Beyond "Taming Electric Scooters": Disentangling Understandings of Micromobility Naturalistic Riding

MAHAN TABATABAIE, University of Connecticut, USA
SUINING HE*, University of Connecticut, USA
HAO WANG, University of Connecticut, USA
KANG G. SHIN, University of Michigan–Ann Arbor, USA

Electric(e)-scooters have emerged as a popular, ubiquitous, and first/last-mile micromobility transportation option within and across many cities worldwide. With the increasing situation-awareness and on-board computational capability, such intelligent micromobility has become a critical means of understanding the rider's interactions with other traffic constituents (called *Rider-to-X Interactions*, RXIs), such as pedestrians, cars, and other micromobility vehicles, as well as road environments, including curbs, road infrastructures, and traffic signs. How to interpret these complex, dynamic, and context-dependent RXIs, particularly for the rider-centric understandings across different data modalities — such as visual, behavioral, and textual data — is essential for enabling safer and more comfortable micromobility riding experience and the greater good of urban transportation networks.

Under a *naturalistic riding* setting (i.e., without any unnatural constraint on rider's decision-making and maneuvering), we have designed, implemented, and evaluated a pilot **C**ross-modality **E**-scooter **N**aturalistic **R**iding **U**nderstanding **S**ystem, namely CENRUS, from a human-centered AI perspective. We have conducted an extensive study with CENRUS in sensing, analyzing, and understanding the behavioral, visual, and textual annotation data of RXIs during naturalistic riding. We have also designed a novel, efficient, and usable disentanglement mechanism to conceptualize and understand the e-scooter naturalistic riding processes, and conducted extensive human-centered AI model studies. We have performed multiple downstream tasks enabled by the core model within CENRUS to derive the human-centered AI understandings and insights of complex RXIs, showcasing such downstream tasks as efficient information retrieval and scene understanding. CENRUS can serve as a foundational system for safe and easy-to-use micromobility rider assistance as well as accountable use of micromobility vehicles.

CCS Concepts: • **Human-centered computing** → *Ubiquitous and mobile computing*.

ACM Reference Format:
Mahan Tabatabaie, Suining He, Hao Wang, and Kang G. Shin. 2024. Beyond "Taming Electric Scooters": Disentangling Understandings of Micromobility Naturalistic Riding. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 3, Article 129 (September 2024), 24 pages. https://doi.org/10.1145/3678513

*Corresponding author.

Authors' addresses: Mahan Tabatabaie, Department of Computer Science and Engineering, University of Connecticut, Storrs, CT, USA, mahan.tabatabaie@uconn.edu; Suining He, Department of Computer Science and Engineering, University of Connecticut, Storrs, CT, USA, suining.he@uconn.edu; Hao Wang, Department of Computer Science and Engineering, University of Connecticut, Storrs, CT, USA, hao.3.wang@uconn.edu; Kang G. Shin, Department of Electrical Engineering and Computer Science, University of Michigan–Ann Arbor, Ann Arbor, MI, USA, kgshin@umich.edu.

2474-9567/2024/9-ART129
https://doi.org/10.1145/3678513

1 Introduction

More than 100 years ago, when Mark Twain published the short essay "*Taming the Bicycle*" regarding his funny experience of learning to ride a high-wheel bicycle in New England [49], no one would expect, a century later, another micromobility system, represented by electric(e)-scooters, proliferates within and across many cities in North America, Europe, and around the globe. Thanks to the ease of maneuverability and the reasonably fast speed (e.g., 15mph), e-scooters, seemingly easy to be "tamed", have emerged as an efficient, convenient, and first/last-mile connectivity, serving as a usable and affordable alternative to many other urban mobility options like private cars.

Despite the hype and proliferation of shared/private micromobility, particularly e-scooters [24, 64], in the streets and neighborhoods, there remains an important but largely under-explored question: how does such a new mobility mode serve as an interesting *interface* to elicit the essential but often subtle insights on the *interactions* between the riders and other traffic participants — such as motorized vehicles, pedestrians, and other e-scooters or bikes — and infrastructures and environments including stop signs, traffic lights, and road conditions. We call such interactions *Rider-to-X Interactions* (RXIs). Fig. 1, assuming an accountable e-scooter rider, illustrates three typical RXI scenes, i.e., the interactions of rider-to-car (other traffic constituents), rider-to-environment (road infrastructures), and rider-to-pedestrian. The rider may decelerate when encountering a stop sign, or yield to the pedestrians who are crossing the street.

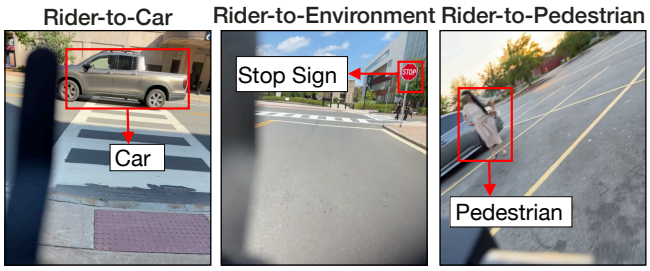

Fig. 1. Example RXIs during a naturalistic riding setting: rider-to-car, environment, and pedestrian.

Understanding these RXIs is essential to the various stakeholders, such as e-scooter manufacturers, in designing and manufacturing interactive micromobility systems, including their brakes and wheel suspension designs [71]. The insights of social-awareness can also benefit planning e-scooter sharing [23] and establishing necessary policies [19]. In addition, the recent rise of the human-centered AI (HCAI) community [43, 45, 60] can develop and deploy AI model designs and applications based on the RXI observations and needs of human riders. In particular, the tangible or foreseeable benefits of understanding RXIs include (1) dissecting the e-scooter rider interactions during, say, the commutes and recreational rides, through the behavioral analysis [64]; (2) incorporating the human interpretation (say, rider feedback) toward semantic and interpretable RXI scene understandings for developing advanced rider assistance [79], including safety and experience enhancement, and future autonomous personal micromobility systems [61]; and (3) enabling various micromobility rider-centered applications [64], such as rider behavior data analytics, insurance decisions, and accident cause analysis, as well as rider-to-car interaction support (e.g., micromobility rider or driver intention prediction) [64].

For RXI understandings with usable and broad implications, we aim to develop a novel micromobility Riding Understanding System (RUS) to interpret and understand the micromobility (e-scooters in particular) rider's maneuver behavior and the decision-making process via an HCAI design. Such an RUS will be based on the *naturalistic riding* (NR) settings, i.e., observing and understanding the riders' natural behaviors from

the unconstrained riding scenarios such as commutes or recreational rides. We want to gain behavior and decision insights from a rider-centered RUS pilot study, particularly on the typical RXIs when the e-scooter riders encounter and engage with the traffic constituents like passing cars, environments like road conditions, and incoming pedestrians.

Toward such an RUS, we have examined the existing studies and practices [40, 64, 65] to find a strong need for carefully addressing three HCAI challenges as follows.

A. How to incorporate human intelligence to augment post-scene situation-awareness and subsequent understanding of RXI scenes: One may consider capturing the rider behaviors and the RXI scenes through the behavioral sensors, such as inertial measurement units (IMUs) — accelerometer and gyroscope — that capture the sequential patterns of human behaviors [64], and visual sensors (say, camera) including video analysis [65]. However, these conventional sensing techniques [36, 59] or modalities, while largely providing on-scene situation-awareness (OSSA), may not necessarily convey the causes and outcomes behind the scenes. They may even lack the human-centered reasoning for the micromobility riders' subtle interactions with the objects in the RXI scenes, such as the implicit interactions of sign language and eye contacts between the riders and pedestrians. Textual annotations provided by the human annotators (e.g., the self-reflection from riders and notes by the observers), on the other hand, can often help incorporate the human intelligence, say, of these annotators, in understanding and reasoning about the complex RXI scenes beyond the above-mentioned OSSA. In other words, such a data modality provides further post-scene situation-awareness (PSSA), similar to the human feedback for Berkeley Deep Drive-X [34] and Tesla Self-Driving [17, 25], but with more subtle characterization of the interactions with various traffic participants encountered. A human textual annotation, "the e-scooter rider decelerates as the pedestrian told her to do so", for instance, can elicit insights on the subtle interaction behaviors (e.g., the intentions expressed by the pedestrian) within an RXI scene. However, the textual modality, unlike the behavioral and visual modalities, can often be complicated to model and integrate due to its heterogeneous formats, structures, and representations. How to incorporate them to construct *semantic* correspondences between OSSA and PSSA, i.e., meaningful and comprehensible connection across the RXI scenes and modalities (semantic OSSA-PSSA interactions), is essential for practical RUS development. Such a study, however, remains largely under-explored.

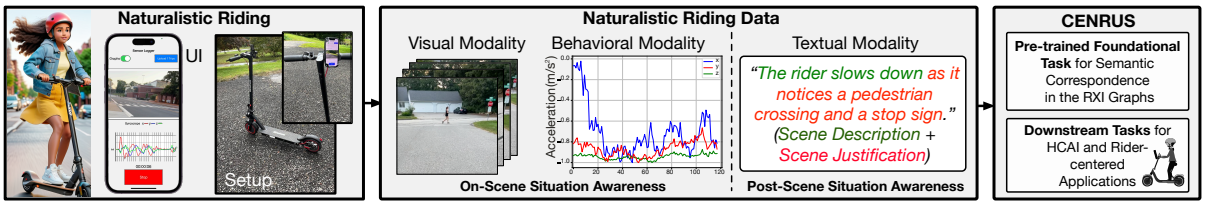

Fig. 2. Our naturalistic riding setup including user interface (UI) and e-scooters, data, and the pre-trained foundation task (PFT) and downstream tasks (DTs).

B. How to extract the semantic correspondences and characterize interactions across real-world NR data modalities: Deriving human-understandable RXI insights hinges upon extracting the semantic correspondences across the heterogeneous data modalities, i.e., visual, behavioral, and textual ones in our pilot study. An RUS needs to differentiate and contrast the feature representations from these input modalities. One may take into account the conventional representation learning [39, 64] in order to find the correlations of different modalities. However, our extensive real-world NR data analytics have revealed that these correspondences can be highly complex, dynamic, and contextually dependent. For instance, in the rider-to-environment interactions,

a rider may either continue cruising given the clear road ahead or slow down given the unsatisfactory or poor road conditions such as curbs or potholes. Modalities such as the behaviors (say, left or right turn) reflected from the IMUs may further elicit the rider's intention and decision in response to the road conditions in addition to the visual observations. We also note that these conventional learning paradigms on human behavioral activities [42, 69, 70] often seek to extract correlations of feature vectors that flatten the representations of modalities. These, however, may overlook diverse, intricate, but potentially interesting interplays that prevail among the visual cues, behavioral patterns, and textual annotations. How to further decompose, or disentangle these complex modalities to dissect correlations that are crucial for useful RXI understanding is essential for a practical RUS.

C. How to achieve generalizable and explainable NR learning to support diverse downstream tasks with HCAI implications: The goal of an RUS is not only to dissect and learn the micromobility rider behaviors in the complex RXIs, but also to enable RXI-aware downstream tasks that provide concrete, tangible, and comprehensible benefits for more rider-centered micromobility vehicles and systems. We note that different modalities, like the above ones providing OSSA and PSSA, may provide complementary information for understanding the RXI scenes when they are all available. In practice, however, the learned correspondences across these modalities during the RUS initialization may not necessarily be transferable and generalizable to the downstream tasks where some modalities may be excluded or unavailable. For instance, a downstream task of scene understanding, such as explaining what happens in the scene, or recognizing the contexts [62, 82], often takes in only the visual and behavioral data in order to derive the human-understandable interpretations. How to enable generalizable HCAI modeling to capture and transfer semantic OSSA-PSSA interactions, without assuming the availability of different modalities (e.g., sparse textual annotations), is essential and worth further exploration. Furthermore, how to explain the HCAI modeling and output results is particularly beneficial to the stakeholders, such as manufacturers, designers, and even policymakers. Conventional approaches [67, 86] usually derive correlations across the input features and the output model results, without interpreting the latent interactions of learning [41] with the physical aspects of the scenes.

To overcome the above-mentioned challenges, we propose CENRUS, a **C**ross-modality **E**-scooter **N**aturalistic **R**iding **U**nderstanding **S**ystem, to derive the HCAI insights and interactive riding behavior understandings for this emerging micromobility mode, as illustrated in Fig. 2. In particular, this paper makes the following contributions by designing, prototyping, and evaluating CENRUS:

(1) Pre-training a foundation model for graph-based semantic correspondences across RXI scenes and the relevant modalities: To enable the learnability of the complex RXIs, we have devised a pre-trained foundation task (PFT) for NR understandings, the first of its kind to our best knowledge, based on the RXI graph representation to initiate and characterize the semantic correspondences across the visual, behavioral, and textual modalities. We have devised a hierarchical RXI graph representation learning architecture for such a PFT and the resulting HCAI model. In particular, our novel RXI graph provides a unified representation that consists of modality nodes, and our PFT helps (i) construct the RXI graph edges characterizing the semantic correspondences across the heterogeneous RXI scenes and modalities of OSSA and PSSA; and (ii) provide global scene contrasting and local concept disentanglement operations upon the RXI graph to enable an effective understanding mechanism of RXI scenes from OSSA and PSSA.

(2) Augmenting semantics and disentangling concepts in the RXI understandings: To strengthen the feature representability of complex RXIs, we have designed a semantic augmentation (SA) mechanism, which introduces both the positive and negative views of different modalities for our hierarchical representation learning on the RXI graphs. In particular, for each of visual, behavioral, and textual modalities we provide the augmentation for positive views (i.e., feature groups respectively representing the modalities) that semantically

correlate with the original ones inside the RXI graphs. Such augmentation includes rephrasing textual annotations without changing its meaning. We have also introduced the negative views that contradict the original views, such as altering the derived turns extracted from the IMU time-series in the behavioral modalities. This way, CENRUS can differentiate the complex semantic OSSA-PSSA interactions across the different modalities given the counterexamples, thus enabling more effective concept disentanglement and RXI understandings.

(3) Designing rider-centered downstream tasks for model generalization, result explanation, and ubiquitous application: We have also designed several downstream tasks (DTs) as the concrete and practical applications of CENRUS, such as behavior-, video-, and text-aware tasks [16, 27, 62], and corroborated the effectiveness, accuracy, and generalizability compared with other baseline models [9, 13, 29, 47, 52, 58, 68, 70]. This is crucial for leveraging RXIs with generalizable applications in complex urban mobility environments, and potentially helping various stakeholders (such as the e-scooter manufacturers) understand, re-imagine, and re-design the various aspects of new micromobility vehicles, and comprehend the HCAI results.

We have conducted extensive studies of DTs and expanded the usability and real-world implications in understanding the complex RXI scenes for potential stakeholders (e.g., micromobility sharing providers [24]). To validate CENRUS, we have conducted extensive experimental NR studies (>11GB raw data with >2.9k RXI scenes and nearly 5.5k maneuver records), the first study to our best knowledge, on a university campus and town (urban environments) in North America. Our results have shown CENRUS, as an HCAI building block, to provide usable and practical insights on micromobility rider behaviors and system designs.

2 Related Work

2.1 Understanding Micromobility Systems

Understanding the operation of the micromobility systems in general and e-scooters in particular, as well as how they interact with the urban traffic environments is essential for re-thinking about their multi-level impacts on urban mobility network [21, 45, 71]. In terms of the macro-level interactions between the micromobility systems and their operating environments, the work in [24] studied the social interactions between the micromobility resource (e.g., e-scooter vehicles) distributions and the urban traffic environments as well as various socioeconomic attributes. The authors of [23] analyzed the dynamic micromobility vehicle redistribution and the flow prediction. On the other hand, prior studies have explored micro-level interactions of the micromobility systems with, for instance, other traffic participants (e.g., through hand-signals [4]), road conditions (e.g., roughness detected from motion sensor data [32, 43]), and pedestrians in the shared public spaces [2].

While the prior studies often focused on singling out the specific interactions (e.g., maneuvers) between the rider and the urban mobility environment, they have not carefully taken into account (a) the holistic and rider-centered study of the complex RXI scenes and the corresponding representation learning insights; and (b) the incorporation of the human intelligence (say, textual modality) for reasoning about and understanding the RXI scenes, and semantically bridging OSSA and PSSA. Our development of CENRUS can pave a new way of establishing a pre-trained foundation model [30] to gain holistic and semantic insights into interactions between humans and mobility systems, and those with the urban mobility network [65, 88]. CENRUS can serve as a foundational building block in supporting the above-mentioned multi-level micromobility usage studies [54] and management operations such as e-scooter sharing [1].

2.2 Understanding Mobility System Users

Understanding how the users interact with their mobility systems, including the micromobility vehicles [22, 64], has attracted attention from the industry and academia [16, 22, 40, 43, 44]. For instance, mobility system maneuvering and decision-making processes can be identified through video recording [33], IMU [64], and other multi-modality mechanisms [34, 65]. Interactions between the users and the micromobility systems can be

captured through the audio-vision integration as well as further context fusion (say, the traffic conditions) [65]. In addition to sensing modalities, various learning paradigms, including meta learning [67], federated learning [67], unsupervised learning [11], and the emerging generative learning approaches [28], have been explored for behavioral insights into mobility system usage. However, how to fuse the heterogeneous modalities for in-depth characterization of the riders' behaviors and decisions in taming the micromobility vehicles remains under-explored.

Despite these prior studies in understanding users' interactions with mobility systems, how to design the HCAI models for micromobility, and generalize the learning models to a variety of rider-centered DTs [30, 47], need in-depth exploration. Our proposed system, CENRUS, through the lens of micromobility (e-scooter) vehicles, can help elicit insights within and across the complex RXI scenes. We note that our studies on the users of micromobility systems differs from those on the autonomous driving scenarios (such as Tesla [17, 25] and Berkeley Deep Drive-X [34]), since the former users group is often less constrained by the driveways and hence more diverse interactions with a much broader spectrum of traffic constituents can be observed. On the other hand, the micromobility vehicles, such as e-scooters, tend to have more diverse choices (as well as human preferences in picking the routes and path conditions) in operation in the various mobility environments other than the roads (e.g., share of roads, bike lanes, kicked along the crosswalk). Furthermore, our proposed PFT also provides a new paradigm to effectively capture the semantic OSSA-PSSA interactions through the RXI graph representation, hence improving the learnability and generalizability across tasks such as behavior recognition and scene understanding.

3 Overview of CENRUS

3.1 Overview of System

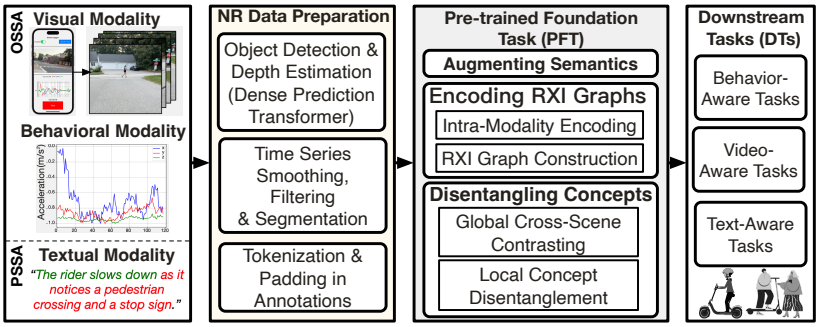

Fig. 3. System overview of CENRUS, which consists of a pre-trained foundation task (PFT) and downstream tasks (DTs).

Fig. 3 shows the overall system architecture of CENRUS that consists of two sets of tasks: a *pre-trained foundation task* (PFT) and *downstream tasks* (DTs). Given the processed NR data (including cleaning and filtering) of the heterogeneous modalities — i.e., visual and behavioral ones for OSSA, and textual ones for PSSA — the PFT first performs the semantic augmentation (SA) to generate positive and negative views beyond the original ones, which, respectively, preserve or contradict the key characteristics of the input data (§4.1). CENRUS then encodes the NR data, and transforms their resulting embeddings into a RXI graph representation, where each node represents the original, positive, and negative views and their edges represent the semantic OSSA-PSSA interactions (§4.2). CENRUS further conducts the hierarchical RXI graph representation learning in order to jointly perform the global scene contrasting and local concept disentanglement operations. The resulting RXI graph

representations and the model from our PFT are then exported to support various DTs for micromobility systems, including behavior-, video-, and text-aware tasks (§5).

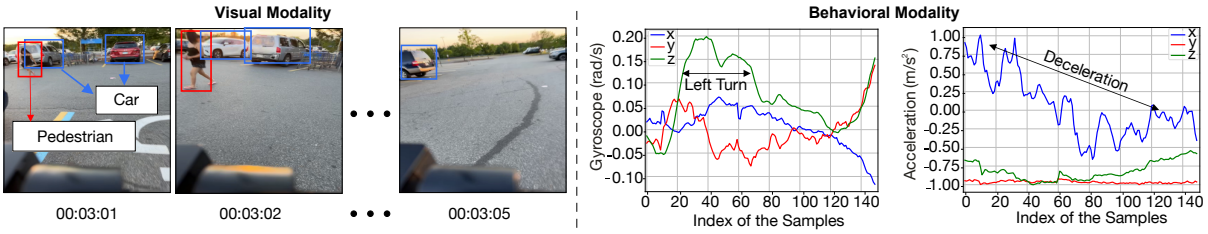

Fig. 4. Visualization of an RXI scene with the textual modality given as "the rider slows down while turning as it notices a pedestrian crossing".

3.2 Collection of Naturalistic Riding (NR) Data

In this prototype study, we have recruited a total of 18 participants (13 males and 5 females of ages 20–35 and heights 4'11"–5'11") for the NR data collection in our university town and campus environments (urban environments as showcased in Fig. 1) with diverse RXI scenes. Despite our current focuses on system and model prototyping, we will expand the participant studies in the future. Recall that Fig. 2 illustrates our experimental setup in which the participants used either GoTrax GXL V2 or iSinwheel i9 Pro e-scooter for collecting the NR data. During the NR data collection, we mount the smartphone (iOS: iPhone 13 Pro, 13 Mini, SE 2022, and 7 Plus; Android: Google Pixel 3 and Xiaomi Redmi Note 8) upon the handlebar of the e-scooter, and collect the NR and RXI scene data in visual and behavioral modalities, i.e., the back-view (main) camera video recordings regarding the visual scenes and road conditions, as well as the IMU measurements (accelerometer and gyroscope in our case) based on our NR data collection app. Note that all the participants, as accountable riders, followed the local regulations and riding ethics, such as wearing a scooter helmet, avoiding rush hours and sidewalks with pedestrian crowds, and observing stop signs, throughout the NR data collection (reviewed and approved by our university Internal Review Board (IRB); with sensitive information like human faces or license plates blurred).

We have collected a total of 11.7GB e-scooter NR trips, resulting in a total of 2,965 RXI scenes (each lasts about 5s). Fig. 4 illustrates an example RXI scene. We have identified 515 key RXI scenes that contain riders'

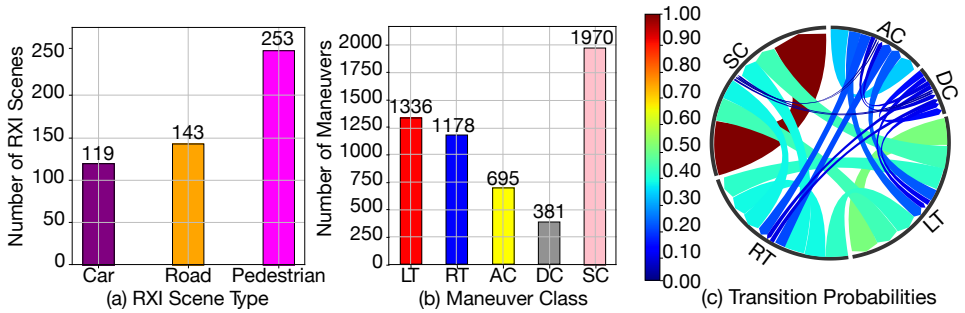

Fig. 5. Statistics of (a) key RXI scenes involving interactions with cars, road environment, or the pedestrians and (b) maneuvers in our NR data collection; and (c) transition probabilities across maneuvers where warmer colors or wider arrows indicate a higher transition probability from one maneuver to another.

interactions with exterior stimuli, i.e., cars, environments, and pedestrians (showcased in Fig. 5(a)), plus the

rest driven by the riders' internal decisions when interacting with her/his scooter (e.g., straight cruising given the clear way). As for cars, we have considered and labeled the RXI scenes (#) with passing (61), parked (20), or incoming cars (38). As for environments (road and infrastructures), we have accounted for stop signs (77), speed humps (51), and potholes (15). Regarding those with pedestrians, we have included the ones with incoming (182), passing (51), and grouped pedestrians (15), and the avoidance of individuals/crowds (5).

In terms of maneuvers that characterize the riders' direct interactions with the scooter and the physical world, we have considered left turn (LT), right turn (RT), acceleration (AC), deceleration (DC), and straight cruising (SC). We have identified a total of 5,560 rider maneuver behaviors (such as left/right turns) within all the identified RXI scenes (detailed statistics can be referred to Fig. 5(b)).

Note that the scene/maneuver distributions result from our participants' naturalistic riding (i.e., not purposely controlled or balanced) process may elicit the natural picture of daily encounters (particularly in a college town or urban environment with more pedestrians; expansion to other city environments will be considered in our future work). The chord chart in Fig. 5(c) further illustrates the relative frequencies of different maneuvers (sizes of the sectors) as well as transition probabilities across them (coded in colors), and shows that the deceleration is often accompanied by cruising and turning actions for the mobility and environment transitions of the riders.

3.3 Preparation of Modalities

Modalities for CENRUS are prepared as follows.

3.3.1 OSSA — Visual Modality $\mathbb{V}$: In preparing the visual modality, we focus on two key aspects — *object detection* and *depth estimation* — regarding the OSSA of the RXI scenes. Fig. 6 shows the pipeline of our visual modality processing. Our key idea is to determine the objects of interest (OoIs), i.e., other traffic participants (such as cars

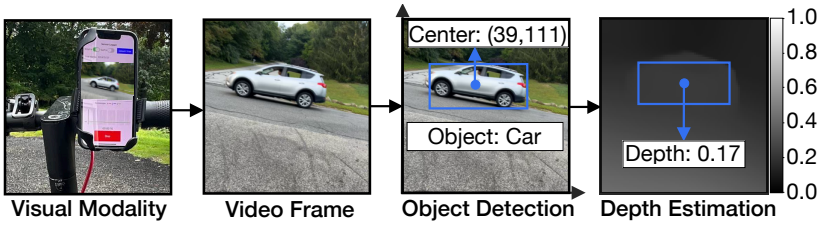

Fig. 6. Illustration of object detection and depth estimation pipeline in CENRUS.

and buses), traffic environments (such as road infrastructures), and pedestrians, involved in the RXI scenes for subsequent scene learning, reasoning, and understanding based on their relative closeness to the rider (ego spot). In particular, similar to the practices of existing auto-piloting systems [46] for computation efficiency [26, 34], we down-sample the video frame rate of 60Hz and size of 480×640 (width $\times$ height; our smartphone is mounted vertically to capture the front horizon) into 1Hz and 90×160, respectively, for ease of processing and later encoding. Our pilot studies leverage the YOLOv5 object detector [55, 72] to identify the type of OoIs and extract the position of the OoIs relative to each frame (the center coordinate of the bounding box returned from YOLOv5; in pixels from the bottom left of the video frame).

For each OoI detected, we further take into account the Dense Prediction Transformer (DPT) [53] for extracting its depth within the frame. In particular, given the center of the objects, we consider the depth of the object's center point in the depth map estimated by DPT as the depth of each OoI. In addition, we normalize the DPT depth output into the range of [0,1], where a value closer to 1 indicates a shorter distance to the ego entity (rider). We select the K closest OoIs (empirically $K = 5$ in our current studies) from each frame. As the relative distances

between the riders and different OoIs are dynamically changing over time, the K nearest OoIs as well as their relative order may vary across frames.

We focus on the following types of OoIs: pedestrians, motorized vehicles (cars, motorcycles, buses, and trucks), bicycles, stop signs, parking meters, and traffic lights. We also account for a special class for the frames with no significant OoI involved. These all sum up to a total of 10 classes of OoIs. We illustrate an example of the object detection and depth estimation for a video frame in Fig. 6. The resulting visual modality data $\mathbb{V}$, which is a series of video frames, consists of the detected object classes (one-hot encoding out of the 10 classes [18]), position coordinates (say, a coordinate of $(39, 111)$ that is relative to the bottom left and normalized by the respective frame width and height) related to each frame, as well as their respective depth (say, 0.17; the gray-scale depth map contains the silhouette of the car).

3.3.2 OSSA — Behavioral Modality $\mathbb{B}$: To characterize the behavioral modality, we focus on the IMU measurements from the accelerometer and gyroscope of the on-board smartphones, which amount to a total of 6 IMU time series (x, y, and z axes for each sensor). In preparing the behavioral modality data as another OSSA, we transform all the axes of time-series from the smartphone coordinate system to the earth coordinate system [64] through the rotation matrices provided by the Apple iOS or Android APIs [6]. We divide each time-series (with 40Hz re-sampling frequency from the raw time-series; smoothed through the moving average [64]) into segments using a sliding window (empirically set as 5s to accommodate the RXI scenes) with no overlap, and we normalize the time series to the range of [-1, 1].

3.3.3 PSSA — Textual Modality $\mathbb{T}$: To incorporate human intelligence within interpretation of RXIs (human reasoning and understanding) for PSSA, we have performed the textual annotations upon the scenes, emulating the scenarios for the rider feedback and the observer notes. For instance, an annotation can be "the rider slows down as noticing a pedestrian", which consists of the *scene description* (i.e., "the rider slows down") and the *scene reasoning* (i.e., "as noticing a pedestrian"). In practice, the object detector (like YOLOv5) might not be able to well recognize a pedestrian (say, if only her/his legs were filmed), while the textual annotation further helps provide such subtle information regarding the RXI scene interpretation. Given the RXI scene annotations, we first convert each sentence to a list of small units called tokens [78, 81]. Such a tokenization process eases our textual modality encoding [65, 88]. Furthermore, we pad the tokens of each textual annotation based on a predefined token PAD [10, 65] as a placeholder. This way, the total token numbers in all textual annotations are equal, and each of the resulting tokens consists of a total of L dimensions ($L = 40$ in our current study in order to accommodate the maximum numbers of tokens in our NR data).

4 Pre-trained Foundation Task (PFT)

Fig. 7 illustrates the flow of our PFT, which consists of three major phases: (i) semantic augmentation (SA); (ii) RXI graph construction (with intra-modality encoding); and (iii) hierarchical RXI graph representation learning for disentangling the NR understandings.

4.1 Augmenting Semantics

Given the complex, originally unstructured, and highly dynamic NR data, the key idea of our semantic augmentation (SA) is to create positive and negative views of the original visual, behavioral, and textual modalities to strengthen the PFT learnability [74, 80, 84]. In particular, the positive SA maintains the key characteristics, such as the contextual, temporal, and sequential dependencies across the riders' decision-making process relevant to the RXI scenes inside the original data. On the other hand, the negative SA provides the views contradicting the original RXI scenes. This way, our PFT can further experience various views to strengthen its ability of interpreting the RXI scenes.

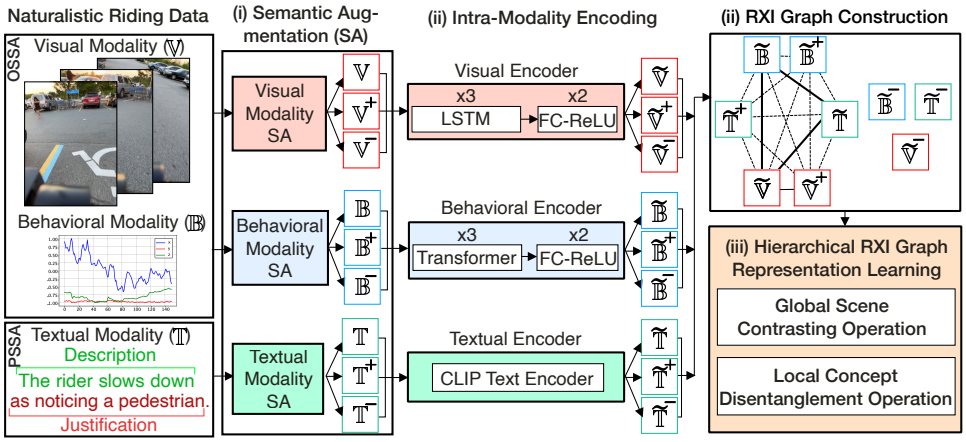

Fig. 7. Information flow of our pre-trained foundation task (PFT) in CENRUS.

4.1.1 Visual Modality SA: For the objects of interest (OoIs), such as cars and traffic lights, detected from the frames, we alter the depth of different objects to generate positive and negative views. Our pilot study focuses on

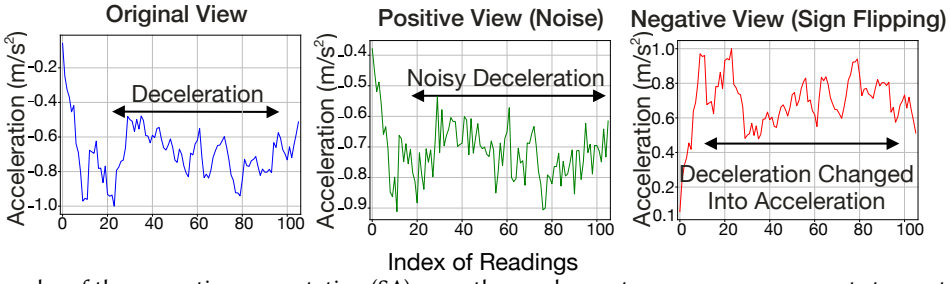

Fig. 8. Examples of the semantic augmentation (SA) upon the accelerometer sensor measurements to create positive and negative views of the behavioral modality by adding Gaussian noise or flipping the sign of the measurements, respectively.

the depth since the closeness of OoIs reflects the longitudinal awareness (say, risk assessment) of the ego rider regarding the RXI scenes [50]. For the positive SA, the resulting views, denoted as $\mathbb{V}^{\mathrm{DP}(+)}$, generally preserve the relative closeness with the added uniform noise (say, $[-0.1, 0.1]$) upon the depth of the OoIs, while for the negative SA, CENRUS contradicts the semantics of the RXI scenes by reversing the relative closeness of important OoIs to the rider. For instance, if a pedestrian is detected close to the rider (e.g., with a normalized depth value above 0.5), our negative augmentation alters the depth inside the frame, and generates $\mathbb{V}^{\mathrm{DP}(-)}$ where the pedestrian is considered far away from the ego spot (rider) with the depth value reversed.

4.1.2 Behavioral Modality SA: CENRUS imposes the Gaussian noise (with zero mean and standard deviation of each original IMU time-series of accelerometer and gyroscope) upon each of the 6-axis IMU time-series independently to obtain the positive view, denoted as $\mathbb{B}^{\mathrm{GN}(+)}$, that preserves the overall behavioral characteristics. On the other hand, for the negative view, denoted as $\mathbb{B}^{\mathrm{F}(-)}$, we provide the counterpart of the IMU time-series by flipping the signs of the measurements from each of the six axes. This is motivated by our experimental observations on the symmetry that prevails across the rider's behavior counterparts (e.g., between LTs and RTs; ACs and DCs). Our behavioral modality SA helps introduce the contrasting semantics for PFT, such as contradicting the scenes when the ego rider steers the scooter to the left to avoid an incoming car from the right.

This also complements, via a horizontal (pitch axis) scope [50], the change of depth (longitudinal axis) in our visual modality SA.

4.1.3 Textual Modality SA: As each textual annotation consists of both scene description and scene justification, we perform SA to jointly account for both, and incorporate the human intelligence when reasoning in original and reverse views. As for positive SA, we replace the scene description or the scene justification parts of the textual annotations with semantically or contextually similar texts, and obtain $\mathbb{T}^{\mathrm{D}(+)}$ and $\mathbb{T}^{\mathrm{J}(+)}$, respectively. We specifically design a rule-based approach which searches within the training data (the original ones only) and leverages the other similar annotations (with most of the shared tokenized words) to replace either the scene description or the scene justification. For instance, given "the rider is slowing down as she notices a pedestrian", we can generate two possible positive views — that is, "the rider is reducing its speed, as s/he notices a pedestrian", and "the rider is slowing down as a pedestrian is crossing the street". For the negative SA, we leverage the antonyms of the rider maneuvers in order to provide counterparts of the scene description and justification denoted by $\mathbb{T}^{\mathrm{D}(-)}$ and $\mathbb{T}^{\mathrm{J}(-)}$, respectively. For instance, the annotation of "the rider turns left to avoid a group of pedestrians blocking the way" can be converted to "the rider turns right to avoid a group of pedestrians blocking the way".

4.2 Encoding RXI Graphs

4.2.1 Intra-Modality Encoding: We have designed three sets of encoding mechanisms to prepare for the RXI graph representation learning. We note that despite the current pilot studies CENRUS is also general enough to be extended to alternative encoding mechanisms, including other pre-trained language-image and time-series models [30]. The three sets of mechanisms are presented as follows.

(a) Regarding the visual modality, CENRUS feeds the augmented $\mathbb{V}$ — that is, OoIs, coordinates, and depths — through a multi-layered long short-term memory (LSTM) [18] mechanism (L_1 layers with B_1 hidden units) followed by dense or fully-connected (FC) [18] layers (L_2 layers with B_2 hidden units) and generate the subsequent visual embeddings. The resulting visual embeddings are denoted as $\widetilde{\mathbb{V}}$ for the original view and $\widetilde{\mathbb{V}}^{\mathrm{DP}(+)}$ and $\widetilde{\mathbb{V}}^{\mathrm{DP}(-)}$ for the augmented views.

(b) CENRUS processes the augmented $\mathbb{B}$ through L_3 consecutive transformer encoders [75], each with B_3 attention heads for attention score calculation. CENRUS also finds the average of the interim behavioral embeddings from the transformers along the time dimension (say, corresponding to the temporal time-series samples within a 5s window) and generates the behavioral embeddings via L_4 FC layers (with B_2 hidden units). The resulting behavioral embeddings are denoted as $\widetilde{\mathbb{B}}$ for the original view and $\widetilde{\mathbb{B}}^{\mathrm{GN}(+)}$ and $\widetilde{\mathbb{B}}^{\mathrm{F}(-)}$ for the augmented views.

(c) For the augmented $\mathbb{T}$, CENRUS takes in the tokens of the human textual annotation and processes them through a contrastive language-image pre-training (CLIP) encoder [52] as an initialized association phase between the visual and textual modalities to generate the textual embeddings. are denoted as $\widetilde{\mathbb{T}} \in \mathbb{R}^{B_2}$ for the original view as well as $\mathbb{T}^{\mathrm{D}(+)}/\mathbb{T}^{\mathrm{J}(+)}$ (descriptions) and $\mathbb{T}^{\mathrm{D}(-)}/\mathbb{T}^{\mathrm{J}(-)}$ (justifications) for the augmented views. We note that the pre-trained CLIP encoder [47, 52] only resides upon the conventional visual and textual tasks, while the PFT in CENRUS aims to provide understandings that are more generalizable to the RXI scenes in NR data beyond these pre-trained models.

4.2.2 RXI Graph Initialization: Based on the resulting visual, behavioral, and textual embeddings, we further initialize the RXI graph, as illustrated in Fig. 7, for each of the above-mentioned different RXI scenes. These graphs will be fine-grained further by our hierarchical RXI graph representation learning. For instance, given an RXI scene with the textual annotation "the rider is turning left to avoid a pedestrian", the resulting RXI graph consists of the nodes of modalities relevant to this particular scene, which represents the visual embeddings of the frames showing the riders' interaction with the encountered pedestrians. In addition, the behavioral embeddings

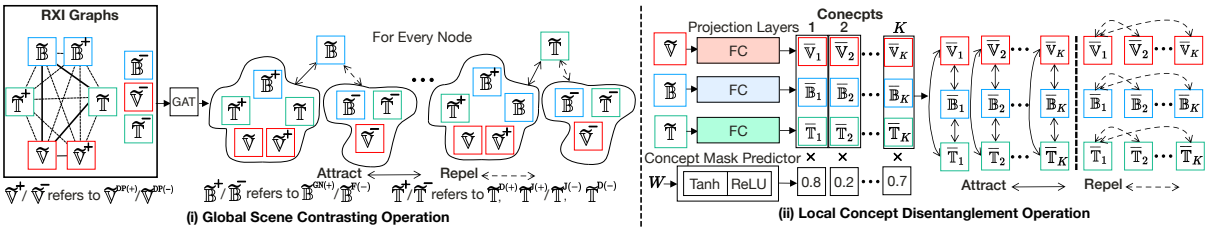

Fig. 9. Illustration of the disentangling understandings in our hierarchical RXI graph representation learning.

incorporate the behavioral patterns as well as sequential dependencies within the RXI scene. Such a graph contains the nodes, represented by the respective modality embeddings, that are semantically corresponding to each other through their weighted edges. On the other hand, CENRUS enforces no edges across the graphs that represent different RXI scenes (say, one related to rider-to-pedestrian interactions, and the other related to rider-to-environment interactions). The thus-constructed RXI graphs are then further fed through the operations of the global scene contrasting and local concept disentanglement to update and retrieve the semantic OSSA–PSSA interactions.

4.3 Disentangling Understandings

4.3.1 Global Scene Contrasting Operation: As illustrated in Fig. 9(i), we first perform the global scene contrasting operation to enable knowledge transfer across the complex RXI scenes and derive more rider-centered and generalizable RXI understandings for DTs. This operation involves actions of *attracting* and *repelling*, i.e., associating the RXI graphs that represent the same RXI scenes; and differentiating those RXI graphs that represent the different RXI scenes and refining the RXI graphs.

In particular, CENRUS first processes the given RXI graph through a graph attention (denoted as GAT [76]) layer to update the node embeddings output from the intra-modality encoders (§4.2). Then, for each node in the given RXI graph, let $\{\mathbf{S}_1, \ldots, \mathbf{S}_Z\}$ be the Z RXI graphs, where $\mathbf{S}_i = \{\mathbf{e}_{(i,1)}, \ldots, \mathbf{e}_{(i,C_i)}\}$ represents the C_i node embeddings in an RXI graph $\mathbf{S}_i$. For the learning operation, we define the Global Scene Contrasting Loss (GSCL) for node $\mathbf{e}_{(i,j)}$, denoted as $\ell_{(i,j)}$, i.e.,

$$\ell_{(i,j)} = -\log\left(\frac{A}{A+R}\right), \tag{1}$$

where A and R are, respectively, the sums of exponential similarities across (a) the RXI graphs that belong to the same RXI scenes and are to be attracted toward each other, and (b) the RXI graphs that correspond to those different RXI scenes and are to be repelled from each other.

Formally, the GSCL with respect to the RXI graphs i and j is given by

$$A \triangleq \sum_{m=1}^{C_i} \exp\left(\frac{\text{sim}\left(\mathbf{e}_{(i,j)}, \mathbf{e}_{(i,m)}\right)}{\tau_{\text{global}}}\right), \quad \text{and} \quad R \triangleq \sum_{p\neq i}^{Z} \sum_{m=1}^{C_p} \exp\left(\frac{\text{sim}\left(\mathbf{e}_{(i,j)}, \mathbf{e}_{(p,m)}\right)}{\tau_{\text{global}}}\right), \tag{2}$$

where $\text{sim}(\cdot,\cdot)$ is the cosine similarity function given two input embeddings and τ represents the temperature parameter [8, 56]. We note that smaller τ encourages the optimizer of CENRUS to focus on the negative pairs of RXI graph nodes that are hard to distinguish. The total GSCL for all RXI scenes is then given by the sum of $\ell_{(i,j)}$

for resulting nodes within all the K RXI graphs, i.e.,

$$\mathcal{L}_{\text{global}} = \sum_{i=1}^{K} \sum_{j=1}^{C_i} \ell_{(i,j)}. \tag{3}$$

Note that the node embeddings result from the GAT and refinements of the node embeddings update the adjacency matrices of the RXI graphs. This way, CENRUS learns and captures the latent semantic OSSA–PSSA interactions across the different RXI scenes.

4.3.2 Local Concept Disentanglement Operation: Even with the scene contrasting, the modalities within each graph of an RXI scene may still contain highly coupled correspondences across each other, making it hard for CENRUS to discriminate the different latent factors within the heterogeneous input modalities. Instead of providing a single representation that flattens all the modalities involved, we have designed the local, independent, and latent concepts that capture the underlying factors responsible for the input modalities [31, 77]. For instance, the disentangled concepts can represent the relative importance of the underlying factors within the RXI scenes. This way, CENRUS provides comprehensive characterization of the local structures, and we will derive HCAI conceptualization and self-explanation in characterizing the diverse RXI scenes.

As shown in Fig. 9(ii), CENRUS projects the embeddings of each of the three modalities into K concepts, each of which has dimension $\frac{B_2}{K}$. This is through a total of $3 \times K$ independent FC layers (each with $\frac{B_2}{K}$ hidden units) and obtains $\left\{\overline{\mathbb{V}}_1, \ldots, \overline{\mathbb{V}}_K\right\}$, $\left\{\overline{\mathbb{B}}_1, \ldots, \overline{\mathbb{B}}_K\right\}$, and $\left\{\overline{\mathbb{T}}_1, \ldots, \overline{\mathbb{T}}_K\right\}$. The subsequent concept disentanglement operation differentiates the latent interactions across different modalities, and identifies their relative importance. We take the i-th concept of the visual modality, denoted as $\overline{\mathbb{V}}_i$, as an example, and leave out other modalities in the interest of limited space available. CENRUS randomly initializes a set of concept mask weights $\mathbf{W} = [W_1, \ldots, W_K] \in \mathbb{R}^K$, where each element is regularized within the range of $[0, 1]$ via a Tanh activation followed by ReLU. A high W_i indicates more importance of the concept i. This way, CENRUS can dynamically determine the number of concepts extracted from the NR data. The concept disentanglement operation for each modality aims to minimize

$$\ell_{\overline{\mathbb{V}}_i} = -W_i \cdot \log\left(\frac{A_{\overline{\mathbb{V}}_i}}{A_{\overline{\mathbb{V}}_i} + D_{\overline{\mathbb{V}}_i}}\right), \tag{4}$$

where $A_{\overline{\mathbb{V}}_i}$ is the similarity of the visual embeddings $\overline{\mathbb{V}}_i$ with respect to (w.r.t.) the i-th behavioral embedding $\overline{\mathbb{B}}_i$ and textual embedding $\overline{\mathbb{T}}_i$ in the same concept i, i.e.,

$$A_{\overline{\mathbb{V}}_i} \triangleq \exp\left(\frac{\text{sim}\left(\overline{\mathbb{V}}_i, \overline{\mathbb{B}}_i\right)}{\tau_{\text{local}}}\right) + \exp\left(\frac{\text{sim}\left(\overline{\mathbb{V}}_i, \overline{\mathbb{T}}_i\right)}{\tau_{\text{local}}}\right). \tag{5}$$

$D_{\overline{\mathbb{V}}_i}$ represents the similarity of $\overline{\mathbb{V}}_i$ w.r.t. the other visual concepts, i.e.,

$$D_{\overline{\mathbb{V}}_i} \triangleq \sum_{m \neq i}^{K} \exp\left(\frac{\text{sim}\left(\overline{\mathbb{V}}_i, \overline{\mathbb{V}}_m\right)}{\tau_{\text{local}}}\right), \tag{6}$$

where τ_{local} represents its temperature parameter. In summary, this operation jointly (a) determines the number of the concepts that can be extracted from the NR data (some W_i's are close to 0); (b) aligns the visual, textual, and behavioral embeddings relevant to the same concept; and (c) differentiates the resulting concepts via W_i's.

Similar to the global scene contrasting operation, this operation will further update the node embeddings and result in refined RXI graphs (characterized by their adjacency matrices) in terms of their semantic OSSA-PSSA interactions through GAT. We can similarly calculate $\ell_{\overline{\mathbb{B}}_i}$ and $\ell_{\overline{\mathbb{T}}_i}$ losses, which makes the total local concept

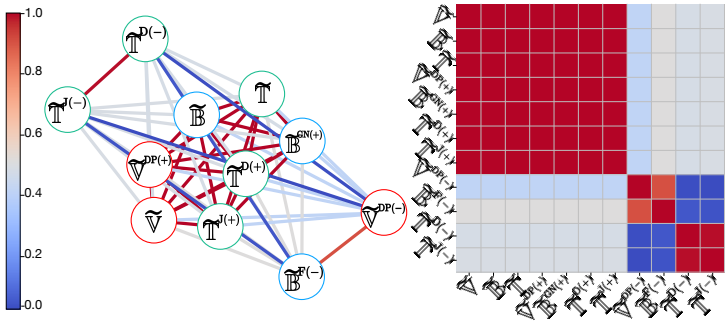

Fig. 10. Illustration of a learned RXI graph (i.e., weighted edges of the graph and its corresponding adjacency matrix) for an instance of rider-to-pedestrian interaction. Warmer colors indicate the higher edge weights and stronger semantic correspondences.

disentanglement loss (LCDL) as

$$\mathcal{L}_{\text{local}} = \sum_{i=1}^{K} \left(\ell_{\overline{\mathbb{V}}_i} + \ell_{\overline{\mathbb{B}}_i} + \ell_{\overline{\mathbb{T}}_i} \right). \tag{7}$$

The final training objective of CENRUS is, therefore, to jointly minimize the weighted sum of GSCL and LCDL, i.e.,

$$\mathcal{L} = \lambda_{\text{global}} \cdot \mathcal{L}_{\text{global}} + \lambda_{\text{local}} \cdot \mathcal{L}_{\text{local}}, \tag{8}$$

where λ_{global} and λ_{local} are the hyperparameters determining their relative importance.

Fig. 10 illustrates an example of the learned RXI graph and the corresponding adjacency matrix for an RXI scene. We can observe that CENRUS has learned strong semantic OSSA-PSSA interactions (warmer colors) among the positive and the original nodes, while the cold colors indicate weak ones between a positive/original node and a negative one (say, the textual modality belonging to a different RXI scene).

5 Downstream Tasks (DTs)

The usability of CENRUS resides beyond the PFT that learns the semantic OSSA-PSSA interactions. Using the globally-contrasted scene dependencies and locally-disentangled concepts, we have further designed multiple usable DTs to enable concrete rider-centered AI applications. The usability of our PFT in CENRUS is corroborated through the following DTs: (i) behavior-aware task (BAT), such as the maneuver recognition ($\mathbb{B} \rightarrow \mathbb{B}$) [5]; (ii) vision-aware task (VAT), such as video information retrieval ($\mathbb{T}$ or $\mathbb{B} \rightarrow \mathbb{V}$) [7]; and (iii) text-aware task (TAT), such as annotation classification ($\mathbb{T} + \mathbb{V} + \mathbb{B} \rightarrow \mathbb{T}$) and context recognition ($\mathbb{V} + \mathbb{B} \rightarrow \mathbb{T}$).

5.1 Behavior-Aware Tasks (BATs)

Given the behavioral modality data (IMU time-series), BAT or $\mathbb{B} \rightarrow \mathbb{B}$ aims to identify five maneuver classes: left turn (LT), right turn (RT), acceleration (AC), deceleration (DC) or braking, and straight cruising (SC). Note that many of today's e-scooters have cruise control (CC) functionality, and when the rider presses the CC button, the e-scooter maintains a certain speed (say, 15mph) in the straight cruising. Despite the seemingly simplicity of BAT, our DT is to classify the multiple labels instead of one per RXI scene, since multiple maneuvers may be captured within each scene. When the pre-trained foundation task is done, we leverage the behavioral encoder (see §4.2) and fine-tune with the IMU time-series as the input and maneuver classes as the output. To evaluate BAT, we compare the maneuver classification performance with (w/) and without (w/o) our PFT, emulating the scenarios for real-world applications such as rider activity recognition.

5.2 Vision-Aware Tasks (VATs)

Given the textual or behavioral data, this downstream task ($\mathbb{T}$ or $\mathbb{B} \rightarrow \mathbb{V}$) aims to retrieve the closest visual frames from the NR data. Specifically, given the textual or behavioral embeddings encoded by CENRUS as the query, as well as all the N visual modality data (embeddings) as the database, CENRUS finds the cosine similarity for every pair of textual/behavioral and visual embeddings, and then returns the visual data (frames) with the highest similarity [20]. VATs can emulate the application scenarios regarding rider NR data management of RUS, such as retrieval of the RXI scenes for managing e-scooter sharing [40], accident analysis, and insurance rate determination [63].

5.3 Text-Aware Tasks (TATs)

Given the PFT model, one TAT, annotation classification or $\mathbb{T} + \mathbb{V} + \mathbb{B} \rightarrow \mathbb{T}$ [65], is to fine-tune the visual, behavioral, and textual encoders (§3.3) to identify the following RXI scenes from the textual annotations (e.g., recognizing the category of a certain portion in the sentence): (1) rider maneuvering status, such as left/right turns; (2) traffic environment status, including annotations related to clear way or heavy traffic; (3) interacting with cars, such as steering to the right side of the lane to yield the road to a following car; (4) interacting with pedestrians, like slowing down when a pedestrian is crossing the street; and (5) interacting with the traffic environments including stop signs and speed bumps. Given each token, this TAT outputs the labels regarding whether the tokens inside the textual annotations relate to any of the above five categories or not. For instance, given the textual annotation "the rider slows down as he notices a pedestrian crossing the street", the textual annotation identifies the concept belonging to the concept of (1) rider maneuvering status from the tokens of "The rider slows down", and another one related to (4) interacting with pedestrians from the tokens of "he notices a pedestrian crossing the street". This task then assigns the labels (1) and (4) to the corresponding texts within the textual annotation.

Another TAT, context recognition or $\mathbb{V} + \mathbb{B} \rightarrow \mathbb{T}$ [48], is to take in only the visual and behavioral modalities and classify the contexts of interacting with (1) pedestrian(s); (2) other vehicles (e.g., cars, scooters); (3) road environment (e.g., stop signs); and (4) no important or meaningful interactions. The TAT of context recognition ($\mathbb{V} + \mathbb{B} \rightarrow \mathbb{T}$) needs to capture the correspondences between the RXI scenes and the implicit semantics, thus creating more challenges than the TAT of annotation classification ($\mathbb{T} + \mathbb{V} + \mathbb{B} \rightarrow \mathbb{T}$).

These two TATs represent our core initiatives of understanding the RXI scenes, and can enable other important socially-aware applications, such as the rider assistance systems [79] and explainable autonomous micromobility [66].

6 Experimentation

6.1 Evaluation Settings

6.1.1 Baseline Approaches: To corroborate the PFT model within CENRUS for the concrete DTs such as BAT, VAT, and TAT, we have implemented and compared CENRUS with the following state-of-the-art and baseline approaches. In particular, our comparative evaluation includes:

- IMU2CLIP [47]: contrastive learning on visual, behavioral, and textual data;
- CLIP [52]: which focuses on extracting similar representations for visual and textual data;
- SSIMU [70]: focusing on extracting similar representations for visual and behavioral data;
- ALIGN [29]: using EfficientNet [68] and BERT [13] as the visual and textual encoders;
- BLIP [38]: which utilizes noisy data for more generalizable representation learning on visual and behavioral data;
- AltCLIP [9]: adapted from CLIP and augmented with bidirectional attention mechanism;
- FLAVA [58]: with the visual and textual fusion architecture;

- LIMU [83]: which applies the BERT [13] mechanism on behavioral data;
- TS2Vec [85]: with contrastive learning on behaviors; and
- TFC [87]: focusing on time-frequency consistency on behaviors.

6.1.2 Default Parameter Settings: The IMU sampling frequencies of our iOS and Android devices are empirically set to 40Hz and 240Hz, respectively. To get the moving average for smoothing IMU time-series, we adopt a sliding window of 1.25s. We use 80% of the NR data for model training and validation, and the rest for evaluation. Furthermore, we perform PFT for 300 epochs with a learning rate of 1e-4 and $\lambda_{\text{global}} = \lambda_{\text{local}} = 1$. Besides, we evaluate and empirically set the temperature values τ_{global} and τ_{local} as 1e-2. For intra-modality encoding, we use $L_1 = 1$ LSTM layer with $B_1 = 32$ to process the visual modality data. For $\mathbb{B}$, we use $L_3 = 1$ transformer layers [75] with $B_3 = 4$ attention heads. For all the modalities, we use $B_2 = 1$ FC layer with $L_2 = L_4 = 512$ hidden units and a LeakyReLU activation function to generate their embeddings [18]. When applying the PFT model for the DTs (BAT, VAT, and TAT), we freeze the PFT model and introduce the output embeddings as the input for the neural networks (see §5) operating the DTs.

6.1.3 DT Settings and Evaluation Metrics: As for the BAT, we adopt the same PFT architecture to process $\mathbb{B}$. We provide the baseline approaches with the modality inputs indicated in their original designs to pre-train the models. We fine-tune each pre-trained model with a total of 1e+3 epochs with a learning rate lr=1e-3 and a batch size of 32. As for VAT, we follow the settings as in §5 to search for and return the nearest embeddings. As for annotation classification (TAT), we introduce an additional classification head that has a total of three FC layers to merge the output embeddings of all the encoding models to predict the tokens' categories within the textual annotations. The first two layers of the classification head comprise 64 hidden units and the ReLU activation function, while the last one has a total of 40 units (versus the number of the tokens) and the Sigmoid function. We train the resulting TAT model for 1e+3 epochs with lr=1e-2. For context recognition TAT, we use a similar classification head and train with lr=1e-5 with early stopping. In addition, CENRUS converts the probability outputs of the textual tokens into 1s if the probability from the Sigmoid function output is greater than a certain value (empirically evaluated and set as 0.6). We use the Adam optimizer [18] and 20% Dropout rates for all the foundational models and the DTs (based on PyTorch 2.0.1).

All baselines are evaluated on a deep learning server with 4×NVIDIA RTX3090 24GB GDDR5, 1×AMD Ryzen Threadripper 3960X 24-Core 48-Thread CPU, and 128GB RAM. Model pre-training, forward passing for embedding generation, RXI graph initialization, and SA take 70.17ms, 41.66ms, 0.01ms, and 0.02ms, respectively, on average per RXI scene. We note that once CENRUS is pre-trained and fine-tuned, various DTs can be efficiently executed based on the pre-trained model. Further efficiency enhancement (such as network pruning for resource-contrained embedded devices) will be considered in our future study [57].

We measure the benefits of CENRUS's PFT through the performance of various DTs. For our BAT and TATs, F1 score [20] is used to evaluate the performance of CENRUS and other baseline approaches in terms of classifying the maneuvering behaviors and human-mobility interaction status, respectively. For VATs, we find the percentage of times when the retrieved visual embeddings match the true ones is among the top-1 (denoted as R@1), top-10 (R@10), or top-50 (R@50) records (in the descending order based on cosine similarity). We also find the mean reciprocal rank (MMR) for VATs as

$$\text{MMR} = \left(\frac{1}{Q} \sum_{i=1}^{Q} \frac{1}{r_i} \right) \times 100, \tag{9}$$

where Q is the total number of queries and r_i is the rank of the most relevant embeddings in the sorted list. In other words, a small r_i implies high MMR and accurate information retrieval.

6.2 Observations and Results

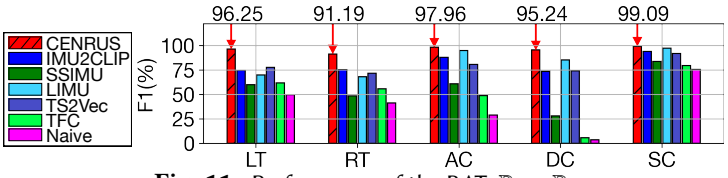

Fig. 11. Performance of the BAT: $\mathbb{B} \rightarrow \mathbb{B}$.

6.2.1 Overall Performance: As for BAT, we showcase the performance of CENRUS and other baselines in Fig. 11. Our evaluation further includes the BAT without CENRUS's PFT model — traditional practices of behavior learning [64] — denoted as "Naive". CENRUS, IMU2CLIP, SSIMU, LIMU, TS2Vec, and TFC, and Naive are shown to achieve 98.41%, 83.09%, 63.31%, 83.09%, 81.31%, 59.41%, and 70.19%, respectively. One can observe that Naive cannot identify such maneuvers as DC due to its misclassification of other patterns within the maneuvers of SC and AC. Note that the baselines such as SSIMU, IMU2CLIP, and LIMU may not capture the heterogeneous and cross-modality patterns, thus failing to achieve generalizability in our BAT. Other baselines such as CLIP, BLIP, ALIGN, AltCLIP, and FLAVA focus on vision-language model learning, and may not adapt themselves to the complex semantic correspondences across heterogeneous modalities in the NR settings. As a result, they are removed from the showcasing in Fig. 11. In contrast to the baselines, CENRUS disentangles the RXI scene understandings by differentiating RXI scenes and concepts, and also achieves the generalizability (25.01% higher F1 on average) for BAT. We also note that our CENRUS achieves generalizable behavior recognition particularly with our unbalanced RXI scenes and unconstrained maneuvers in our NR dataset (Figs. 5(a) and 5(b)).

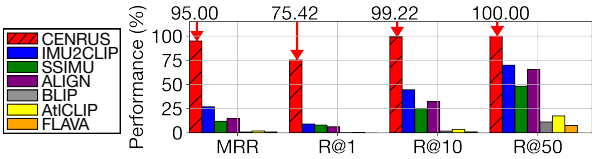

Fig. 12. Performance of the VAT (video retrievals given textual inputs): $\mathbb{T} \rightarrow \mathbb{V}$.

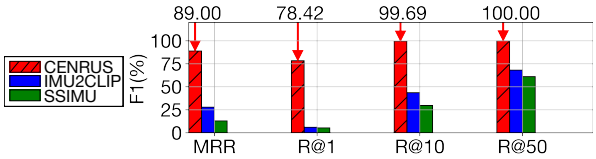

Fig. 13. Performance of the VAT (video retrievals given the behavioral inputs): $\mathbb{B} \rightarrow \mathbb{V}$.

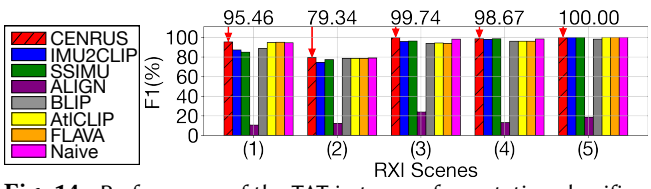

Fig. 14. Performance of the TAT in terms of annotation classification: $\mathbb{T}, \mathbb{V}, \mathbb{B} \rightarrow \mathbb{T}$.

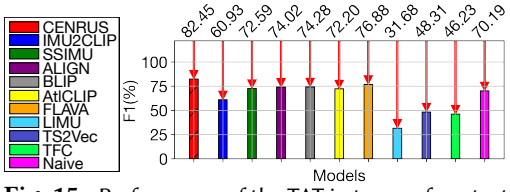

Fig. 15. Performance of the TAT in terms of context recognition: $\mathbb{B}, \mathbb{V} \rightarrow \mathbb{T}$.

As for VAT, Figs. 12 and 13 show the evaluation results of CENRUS and other baselines. Regarding the video retrieval task given $\mathbb{T}$ ($\mathbb{T} \rightarrow \mathbb{V}$) in Fig. 12, we can observe at best 27%, 9.23%, 44.57%, and 69.98% in terms of MMR, R@1, R@10, and R@50 (from most to least difficult) for the baselines, while CENRUS achieves much better overall performance (54.15% on average) than the baselines. For the video retrieval task given $\mathbb{B}$ ($\mathbb{B} \rightarrow \mathbb{V}$) in Fig. 13, the baselines achieve, at best, 28%, 6.05%, 43.62%, and 68.21% in terms of MMR, R@1, R@10, and R@50, while

CENRUS achieves 42.95% better, on average, in each metric than the baselines. CLIP, BLIP, ALIGN, AltCLIP, and FLAVA are excluded due to failing to adapt to behavioral data (time-series).

As for TAT, Figs. 14 and 15 illustrate the overall performance of CENRUS in annotation classification ($\mathbb{T}+\mathbb{V}+\mathbb{B} \rightarrow \mathbb{T}$) and context recognition ($\mathbb{V}+\mathbb{B} \rightarrow \mathbb{T}$), respectively. CENRUS is found to outperform the baselines (including IMU2CLIP and SSIMU dedicated to $\mathbb{B}$) by 13.38% on average, demonstrating its important generalizability toward DTs in processing the complex behavioral modalities for the text-aware learning. Furthermore, in terms of the context recognition within Fig. 15, CENRUS is shown to be able to make overall 19.71% improvements on average over the baselines in understanding (classifying) the correct contexts in the RXI scenes. One may leverage such a PFT outcome to enable future context-aware advanced rider assistance systems [64] (e.g., enhancing rider safety given recognized contexts and assessed risks) or RXI scene interpretation for causal analysis of micromobility accidents [43].

6.2.2 PFT Ablation Studies: We have considered the BAT as an instance to perform the model ablation studies upon the components within the PFT in Fig. 16(a). Experimentation on other DTs is left out due to space

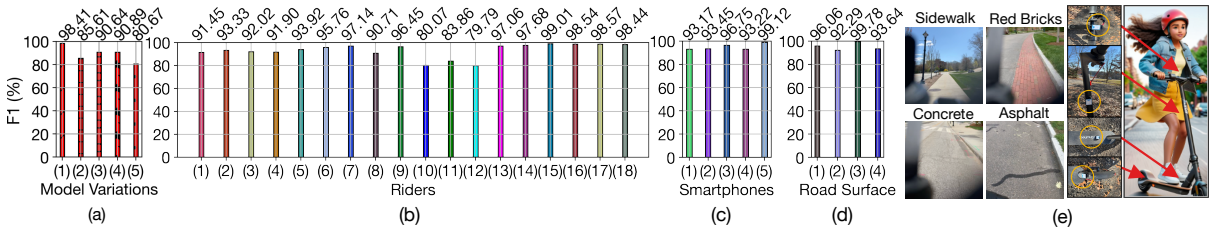

Fig. 16. (a) Performance results of CENRUS's ablation studies; (b)–(d) Impacts of different riders, devices, and road conditions on CENRUS's performance; (e) Showcased different road conditions and IMU placements.

limitation. Specifically, we have compared the complete CENRUS (labeled as (1)) with respect to variations (removal of a certain design component) of: (2) w/o SA; (3) w/o the global scene contrasting; (4) w/o the local concept disentanglement; and (5) w/o the RXI graphs.

The performance drop (12.80%) of (2) compared to (1) corroborates the importance of SA in expanding the PFT's learnability upon the complex RXI scenes. We also note the performance degradation of 7.77% or 7.52% for removal of either (3) global scene contrasting or (4) local concept disentanglement. Both operations are essential for differentiating the RXI scenes and modalities within the NR data. In addition, the 17.74% reduction without the RXI graph representations demonstrates the importance of our representation learning in constructing the semantic interactions.

6.2.3 Sensitivity Studies: Despite an unconstrained and naturalistic setting, we have also examined the impacts of individual differences (18 riders; leave-one-out), smartphones plus an on-board IMU sensor, as well as road surface conditions (sidewalk, red bricks, concrete, and asphalt) in Figs. 16(b), 16(c), and 16(d). Fig. 16(b) shows CENRUS to achieve in general high accuracy across the involved participants. In Fig. 16(c), regarding different smartphone/sensor types, we let a rider use iPhone 13 Plus, 7 Plus, SE 2022, Xiaomi Note 8, and an on-board 9DOF IMU TDK MPU9250 [12] as a reference (labeled as (1)—(5)). From the three studies, we have observed overall minor standard deviations (4.53, 2.13, and 2.47) in terms of F1 scores, demonstrating the overall generalizability of our PFT in adapting to complex NR data. Such a result also implies the usability of CENRUS for ubiquitous and rider-centered applications, such as adapting to various riders' NR preferences and collection settings [3].

As illustrated in Figs. 16(d) and (e), we have also empirically studied various placement locations of MPU9250 IMU upon the e-scooter, i.e., handlebar, pillar, board, and rear end, and observed that CENRUS achieves 99.64%±0.31% F1 score for BAT, on average, across different IMU sensor placement positions. Such generalizability of CENRUS provides implications of integrating CENRUS with on-board IMU sensors for ubiquitous e-scooter sharing systems or autonomous micromobility vehicles for last-mile delivery and elderly people. Further generalizability can be achieved through various measures, such as meta or transfer learning [67], and lifelong learning [51], which are left as our future work.

6.2.4 Visualization and Interpretation: In order to further close the loop between the interior model learning and the exterior physical-world interpretation, we have included Fig. 17 to provide insights on CENRUS's disentanglement of understandings from the human NR data. In particular, we have conducted the model explanation studies upon the RXI scenes of (a) rider-to-car (954 samples), (b) rider-to-pedestrian (128 samples), and (c) rider-to-environment (63 samples). We first show in the columns (i) and (ii) with the 2-D t-SNE projection [73] of visual, behavioral, and textual embeddings before and after the PFT. We can observe, in terms of feature representation (embeddings) level [18], our CENRUS aligns the previously-scattered heterogeneous modalities (green: textual; blue:

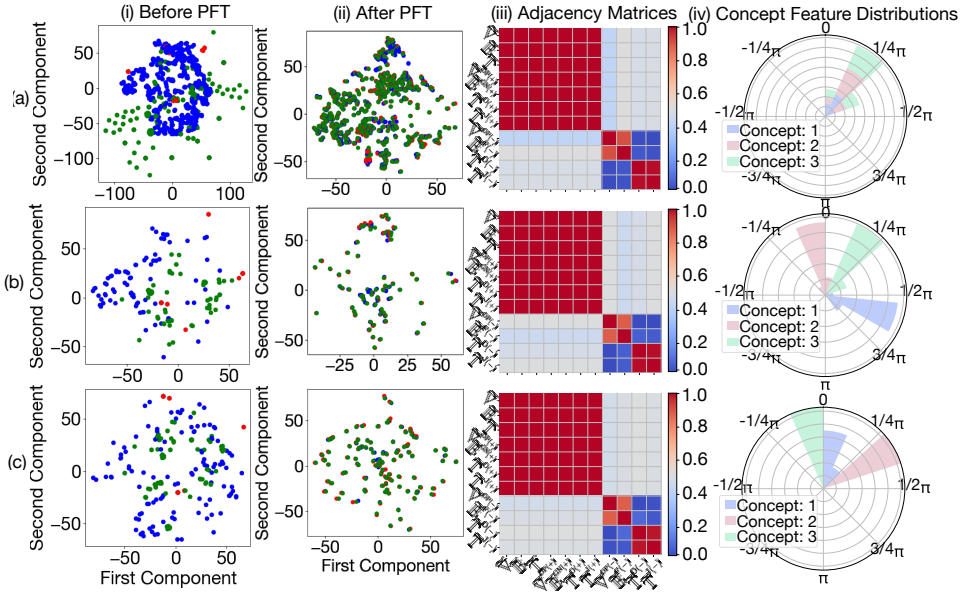

Fig. 17. Disentangled understandings from three RXI scenes in terms of (i) projected embeddings before PFT; (ii) projected embeddings after PFT; (iii) learned adjacency matrices of the RXI graphs; and (iv) feature distributions of the learned concepts.

behavioral; red: visual) for each of the RXI scenes (a scene is represented by three dots of different modalities) in the projected space. This demonstrates CENRUS's learnability, thanks to the hierarchical RXI graph representation learning, in disentangling the understanding of the complex NR data by grouping modalities towards RXI scenes.

We also include examples of learned RXI graph adjacency matrices in column (iii), visualizing the semantic OSSA-PSSA interactions learned. In column (iv), we demonstrate the top three concepts (the highest weights in $\mathbf{W}$) learned within the local concept disentanglement operation. As for each RXI scene category, we project each embedding of the top three concepts (1 to 3) disentangled through t-SNE [73], and find the angle distributions of the resulting 2-D vectors in an angular histogram (15 bins in total; larger radius means more vectors grouped). We have found for each type of RXI scene, i.e., rider-to-car, to-pedestrian, and to-infrastructure, the average

angles (unit: rad) of vectors representing the top three concepts as [0.803, 0.909, 0.884], [2.139, 0.051, 0.920], and [0.576, 1.279, 0.030] (we note that the majority of some concepts like 3 in (c) is barely below zero with a few positive angles). These are differentiated, disentangled, and vectorized based on our current NR data. That is, our PFT derives disentangled understandings for differentiating the RXI scenes, and characterizes (fingerprints) the three types of RXI scenes via the important concepts, and hence can provide semantic correspondences and support various DTs.

7 Deployment Discussion

7.1 NR Data Collection Ethics and Privacy

While in our NR studies, all the participants, as accountable riders, followed the local regulations and riding ethics throughout the studies, we also notice the needs of reducing conflicts and micromobility-related accidents, as well as the imperatives for the stakeholders in regulating the riding behaviors [3]. Our NR data collection studies for CENRUS and the potential integration with the existing micromobility infrastructures will benefit the manufacturers, city planners, sharing service providers (e.g., the e-scooter sharing), and other stakeholders (say, insurance companies) in (a) promoting safer and more accountable riding behaviors and (b) improving vehicle interfaces with the riders for enhanced riding experience and informed decision-making. Furthermore, despite the current focuses on the ubiquitous and human-centered AI studies, in future we will also consider data security and privacy-preserving measures (including privacy-preserving federated learning [14, 15]).

7.2 System Optimization and Integration

Given the rise of ubiquitous computing and Internet of Things (IoTs), contemporary micromobility vehicles are capable of providing situational awareness, performing computation and communication with the central server (in the context of micromobility sharing) or the riders' smartphones. One may envision that future micromobility systems will become smarter, and many of them will be equipped with the better battery and motor designs to power all these sensors and computing capabilities. We envision that the potential increase of the IoT costs, as well as power consumption, will be outweighed by the tangible benefits in enhancing the safety, efficiency, and fun in the future micromobility systems. Also, there exist various efforts (lowered frame rates and IMU sampling rates) and techniques [35, 37] to reduce computationally-intensive model training upon the resource-constrained embedded devices. CENRUS can be easily extended to include the computation and battery optimization and integration of low-power sensing and computing modules [35, 37]. However, they are orthogonal to the scope of this paper and will be part of our our future work.

7.3 Broader Experimental Studies:

CENRUS is general enough to be extended to broader NR data collection settings. We have observed CENRUS to achieve overall generalizable performance across various settings (e.g., riders, sensor installation positions, road conditions). Our current experimental studies in a college town environment, which resembles the urban settings and already involves a diverse set of common RXI scenes. Since varying the test environments (say, more cities and towns), cultural contexts (e.g., different neighborhoods), and traffic conditions will be subject to further compliance reviews relevant to local transportation regulations [3], we will expand these studies in our future work.

8 Conclusion

We have designed CENRUS, a novel cross-modality e-scooter naturalistic riding understanding system for HCAI insights on the rider-to-X interactions (RXIs). We have designed a novel PFT based on hierarchical RXI graph representation learning to characterize the semantic OSSA-PSSA interactions across the RXI scenes for various

DTs with awareness of behavioral, visual, and textual modalities. We have conducted extensive e-scooter NR experimental studies, and corroborated the effectiveness of CENRUS in disentangling NR understandings, and its potential of hailing for and supporting future human-centered micromobility systems.

ACKNOWLEDGMENT

We would like to thank the editors and the anonymous reviewers for their constructive comments. We would like to thank all the participants for their assistance in collecting the naturalistic riding data. This project is supported, in part, by the National Science Foundation (NSF) under Grant No. 2239897 and 2245223, Google Research Scholar Program Award (2021–2022), and NVIDIA Applied Research Accelerator Program Award (2021–2022). We would like to thank Connecticut Transportation Institute (CTI) for their support of our research projects. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

References

[1] Rusul L Abduljabbar, Sohani Liyanage, and Hussein Dia. 2021. The role of micro-mobility in shaping sustainable cities: A systematic literature review. *Transportation Research Part D: Transport and Environment* (2021).

[2] Alsaleh, Rushdi and Hussein, Mohamed and Sayed, Tarek. 2020. Microscopic behavioural analysis of cyclist and pedestrian interactions in shared spaces. *Canadian Journal of Civil Engineering* (2020).

[3] Cynthia Bennett, Emily Ackerman, Bonnie Fan, Jeffrey Bigham, Patrick Carrington, and Sarah Fox. 2021. Accessibility and the crowded sidewalk: Micromobility's impact on public space. In *Proc. ACM DIS*. 365–380.

[4] Brunner, Pascal and Löcken, Andreas and Denk, Florian and Kates, Ronald and Huber, Werner. 2020. Analysis of experimental data on dynamics and behavior of e-scooter riders and applications to the impact of automated driving functions on urban road safety. In *IEEE IVS*.

[5] German Castignani, Thierry Derrmann, Raphaël Frank, and Thomas Engel. 2017. Smartphone-based adaptive driving maneuver detection: A large-scale evaluation study. *IEEE T-ITS* (2017).

[6] Dongyao Chen, Kyong-Tak Cho, Sihui Han, Zhizhuo Jin, and Kang G. Shin. 2015. Invisible Sensing of Vehicle Steering with Smartphones. In *Proc. ACM MobiSys*. 1–13.

[7] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. 2020. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proc. IEEE/CVF CVPR*.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proc. ICML*. PMLR, 1597–1607.

[9] Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Fulong Ye, Qinghong Yang, and Ledell Wu. 2022. AltCLIP: Altering the language encoder in CLIP for extended language capabilities. *arXiv preprint arXiv:2211.06679* (2022).

[10] Sanghyun Choo and Wonjoon Kim. 2023. A study on the evaluation of tokenizer performance in natural language processing. *Applied Artificial Intelligence* 37, 1 (2023), 2175112.

[11] Sook-Ling Chua, Stephen Marsland, and Hans Guesgen. 2011. Unsupervised learning of human behaviours. In *Proc. AAAI*, Vol. 25. 319–324.

[12] TDK Corporation. 2023. 9DOF Motion Sensor TDK MPU9250. https://store.rakwireless.com/products/9dof-motion-sensor-tdk-mpu9250-rak1905.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[14] Fatima Elhattab, Sara Bouchenak, and Cédric Boscher. 2024. PASTEL: Privacy-Preserving Federated Learning in Edge Computing. *Proc. ACM IMWUT* 7, 4 (2024), 1–29.

[15] Jie Feng, Can Rong, Funing Sun, Diansheng Guo, and Yong Li. 2020. PMF: A privacy-preserving human mobility prediction framework via federated learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–21.

[16] Christina MR Garman, Steven G Como, Ian C Campbell, Jeffrey Wishart, Kevin O'Brien, and Scott McLean. 2020. *Micro-mobility vehicle dynamics and rider kinematics during electric scooter riding*. Technical Report. SAE Technical Paper.

[17] Samineh C Gillmore and Nathan L Tenhundfeld. 2020. The good, the bad, and the ugly: Evaluating Tesla's human factors in the wild west of self-driving cars. In *Proc. the Human Factors and Ergonomics Society Annual Meeting*, Vol. 64. SAGE Publications Sage CA: Los Angeles, CA, 67–71.

[18] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT press.

[19] Stefan Gössling. 2020. Integrating e-scooters in urban transportation: Problems, policies, and the prospect of system change. *Transportation Research Part D: Transport and Environment* (2020).
[20] Jiawei Han, Jian Pei, and Hanghang Tong. 2022. *Data Mining: Concepts and Techniques*. Morgan kaufmann.
[21] Zengyi Han, Xuefu Dong, Yuuki Nishiyama, and Kaoru Sezaki. 2023. HeadSense: Visual Search Monitoring and Distracted Behavior Detection for Bicycle Riders. In *Proc. IEEE WoWMoM*. IEEE.
[22] Zengyi Han, Liqiang Xu, Xuefu Dong, Yuuki Nishiyama, and Kaoru Sezaki. 2023. HeadMon: Head Dynamics Enabled Riding Maneuver Prediction. In *Proc. IEEE PerCom*. 22–31.
[23] Suining He and Kang G. Shin. 2020. Dynamic Flow Distribution Prediction for Urban Dockless E-Scooter Sharing Reconfiguration. In *Proc. WWW*. 133–143.
[24] Suining He and Kang G Shin. 2022. Socially-Equitable Interactive Graph Information Fusion-based Prediction for Urban Dockless E-Scooter Sharing. In *Proc. WWW*. 3269–3279.
[25] Will Douglas Heavenarchive. 2023. This driverless car company is using chatbots to make its vehicles smarter. https://www.technologyreview.com/2023/09/14/1079458/this-driverless-car-company-is-using-chatbots-to-make-its-vehicles-smarter/.
[26] Gim Hee Lee, Friedrich Faundorfer, and Marc Pollefeys. 2013. Motion estimation for self-driving cars with a generalized camera. In *Proc. IEEE/CVR CVPR*. 2746–2753.
[27] Peng Hu, Liangli Zhen, Dezhong Peng, and Pei Liu. 2019. Scalable deep multimodal learning for cross-modal retrieval. In *Proc. ACM SIGIR*. 635–644.
[28] Boris Ivanovic, Edward Schmerling, Karen Leung, and Marco Pavone. 2018. Generative modeling of multimodal multi-human behavior. In *Proc. IEEE/RSJ IROS*. IEEE, 3088–3095.
[29] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proc. ICML*. PMLR.
[30] Ming Jin, Qingsong Wen, Yuxuan Liang, Chaoli Zhang, Siqiao Xue, Xue Wang, James Zhang, Yi Wang, Haifeng Chen, Xiaoli Li, et al. 2023. Large Models for Time Series and Spatio-Temporal Data: A Survey and Outlook. *arXiv preprint arXiv:2310.10196* (2023).
[31] Peng Jin, Hao Li, Zesen Cheng, Jinfa Huang, Zhennan Wang, Li Yuan, Chang Liu, and Jie Chen. 2023. Text-Video Retrieval with Disentangled Conceptualization and Set-to-Set Alignment. *Proc. IJCAI* (2023), arXiv–2305.
[32] Eunji Kim, Hanyoung Ryu, Hyunji Oh, and Namwoo Kang. 2022. Safety monitoring system of personal mobility driving using deep learning. *Journal of Computational Design and Engineering* 9, 4 (2022).
[33] Jinkyu Kim and John Canny. 2017. Interpretable learning for self-driving cars by visualizing causal attention. In *Proc. IEEE ICCV*. 2942–2950.
[34] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. 2018. Textual explanations for self-driving vehicles. In *Proc. ECCV*. 563–578.
[35] Jaejun Ko, Jongwon Lee, and Young-June Choi. 2017. Computation offloading for energy efficiency of smart devices. In *Proc. ACM UbiComp*. 109–112.
[36] Pantelis Kopelias, Elissavet Demiridi, Konstantinos Vogiatzis, Alexandros Skabardonis, and Vassiliki Zafiropoulou. 2020. Connected & autonomous vehicles–Environmental impacts–A review. *Science of the Total Environment* (2020).
[37] Seulki Lee, Bashima Islam, Yubo Luo, and Shahriar Nirjon. 2019. Intermittent learning: On-device machine learning on intermittently powered system. *Proc. ACM IMWUT* 3, 4 (2019), 1–30.
[38] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proc. ICML*. PMLR.
[39] Sheng Li and Handong Zhao. 2021. A survey on representation learning for user modeling. In *Proc. IJCAI*. 4997–5003.
[40] Li, Max Guangyu and Jiang, Bo and Che, Zhengping and Shi, Xuefeng and Liu, Mengyao and Meng, Yiping and Ye, Jieping and Liu, Yan. 2019. DBUS: Human Driving Behavior Understanding System.. In *ICCV Workshops*.
[41] Andreas Löcken, Pascal Brunner, and Ronald Kates. 2020. Impact of hand signals on safety: Two controlled studies with novice e-scooter riders. In *Proc. ACM AutomotiveUI*.
[42] Haojie Ma, Zhijie Zhang, Wenzhong Li, and Sanglu Lu. 2021. Unsupervised human activity representation learning with multi-task deep clustering. *Proc. ACM IMWUT* (2021).
[43] Qingyu Ma, Hong Yang, Alan Mayhue, Yunlong Sun, Zhitong Huang, and Yifang Ma. 2021. E-Scooter safety: The riding risk analysis based on mobile sensing data. *Accident Analysis & Prevention* 151 (2021), 105954.
[44] Qingyu Ma, Hong Yang, and Zizheng Yan. 2023. Use of Mobile Sensing Data for Assessing Vibration Impact of E-Scooters with Different Wheel Sizes. *Transportation Research Record* (2023).
[45] Matviienko, Andrii and Müller, Florian and Schön, Dominik and Fayard, Régis and Abaspur, Salar and Li, Yi and Mühlhäuser, Max. 2022. E-ScootAR: Exploring Unimodal Warnings for E-Scooter Riders in Augmented Reality. In *Proc. ACM CHI*.
[46] Thomas Monninger, Julian Schmidt, Jan Rupprecht, David Raba, Julian Jordan, Daniel Frank, Steffen Staab, and Klaus Dietmayer. 2023. SCENE: Reasoning about traffic scenes using heterogeneous graph neural networks. *IEEE RA-L* 8, 3 (2023), 1531–1538.

[47] Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Alireza Dirafzoon, Aparajita Saraf, Amy Bearman, and Babak Damavandi. 2022. IMU2CLIP: Multimodal Contrastive Learning for IMU Motion Sensors from Egocentric Videos and Text. *arXiv preprint arXiv:2210.14395* (2022).
[48] Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal named entity recognition for short social media posts. *arXiv preprint arXiv:1802.07862* (2018).
[49] Library of America. 2017. *Taming the Bicycle*, From Mark Twain (1835–1910): Collected Tales, Sketches, Speeches, & Essays 1852–1890. https://storyoftheweek.loa.org/2017/11/taming-bicycle.html.
[50] Society of Automotive Engineers (SAE). 2022. Vehicle Dynamics Terminology J670_202206. https://www.sae.org/standards/content/j670_202206/.
[51] Oleksandra Poquet and Maarten De Laat. 2021. Developing capabilities: Lifelong learning in the age of AI. *British Journal of Educational Technology* 52, 4 (2021), 1695–1708.
[52] Radford, Alec and Kim, Jong Wook and Hallacy, Chris and Ramesh, Aditya and Goh, Gabriel and Agarwal, Sandhini and Sastry, Girish and Askell, Amanda and Mishkin, Pamela and Clark, Jack and others. 2021. Learning transferable visual models from natural language supervision. In *Proc. ICML*.
[53] Ranftl, René and Lasinger, Katrin and Hafner, David and Schindler, Konrad and Koltun, Vladlen. 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI* (2020).
[54] Daniel J Reck, He Haitao, Sergio Guidon, and Kay W Axhausen. 2021. Explaining shared micromobility usage, competition and mode choice by modelling empirical data from Zurich, Switzerland. *Transportation Research Part C: Emerging Technologies* (2021).
[55] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proc. IEEE CVPR*.
[56] Nils Rethmeier and Isabelle Augenstein. 2023. A Primer on Contrastive Pretraining in Language Processing: Methods, Lessons Learned, and Perspectives. *ACM Computing Surveys (CSUR)* 55, 10 (2023), 1–17.
[57] Wolfgang Roth, Günther Schindler, Bernhard Klein, Robert Peharz, Sebastian Tschiatschek, Holger Fröning, Franz Pernkopf, and Zoubin Ghahramani. 2024. Resource-efficient neural networks for embedded systems. *JMLR* 25, 50 (2024), 1–51.
[58] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. FLAVA: A foundational language and vision alignment model. In *Proc. IEEE/CVR CVPR*.
[59] David Sirkin, Nikolas Martelaro, Mishel Johns, and Wendy Ju. 2017. Toward measurement of situation awareness in autonomous vehicles. In *Proc. CHI*.
[60] Yongjae Sohn, Haeun Lee, Yelim Lee, Taeyun Kim, Youkeun Oh, and Dokshin Lim. 2022. Interaction Design of a Smart Helmet for Micro-Mobility Riders. *Human Interaction & Emerging Technologies: Artificial Intelligence & Future Applications* (2022).
[61] Raffaele Soloperto, Philipp Wenzelburger, David Meister, Dominik Scheuble, Veronika SM Breidohr, and Frank Allgöwer. 2021. A control framework for autonomous e-scooters. *IFAC-PapersOnLine* (2021).
[62] Jithesh Gugan Sreeram, Xiao Luo, and Renran Tian. 2021. Contextual and Behavior Factors Extraction from Pedestrian Encounter Scenes Using Deep Language Models. In *International Conference on Big Data Analytics and Knowledge Discovery*. Springer.
[63] Stigson, H and Malakuti, Iman and Klingegård, M. 2021. Electric scooters accidents: Analyses of two Swedish accident data sets. *Accident Analysis & Prevention* (2021).
[64] Mahan Tabatabaie and Suining He. 2023. Naturalistic E-Scooter Maneuver Recognition with Federated Contrastive Rider Interaction Learning. *Proc. ACM IMWUT* 6, 4, Article 205 (jan 2023), 27 pages.
[65] Mahan Tabatabaie, Suining He, and Kang G. Shin. 2023. Cross-Modality Graph-Based Language and Sensor Data Co-Learning of Human-Mobility Interaction. *Proc. ACM IMWUT* 7, 3, Article 125 (sep 2023), 25 pages.
[66] Mahan Tabatabaie, Suining He, and Kang G Shin. 2023. Interaction-Aware and Hierarchically-Explainable Heterogeneous Graph-based Imitation Learning for Autonomous Driving Simulation. In *Proc. IEEE/RSJ IROS*.
[67] Mahan Tabatabaie, Suining He, and Xi Yang. 2022. Driver Maneuver Identification with Multi-Representation Learning and Meta Model Update Designs. *Proc. ACM IMWUT* 6, 2, Article 74 (July 2022), 23 pages.
[68] Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proc. ICML*. PMLR.
[69] Shuhan Tan, Tushar Nagarajan, and Kristen Grauman. 2023. EgoDistill: Egocentric Head Motion Distillation for Efficient Video Understanding. *arXiv preprint arXiv:2301.02217* (2023).
[70] Satoshi Tsutsui, Ruta Desai, and Karl Ridgeway. 2021. How You Move Your Head Tells What You Do: Self-supervised Video Representation Learning with Egocentric Cameras and IMU Sensors. *arXiv preprint arXiv:2110.01680* (2021).
[71] Sylvaine Tuncer and Barry Brown. 2020. E-scooters on the ground: Lessons for redesigning urban micro-mobility. In *Proc. ACM CHI*.
[72] Ultralytics. 2022. Yolov5. Retrieved October 3, 2022 from https://github.com/ultralytics/yolov5
[73] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *JMLR* 9, 11 (2008).
[74] David A Van Dyk and Xiao-Li Meng. 2001. The art of data augmentation. *Journal of Computational and Graphical Statistics* 10, 1 (2001), 1–50.

[75] Vaswani, Ashish and Shazeer, Noam and Parmar, Niki and Uszkoreit, Jakob and Jones, Llion and Gomez, Aidan N and Kaiser, Łukasz and Polosukhin, Illia. 2017. Attention is all you need. *Proc. NeurIPS* (2017).
[76] Veličković, Petar and Cucurull, Guillem and Casanova, Arantxa and Romero, Adriana and Lio, Pietro and Bengio, Yoshua. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).
[77] Xin Wang, Hong Chen, Yuwei Zhou, Jianxin Ma, and Wenwu Zhu. 2023. Disentangled Representation Learning for Recommendation. *IEEE T-PAMI* 45, 1 (2023), 408–424.
[78] Jonathan J Webster and Chunyu Kit. 1992. Tokenization as the initial phase in NLP. In *Proc. ACL*.
[79] Philipp Wenzelburger and F Allgower. 2020. A first step towards an autonomously driving e-scooter. In *Proc. IFAC World Congress*.
[80] Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. 2021. Self-supervised graph learning for recommendation. In *Proc. ACM SIGIR*. 726–735.
[81] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144* (2016).
[82] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. 2018. Unified perceptual parsing for scene understanding. In *Proc. ECCV*. 418–434.
[83] Huatao Xu, Pengfei Zhou, Rui Tan, Mo Li, and Guobin Shen. 2021. LIMU-BERT: Unleashing the potential of unlabeled data for IMU sensing applications. In *Proc. ACM SenSys*. 220–233.
[84] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. Graph contrastive learning with augmentations. *Proc. NeurIPS* 33 (2020), 5812–5823.
[85] Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. 2022. TS2VEC: Towards universal representation of time series. In *Proc. AAAI*. 8980–8987.
[86] Ming Zeng, Haoxiang Gao, Tong Yu, Ole J Mengshoel, Helge Langseth, Ian Lane, and Xiaobing Liu. 2018. Understanding and improving recurrent networks for human activity recognition by continuous attention. In *Proc. ACM ISWC*.
[87] Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. 2022. Self-supervised contrastive pre-training for time series via time-frequency consistency. *Proc. NeurIPS* (2022), 3988–4003.
[88] Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. 2023. A comprehensive survey on pretrained foundation models: A history from BERT to ChatGPT. *arXiv preprint arXiv:2302.09419* (2023).