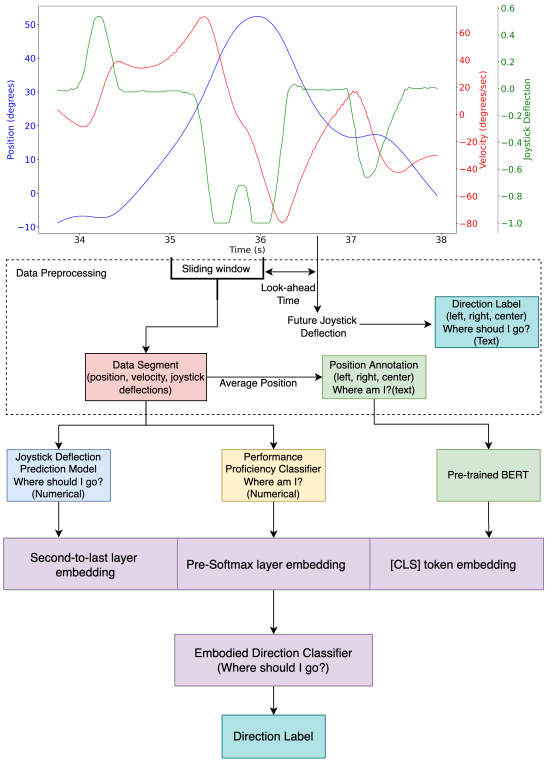

50
40
30
20
10
0
−10
Position (degrees)
60
40
20
0
−20
−40
−60
−80
Velocity (degrees/sec)
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
Joystick Deflection
34
35
36
37
38
Time (s)
Data Preprocessing
Sliding window
Look-ahead Time
Future Joystick Deflection
Direction Label (left, right, center) Where shoud I go? (Text)
Data Segment (position, velocity, joystick deflections)
Average Position
Position Annotation (left, right, center) Where am I?(text)
Joystick Deflection Prediction Model Where should I go? (Numerical)
Performance Proficiency Classifier Where am I? (Numerical)
Pre-trained BERT
Second-to-last layer embedding
Pre-Softmax layer embedding
[CLS] token embedding
Embodied Direction Classifier (Where should I go?)
Direction Label

Figure 3: Overview of the embodied model architecture.

The model architecture, shown in Fig. 3, can be divided into five parts: (1) **data preprocessing**; (2) a **joystick-deflection predictor** of immediate future action; (3) a **performance proficiency classifier**, which provides a high-level view of the subject's task performance; (4) **BERT annotation embeddings**, which provide real-valued semantic representations that the outputs of previous two modules are correlated to, and (5) the combined model, or **embodied direction classifier** (EDC).

4.1 Data Preprocessing

For each trial in the data, we use a fixed sliding window technique to extract segments of joystick deflections, angular velocity and position where the user was in control and no crashes occurred for the given look-ahead time y seconds in the future.

For each viable window extracted, we assign a random sentence annotation for the 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."

The processed data has two parts for each sample, (1) the MARS machine features i.e. joystick deflections, position and velocity and (2) the grounded position annotations.

4.2 Joystick-Deflection Prediction Model

Using the processed data on angular position, angular velocity and joystick deflections, we train a deep feedforward neural network model (see Sec. 5 for hyperparameters) to predict how much the joystick should be deflected to keep the user balanced. Inputs are the 1000ms segments of joystick deflections, positions and velocities, and target values are the joystick deflections made 400ms in the future. Essentially, once operationalized, this model should tell how a user should deflect their joystick to balance themselves[1].

4.3 Performance Proficiency Classifier

To account for how well a user is performing the balancing task, we build a neural performance classifier that is able to tell us the user's ability to discern and gauge where they are in terms of position and where they should go. The proficiency labels are obtained from Vimal et al. (2020) (described in Sec. 3.2). We train a deep feedforward neural network model (see Sec. 5 for hyperparameters) using the same inputs as those to the Joystick-Deflection Prediction Model (Sec. 4.2). However, here the target labels are discrete proficiency labels of the participant for each sample in turn; ***Proficient***, ***Somewhat Proficient***, and ***Not Proficient***. This model should output a proficiency label for each segment, reflecting how proficient the participant is behaving at that time. The final pre-classification layer of this model outputs embeddings that are situated within the task phase space of the task by preserving high-dimensional similarity relations between actual direction and velocity values and task proficiency.

4.4 BERT Sentence Embeddings

We use pretrained BERT to produce the pooled sentence embedding (the embedding of the [CLS] token) for the the position annotations for each sample. This natural language representation serves as a rather literal "thought vector," representing the "where am I?" grounded positional label input to our embodied directional classifier.

4.5 Embodied Direction Classifier

Our task is now to take the numerical models learned from embodied human performance, and the linguistic representations from BERT, and train a model, the embodied direction classifier, that grounds the linguistic representation to circumstances described by the numerical data.

We combine the three aforementioned models and build a classification model that has essentially embodied the operational physics of the disorienting balancing task through human performance data, and has grounding annotations of positional language ("where am I?"). This classifier takes these inputs to predict the grounded directional label, "where should I go?" for better balance.

Input to the EDC is three-fold. **Joystick-Deflection Embeddings** are extracted for each sample from the penultimate layer of the Joystick-Deflection Prediction Model. These vector embeddings represent how much and in which direction the user should deflect their joystick to maintain balance. **Performance Embeddings** are also extracted from the pre-softmax layer of the Performance Proficiency Classifier to represent how well the user can gauge their position and direction. Finally, the **BERT Sentence Embeddings** for the positional thought vectors are extracted. For each sample, these three vector embeddings are concatenated and passed to the model.

The EDC is trained to predict the grounded directional labels, i.e., *left*, *right*, and *center*, which represent the "where should I go?" aspect in the balancing task. In operation, this would be a cue to a guide a human participant through linguistic instruction to either deflect to the left, deflect to the right, or do nothing with the joystick. Here we simply assess the performance of the model and how it compares to humans.

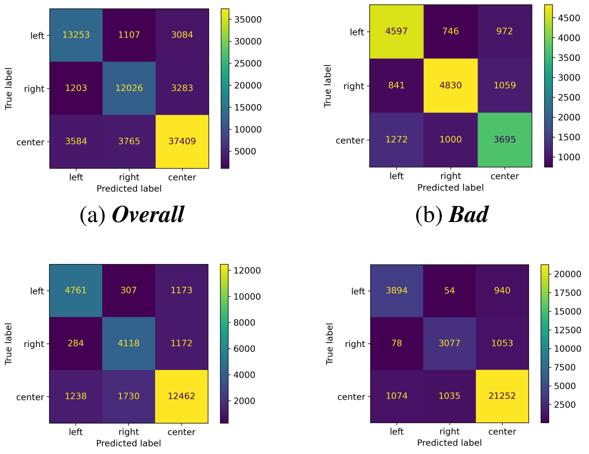
5 Evaluation

We randomly selected 12 participants from the dataset—4 participants of each proficiency. We use 38 of each participant's 40 trials for the train set and 2 for the test set. As described in Sec. 4.1, we use a sliding window of 1000ms and a look-ahead time of 400ms. After data processing, we end up with about 1.7 million training samples and 80,000 testing samples, for a ∼95:5 train-test split.

All neural networks have 3 layers (100 units each, $tanh$ activation), and are trained with Adam optimization for 50,000 epochs. The Joystick-Deflection Prediction Model was trained with MSE Loss and both the Performance Proficiency Classifier and EDC were trained with Cross Entropy Loss and a final softmax layer. To evaluate the performance/competence of the EDC we examine:

1. How well the model performs on average and for each proficiency group.
2. Misclassified samples where the model "disagrees" with the apparent ground truth, or the decision the human participant had made.

[1] 400ms is slightly below the reaction time of average humans (Nagler and Nagler, 1973) and well above the reaction time of trained pilots (Binias et al., 2020).



(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.

6 Results

Table 1 illustrates the performance of the EDC overall and for each of the three proficiency groups. We also show the EDC's precision, recall, and F1 for the three target labels, i.e., *left*, *right*, and *center*. Here a "correct" answer is one where the human participant made the correct movement choice with respect to their angular position and velocity, and the model predicted the same movement choice.

		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 %.

7 Discussion

7.1 Proficiency Breakdown

In Table 1, we can see that the EDC's performance increases as the proficiency of the participant increases. We see that the Bad proficiency group shows lower performance on correctly grounding the center label, i.e., these participants think they are in the center region, but the model thinks otherwise. They do appear to have a better understanding of whether they are in the left or right region and balance themselves accordingly. The Medium & Good proficiency groups have a better

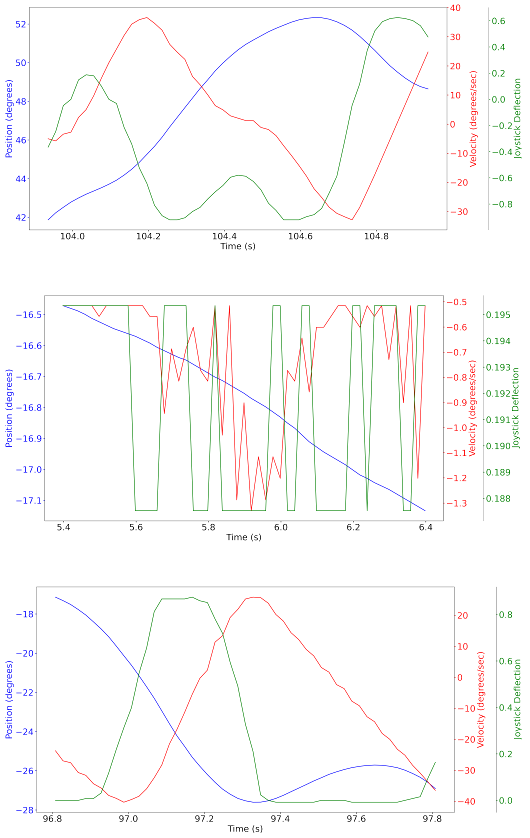


Figure 5: Misclassified test samples from each proficiency group (following conventions from Fig. 2). Top: Bad participant in the right region, truth label *center*, predicted label *left*. Middle: Medium participant drifting toward left region, truth label of *center*, predicted label *right*. Bottom: Good participant in the left region, truth label *center*, predicted label *right*.

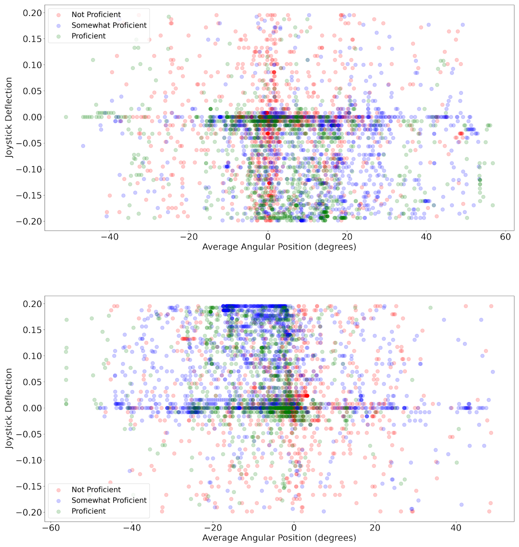


Figure 6: Misclassified test samples where the ground truth labels were center but predicted as *left* (top) and *right* (bottom), showing the spread of actual joystick deflection vs. sample average position when the EDC "disagrees" with the participant's movement.

understanding of where they are in the problem space than the Bad group, especially when the participants think they are in the center region. For the Good proficiency group, we see that the EDC had an F1 score of 91% for the center label, which means that the model agrees with their decision to do nothing drastic when they are in the center region roughly 91% of the time. This is likely due in part to the fact that many Good (or proficient) participants are able to remain balanced within the center region for most of their trials.

Fig. 4 provides a deeper insight into the what kinds of samples are commonly confused with each other by the EDC. Regardless of proficiency group, the center labels is more often misclassified as left or right than the reverse. This is likely due in part to there being more center labels in the dataset overall (due to Medium and especially Good participants successfully keeping themselves balanced), however the confusion matrices further validate the performance of the model for each of the three proficiency groups: the Bad group has the most confusions and the Good group has the least. The EDC is able to combine the embodied numerical and language representation channels and determine that when a person is in the central region, they should not attempt to move out of it.

Bad participants, meanwhile, are all over the place, and spend ~72% of the time moving either left or right (for the correctly classified samples) whereas Medium and Good participants spend an average of 42% and 25% of their time, respectively, moving left or right. The rest of the time is spent making slight, intermittent movements to remain in the center. They do better at avoiding destabilizing deflections, which the EDC picks up and outputs as directional labels that describe doing just that. Our model, which is trained on data from all proficiency groups, makes decisions that align, in aggregate, with those of a Somewhat Proficient participant.

7.2 Analysis of Misclassified Labels

While the overall metrics for the EDC's performance are promising, and it performs particularly strongly on Good participants, those numbers do not tell the whole story. Fig. 5 shows one sample from each proficiency group that have a ground truth label of *center* but are predicted as *left* or *right* by the model. Fig. 5 (top) shows a participant from the Bad proficiency group positioned in the right region, closer to the crash boundary, velocity increasing as they deflect the joystick to the right as well (a destabilizing joystick deflection). The truth label here is *center* as the participant does not move the joystick for 400ms after the end of this sample, but the model predicts that the participant should deflect to the left, which appears to be objectively more correct than the "ground truth" label is. Therefore the training data itself may actually include noise introduced by subpar participants' suboptimal movements, but the EDC is actually able to learn better intuitive representations from the combination of embodied data and language data from better participants. Fig. 5 (middle and bottom) shows that participants from the Medium and Good proficiency groups respectively, are also occasionally prone to the same situations faced by the participant in top sample, and sometimes make mistakes. Here, the Medium and Good participants are both either in or moving closer to the left region and classifier predicts that the participant should deflect to the right, despite a ground truth label of *center*. This shows that the EDC does learn a better model of both disoriented balancing task performance and in-the-moment guidance through language by learning from multiple participants. If the model were reevaluated against expert/commonsense judgments of optimal human actions, the metrics in Table 1 could rise substantially. In addition, by accurately predicting subpar actions, the EDC may be used to guard against them.

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). The graphs themselves show the joystick deflection on the Y-axis vs. sample average position on the X-axis. In Fig. 6 (top), many samples are clustered just right of center with joystick deflection to the left (bottom part of the plot). The opposite is true for the bottom plot, with deflections clustered right of center while average position is just left of the DOB.

If we examine these plots by participant proficiency, the Proficient and Somewhat Proficient samples remain mostly in the center region, close to the DOB. These participants make slight joystick deflections to remain within 20° of the DOB, but the model predicts that the best move is a stronger deflection in one direction. These may be cases where the participant is technically within the center region but perhaps close to a left/right boundary. The Not Proficient participants have a much wider spread of average positions where they make close to no deflection of the joystick. The EDC disagrees with them, demonstrating both the noise in the data when non-proficient participants' actions are taken as ground truth, and the ability of the EDC, despite this, to make objectively "good" decisions in the context of this task. The numerical performance of the model (Table 1) goes up as participant proficiency goes up, but in fact this reveals that the model is already able to make objectively good decisions, and as human performance improves and participants get better at balancing and become more likely to remain in the center region or recover from drifts, the human decisions are more likely to match these. This suggests that a combined embodied-linguistic method as demonstrated here may be suitable for guiding humans in such a task in real time. The EDC appears to actually display some understanding of the correlation between position and velocity in the problem space, and discrete directional labels.

8 Conclusion

The ultimate goal of this work is to train an AI model that can give guiding cues to a human participant in real time to improve their performance in an embodied task such as the MARS balancing or similar. Successful guidance of a human through language requires that the AI "embody" the relation between linguistic terms and the situation inhabited by the human. Here we have presented evidence that an AI model can be trained to ground directional labels to embedding-level representations of angular position and velocity, and can do so in a way that is sensitive to the proficiency level of a participant in this task, if that information is provided as input. These grounded labels can serve as cues to a human participant, as the AI considers the situation and answers "where am I?" with an answer to "where should I go?" (e.g., "I am drifting to the left. I should deflect more to the right.").

Our model, EDC, trained on data from participants of all proficiencies, displays apparent performance on par with a Somewhat Proficient participant, but a deeper dive into misclassifications reveals that even though the training data itself is noisy, as the ground truth is taken to be the actual actions of the participants, even non-proficient ones, our model's apparent mislabels may actually be better decisions than those of study participants.

8.1 Future Work

Given the nature of the task and the need for immediate response by humans, is a linguistic cue really the best cue to use in this case? While disoriented, humans may not respond as quickly to language cues; perhaps visual or vibrotactile cues are more apt for prompting faster responses. Further experiments need to be carried out in real time human-AI collaboration in this task (e.g., what kind of AI cues help humans perform better?). Nonetheless, the language input seems to be important to the model for predicting directional guidance, regardless of how that guidance is ultimately expressed. Another feature that could improve our situated embodied model is speed of the MARS, i.e., adding thought vectors representing things like "too fast" or "in control" to positional thought vectors could bolster the combined model's effectiveness as a countermeasure to disorientation by factoring in gradations for things like speed or amount of deflection, which would be important for actually guiding humans in the MARS task where continuous joystick deflecton is being applied.

In future work, we plan ablation studies to quantify the effect of each type of embedding, in particular the precise role of language. By taking the existing sentence annotations and automatically transforming them into alternate phrasings (e.g., "I think I am somewhere in the center" → "I *am* somewhere in the center"), we can quantify the differences in sentence and contextualized word embeddings, and the resultant predictive power of the EDC. We are also adapting the virtual inverted pendulum environment of Vimal et al. (2020) to facilitate additional high-throughput studies where we can experiment further with language, e.g., by having subjects call out their perceived direction in real time, or having other trained humans give a subject real-time linguistic guidance. The intermediate models themselves—the joystick-deflection predictor and proficiency classifier—can be improved using techniques like LSTMs and GRUs to pick up on time-series patterns. Furthermore, to be an effective partner for an average human, our models would need to be trained to predict directions for lookahead times greater than 400ms to account for different human reaction times.

Acknowledgements

Thanks to Vivekanand Pandey Vimal, Paul DiZio, and James R. Lackner for thoughts, discussion, and providing access to the original data. Thanks to our reviewers for their helpful comments on the paper.

References

Muhannad Alomari, Paul Duckworth, Nils Bore, Majd Hawasly, David C Hogg, and Anthony G Cohn. 2017a. Grounding of human environments and activities for autonomous robots. In *IJCAI-17 Proceedings*, pages 1395–1402. Lawrence Erlbaum Associates, Inc.

Muhannad Alomari, Paul Duckworth, David Hogg, and Anthony Cohn. 2017b. Natural language acquisition and grounding for embodied robotic systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Emily M Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198.

Bartosz Binias, Dariusz Myszor, Henryk Palus, and Krzysztof A Cyran. 2020. Prediction of pilot's reaction time based on EEG signals. *Frontiers in neuroinformatics*, 14:6.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer.

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.

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.

Randy Gibb, Bill Ercoline, and Lauren Scharff. 2011. Spatial disorientation: decades of pilot fatalities. *Aviation, space, and environmental medicine*, 82(7):717–724.

Monika Hengstler, Ellen Enkel, and Selina Duelli. 2016. Applied artificial intelligence and trust–The case of autonomous vehicles and medical assistance devices. *Technological Forecasting and Social Change*, 105:105–120.

Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. 2020. Iterative answer prediction with pointer-augmented multimodal transformers for TextVQA. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9992–10002.

Nikolai Ilinykh and Simon Dobnik. 2022. Attention as Grounding: Exploring Textual and Cross-Modal Attention on Entities and Relations in Language-and-Vision Transformer. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4062–4073.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. SIMMC 2.0: A Task-oriented Dialog Dataset for Immersive Multimodal Conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4903–4912.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Seungwhan Moon, Satwik Kottur, Paul A Crook, Ankita De, Shivani Poddar, Theodore Levin, David Whitney, Daniel Difranco, Ahmad Beirami, Eunjoon Cho, et al. 2020. Situated and Interactive Multimodal Conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1103–1121.

Charles Arthur Nagler and William Merle Nagler. 1973. Reaction time measurements. *Forensic science*, 2:261–274.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Gert Rickheit and Ipke Wachsmuth. 2006. *Situated communication*. Mouton de Gruyter.

Angus H Rupert. 2000. An instrumentation solution for reducing spatial disorientation mishaps. *IEEE Engineering in Medicine and Biology Magazine*, 19(2):71–80.

Lanbo She and Joyce Chai. 2017. Interactive learning of grounded verb semantics towards human-robot communication. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1634–1644.

Lanbo She, Shaohua Yang, Yu Cheng, Yunyi Jia, Joyce Chai, and Ning Xi. 2014. Back to the blocks world: Learning new actions through situated human-robot dialogue. In *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, pages 89–97.

Mark Shelhamer. 2015. Trends in sensorimotor research and countermeasures for exploration-class space flights. *Frontiers in Systems Neuroscience*, 9:115.

Kartik Talamadupula, J Benton, Subbarao Kambhampati, Paul Schermerhorn, and Matthias Scheutz. 2010. Planning for human-robot teaming in open worlds. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 1(2):1–24.

Vivekanand Pandey Vimal. 2017. *The Role of Gravitational Cues in the Learning of Balance Control*. Brandeis University.

Vivekanand Pandey Vimal, Paul DiZio, and James R Lackner. 2017. Learning dynamic balancing in the roll plane with and without gravitational cues. *Experimental brain research*, 235(11):3495–3503.

Vivekanand Pandey Vimal, Paul DiZio, and James R Lackner. 2019. Learning and long-term retention of dynamic self-stabilization skills. *Experimental brain research*, 237(11):2775–2787.

Vivekanand Pandey Vimal, Paul DiZio, and James R Lackner. 2022. The role of spatial acuity in a dynamic balancing task without gravitational cues. *Experimental brain research*, 240(1):123–133.

Vivekanand Pandey Vimal, James R Lackner, and Paul DiZio. 2016. Learning dynamic control of body roll orientation. *Experimental brain research*, 234(2):483–492.

Vivekanand Pandey Vimal, James R Lackner, and Paul DiZio. 2018. Learning dynamic control of body yaw orientation. *Experimental brain research*, 236(5):1321–1330.

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.

Yonglin Wang, Jie Tang, Vivekanand Pandey Vimal, James R Lackner, Paul DiZio, and Pengyu Hong. 2022. Crash prediction using deep learning in a disorienting spaceflight analog balancing task. *Frontiers in physiology*, page 51.

Martin Weber. 1987. Decision making with incomplete information. *European journal of operational research*, 28(1):44–57.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings.