Temporal Specification Construct Undercoverage in Pedestrian Behavioral Modeling

Calvin Breseman[1*], Igor Moiseev[1], and David M. Woollard[1]

Abstract— **Classification of behavior into discrete categories from human trajectories is widely used in pedestrian modeling, including robotics. Higher-level behaviors are inferred from motion primitives such as velocity, dwell time, or proximity to environmental features and are used to structure trajectory datasets and prediction tasks, compressing uncertainty about future motion into the likelihood of a particular behavioral regime. This can be useful in applications where future intent is more predictive of pedestrian motion than path dependence. Yet classifying behavior that is continuous and heterogeneous presents difficulty. Previous research has shown that operational definitions of behavioral classes shape the resulting representation of behavior by embedding assumptions about how a target construct is expressed. We extend this research by showing that even when the behavioral definition is held constant, temporal ordering choices within the classification pipeline can systematically exclude coherent behavioral episodes that satisfy that definition. Using pose-derived pedestrian trajectories collected in a live retail environment, we examine browsing, a common exploratory behavior in indoor spaces. Comparing indicator-first classification pipelines with a structure-first trajectory segmentation approach, we show that a change in order of operations alone can push coherent episodes outside the classifiable behavioral space, a phenomenon we characterize as temporal specification construct undercoverage. Our results highlight how modest design decisions in pedestrian behavior classification can materially shape what behaviors a system represents, with downstream consequences for inference.**

I. INTRODUCTION

Understanding human behavior from motion trajectories is central to pedestrian modeling and prediction [1]. In robotics, autonomous navigation, human–robot interaction, and intelligent environments, behavioral abstractions are often used to anticipate intent, structure interaction, and condition downstream predictions [2], [3]. They help determine what kinds of human motion are legible to a system and, therefore, the behavior a robot can learn to expect [1]. Because behavioral constructs are not directly observable in trajectories, they must be inferred from indicators whose operationalization shapes what becomes measurable [4], [5], [6], [7]. We extend that reasoning by assessing an important yet often overlooked decision point in behavioral classification: how to handle time. This issue arises across pipelines that map observable trajectory features to higher-order behavioral categories [8], [9]. Such approaches offer interpretable links between continuous motion data and behavioral states, and variants of them appear across pedestrian behavior recognition, indoor activity modeling, and retail analytics [3], [10], [11] Yet since behavior extends through time, these systems must also decide what constitutes a behavioral episode: whether a behavior is built from instants, or whether it is a property of a larger temporal structure. Although this choice is often treated as procedural, or downstream of model choice, we argue that it is substantive: it constrains the class of behavioral realizations that the resulting pipeline can represent even when the behavioral definition remains fixed.

Behaviors vary with environment and individual. A pipeline that classifies framewise indicators and aggregates them treats behavior differently than one that classifies segmented temporal units. In effect, the order of operations influences the feasible manifold: it determines which regions of trajectory space can be captured by the classifier as the target behavior. We examine this problem through browsing, a form of exploratory behavior that is both important and temporally diffuse in environments such as retail spaces. Using pose-derived pedestrian trajectories collected in a live retail setting, we compare two classification pipelines applied to identical trajectory data [12]. The first follows an *indicator-first* ordering, in which frame-level trajectory primitives are classified before being aggregated. The second, a *structure-first* ordering, where temporally coherent segments are first identified and semantic interpretation is applied subsequently. Our target is held constant across the two pipelines; the difference is the order evidence is organized.

Our analysis reveals systematic disagreement between the resulting classifications of these treatments. In particular, we identify coherent behavioral episodes that are consistently omitted by indicator-first specifications even after calibration and model variation. We describe this phenomenon as *temporal specification construct undercoverage*: the narrowing of the classifiable space due to decisions that embed assumptions about the forms the target behavior may take. In this sense, apparently modest implementation choices can alter the effective ontology of the classification problem even after the formal behavioral definition has ostensibly been fixed.

This has direct implications for robotics [1]. Operational decisions such as the treatment of temporal association influence which parts of the behavioral space are associated with the target construct. This can push valid forms of the behavior outside the set of trajectories that can be realized as positive classifications under that ordering of operations. Downstream systems may therefore learn an incomplete behavioral world model, in which forms of human conduct are absent not because they do not occur, but because the pipeline cannot represent them. Because these exclusions arise from design-time ordering choices, they cannot be recovered by later calibration, threshold adjustment, or model

[1]Standard AI, San Francisco, California USA
∗ Corresponding author: calvin.breseman@standard.ai

tuning [13]. For robotic systems operating in shared human environments, this matters because representational blind spots introduced upstream can propagate into prediction, planning, and interaction. If a classification pipeline renders some legitimate forms of human motion unclassifiable, then those forms may also become less predictable, less interpretable, or less actionable to the autonomous systems trained on its outputs.

II. RELATED WORK

A. Pedestrian Trajectory Modeling and Prediction

Understanding human behavior from motion trajectories is a central problem in pedestrian modeling and prediction. Recent work has produced a wide range of learning-based approaches that forecast pedestrian motion from observed trajectory histories, including recurrent architectures, graph-based interaction models, and attention-based forecasting systems [1]. Representative examples include recurrent interaction models such as Social-LSTM [2] and graph-based forecasting frameworks such as Trajectron++ [14]. Many pipelines rely on behavioral abstractions derived from trajectory dynamics, such as intent, interaction states, or exploratory motion patterns. These abstractions are useful for structuring prediction tasks, organizing datasets, or interpreting model outputs [2]. They also provide a bridge between low-level motion primitives and high-level behavioral reasoning used in navigation, human–robot interaction, and intelligent environments, reducing uncertainty over plausible future actions by focusing prediction on trajectories associated with the currently inferred state and its likely transitions [15], [16]. In robotics applications, such abstractions help determine which human states are most legible to systems that depend on behavioral prediction in order to plan, anticipate, or adapt.

B. Behavioral Inference from Trajectory Data

A large body of work infers higher-level behavioral states from trajectory-derived signals such as velocity, dwell time, or spatial proximity to environmental structures. Early retail and mobility studies used trajectory analysis to characterize shopping paths and exploratory movement patterns in physical environments [8], [9]. More recent sensing systems similarly infer behaviors from indoor location traces collected using vision, wearable sensors, or RFID infrastructure [3], [10], [11]. Across these domains, behavioral states are typically defined through operational criteria applied to motion primitives, making indicator-based behavioral classification a common approach for interpreting trajectory data.

C. Construct Undercoverage

Our argument also connects to broader work on measurement and classification. Measurement theory emphasizes that many important constructs are not directly observable and must instead be inferred through operational definitions grounded in observable indicators [5]. In parallel, work in sociotechnical systems has shown that classification schemes, category systems, and problem formulations do not just describe the world but actively shape the representations systems produce [6], [7], [13]. These literatures establish that operational decisions embed assumptions about the target construct. Within validation theory, Messick identified *construct underrepresentation* as a core threat to validity: a measurement procedure may be too narrow and fail to include important dimensions of the construct it aims to assess [4]. This idea is especially relevant for higher-order constructs, which are often heterogeneous.

D. Segmentation and Behavioral Classification

Because behavior unfolds over time, trajectory analysis must decide how temporal continuity is handled. One family of approaches classifies behavior locally at the level of frames, windows, or short observations and then aggregates those classifications into longer episodes. Another family first partitions a trajectory into temporal units and then assigns behavioral meaning to those units. This distinction appears across sequence modeling more broadly. In adjacent sequence-analysis literatures, the difference between frame-wise classification and temporal classification over unsegmented sequences has long been recognized as a substantive modeling distinction [17]. Likewise, temporal action segmentation work in computer vision treats the relation between local evidence, temporal continuity, and action boundaries as a central problem in its own right [18]. Related distinctions also appear specifically in trajectory analysis. TRACLUS, for example, explicitly partitions trajectories into line segments in order to recover common sub-trajectories that whole-trajectory methods may miss [12]. More recent work on semantic segmentation of pedestrian trajectories seeks to divide trajectories into semantically meaningful segments before higher-level interpretation [19]. Temporal partitioning is often indispensable for discovering local motion structure, isolating episodes, or improving interpretability.

III. TEMPORAL SPECIFICATION FRAMEWORK

A. Research Gap and Experimental Motivation

Existing discussions of undercoverage usually focus on inexact construct definitions or mismatch between a construct and indicators. Less apparent are decisions beyond the choice of behavioral class itself which can shape what forms of behavior are representable. Even work that focuses on temporal trajectory segmentation is typically motivated in terms of pattern discovery, localization, compression, or predictive performance. Much less attention has been paid to the possibility that the temporal order of operations itself changes the effective class of behavioral realizations available to the classifier. In other words, whether a system classifies local observations and aggregates afterward, or identifies coherent spans and classifies them as units, is usually treated as a modeling choice about convenience or performance. We show this choice is a representational constraint, altering which coherent behavioral episodes fall within the classifiable behavioral space induced by the pipeline.

B. Browsing as an example behavior

We study browsing as a concrete instance of an exploratory pedestrian behavior commonly observed in indoor environments such as retail spaces. While intuitively recognizable, browsing does not correspond to a single, uniquely specified pattern of motion. Plausible browsing episodes may differ in speed, pausing, revisitation, and orientation due to local geometry and individual interaction style. As previously mentioned, behavioral constructs used in pedestrian modeling are rarely directly observable from motion trajectories. Instead, they are specified through operational definitions that map observable trajectory indicators to latent behavioral constructs [5]. Browsing is well suited to represent behavioral abstractions that are practically useful and intuitively recognizable, but whose breadth is inexact.

C. Indicator-First vs. Structure-First Specifications

Browsing is well suited to examine the assumptions on behavioral construction imposed by the necessity of incorporating time. We focus on two common families of temporal specification for this behavior.

Indicator-first specifications first classify trajectory evidence using motion primitives (e.g., thresholds over velocity, dwell, proximity), then aggregate frame-level detections into temporally extended behavioral segments. This order of operations is common in deployable behavior inference systems because it is interpretable and easy to audit.

Structure-first specifications first segment trajectories into temporally coherent units using kinematic or geometric regularities (e.g., trajectory segmentation and clustering [12]), and apply semantic interpretation (e.g., "browsing") post hoc to selected segment types. This approach defers semantic constraints, but still embeds commitments through segmentation criteria and granularity.

Both approaches involve drawbacks. Structure-first methods risk creating temporal boundaries that do not neatly correspond to the change-points of the behavior in question, even though the resulting classification applies to the entire window. They may also be difficult to productionize in applications with online inference requirements. Indicator-first approaches can miss evidence of a behavior that unfolds subtly across a period rather than appearing in instantaneously recognizable form. The choice of one family of approaches over another, then, necessarily involves privileging certain conceptions of what behavior is over others.

D. Temporal Specification Construct Undercoverage

Applying different specifications to the same trajectories can yield systematic differences in which episodes are classifiable as browsing. Undercoverage denotes trajectory regimes that do not enter the representational space of a pipeline and therefore cannot be recovered by threshold tuning, relabeling, or post hoc calibration once the specification has ruled them out. We refer to *temporal specification construct undercoverage* as the omission of plausible realizations of a target behavior relative to a different temporal specification of that same behavior. Prior work has established the broader significance of operationalization in measurement and classification. What remains underexplored is how temporal order of operations in a behavioral classification pipeline shapes the feasible behavioral manifold. Our work addresses that gap by showing that seemingly modest ordering choices can systematically exclude coherent behavioral episodes even when the definition and data are held constant.

IV. EXPERIMENTAL SETUP

A. Trajectory Dataset, Representation, and Preprocessing

This section describes the empirical setting used to examine how different temporal specifications of behavioral classification operate on identical trajectory data. The goal is not to identify a correct definition of browsing, nor to compare the predictive performance of alternative temporal treatments. Rather, the aim is to test whether changing the temporal order of operations in a time-spanning classification task changes which realizations of a behavior are reachable to the classifier, even when the behavioral target is held constant and modeling choices are matched as closely as possible. The analysis uses pose-derived pedestrian trajectories collected in a live retail environment. We focus on pedestrian movement within a single aisle of a retail store to constrain environmental variation while retaining realistic exploratory behavior. Activity is captured using a multi-camera overhead vision system. Two-dimensional poses are estimated per view and fused through geometric triangulation to recover three-dimensional trajectories aligned to a store coordinate frame. Trajectories capture translational motion and coarse body orientation at 10 Hz over the portion of a visit occurring within the aisle. The sensing pipeline produces anonymized movement traces and does not assign semantic labels or identity-linked attributes; behavioral interpretation is introduced only through downstream behavioral specifications.

Each trajectory is represented as a temporal sequence of pose-derived features in a store-aligned coordinate system, allowing interpretation relative to environmental structures such as shelves and aisle boundaries. A shared preprocessing pipeline (artifact filtering, short-gap interpolation, and normalization) is applied across all trajectories. Both behavioral specifications operate on identical preprocessed trajectories. Consequently, any differences in predicted browsing segments can be attributed to differences in specification rather than sensing, preprocessing, or trajectory representation.

B. Behavioral Specifications Compared

Indicator-first specification. Browsing is inferred from frame-level trajectory primitives, including velocity, dwell time, and orientation relative to fixtures. We evaluate variants of this approach, including deterministic rulesets and a supervised classifier, representing common deployment pipelines. These variants are not intended as unrelated alternative models, but as indicator-first attempts to recover the browsing instances identified under the structure-first specification.

Frame-level predictions are aggregated into contiguous browsing segments. Since browsing prediction is at the frame level for the indicator-first specification, sequences of frames containing positive predictions are evaluated for aggregation based on three factors: the maximum time gap between positive frame level predictions, the minimum length of a merged sequence of 'browsing' frames, and the minimum proportion of frames which are classified as 'browsing' in a larger browsing segment. Positive indicators are merged together by default until the segment they would produce would violate the aggregation parameters. In the baseline deterministic model, these merging rules were hand-tuned. In the XGBoost and tuned deterministic models, these parameters were selected using Optuna to conduct a search of the parameter space for the optimal aggregation ruleset.

Structure-first specification. Trajectories are first segmented into temporally coherent units based on changes in movement structure. Segmentation targets units with internally consistent organization, after which browsing labels are assigned at the segment level through human annotation. These annotations are not treated as universal ground truth for browsing, but as one coherent instantiation of a structure-first behavioral specification against which recoverability by indicator-first pipelines can be tested.

Human labeling is used rather than a learned structure-first classifier in order to minimize confounding from model capacity, optimization, or calibration on the structure-first side. This allows the comparison to focus on whether changing the temporal order of operations alters which browsing instances are reachable when the behavioral target is otherwise held fixed. Likewise, the goal is not to establish the superiority of any particular segmentation method or behavioral architecture, but to examine whether a rationally derived indicator-first methodology can recover the positive set induced by a coherent structure-first specification.

V. Empirical Evidence

We now examine whether the positive set induced by a structure-first specification of browsing is recoverable within an indicator-first pipeline operating on the same trajectories and an unchanged understanding of the behavior. The goal is not to compare the predictive performance of model families. Rather, the comparison is designed to test whether a change in temporal specification alters which realizations of browsing are classifiable.

In this analysis, structure-first segment annotations serve as an empirical approximation of the set of trajectory episodes classifiable as browsing under a structure-first specification. Indicator-first pipelines are then evaluated against that reference set using the same preprocessed trajectory data. Reported metrics therefore quantify agreement with the reference specification rather than absolute correctness of browsing labels.

If disagreement were only a matter of imperfect tuning or model weakness, one would expect improved alignment with sufficient parameterization. If, however, the disagreement reflects a representational mismatch between temporal specifications, then certain structure-first browsing episodes should remain systematically difficult to recover within the indicator-first family.

TABLE I
Indicator-first pipelines against structure-first labels.

Model	Split	Precision	Recall	F1	IoU
Baseline deterministic	Train	0.3960	0.7425	0.5165	0.3482
Baseline deterministic	Test	0.3342	0.6736	0.4467	0.2876
Tuned deterministic	Train	0.3771	0.8192	0.5165	0.3481
Tuned deterministic	Test	0.2807	0.7744	0.4120	0.2595
XGBoost + aggregator	Train	0.6832	0.4703	0.5571	0.3861
XGBoost + aggregator	Test	0.3605	0.2727	0.3105	0.1838

A. Matching the Structure-First Target

To establish whether the disagreement between behavioral specifications can be resolved through improved modeling, we evaluate indicator-first pipelines operating on identical trajectories. These include a deterministic ruleset, a tuned deterministic variant optimized via hyperparameter search, and a supervised gradient-boosted classifier. In all cases, browsing is inferred from frame-level primitives and then aggregated into contiguous segments using the same post-processing procedure. Both the tuned deterministic model and the supervised model are optimized against the structure-first target, so the comparison is not between unrelated label spaces, but between members of the temporal specification family operating against the same reference set. Performance is evaluated against structure-first browsing segments under 3-fold cross validation. Table I summarizes the results.

Across indicator-first variants, agreement with the structure-first target is limited. Tuning and supervision alter precision and recall, but none of the indicator-first pipelines recover the structure-first set with high fidelity on held-out data. Regardless of whether the indicator-first family is hand tuned, trained directly against the segment-level target, or when the model employed changes to a performant ML classifier, substantial portions of the structure-first browsing set remain unmatched by the indicator-first family.

B. Evidence of Structural Disagreement

The next question is whether this mismatch should be interpreted as a classification error or as evidence of a more systematic exclusion. If indicator-first failures were merely residual, one would expect disagreement to be diffused across trajectory space. Instead, closer inspection shows that omitted structure-first browsing episodes are not randomly scattered, but concentrated in particular motion regimes.

To examine this, we analyze the segment-level error structure of the indicator-first pipelines using summary statistics derived from the same pose-based primitives used throughout the comparison, including velocity, acceleration, angular velocity, and distance to fixtures. We then cluster segments in this feature space using k-means to identify recurring motion regimes. Figure 1 summarizes the resulting structure for the XGBoost-based indicator-first pipeline. Several clusters are dominated by false negatives relative to the structure-first reference set, indicating regions of behavioral space that the indicator-first pipeline systematically fails to classify as browsing. Cluster diagnostics indicate that the false-negative-dominated clusters are moderately compact (average dispersion = 1.27), persist under resampling (bootstrap stability =

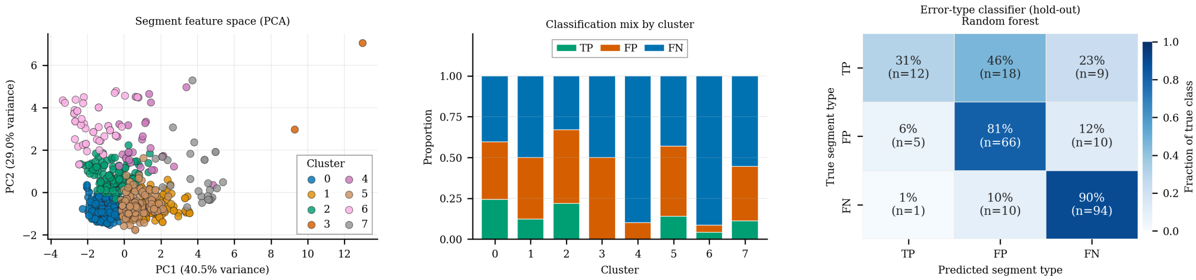

Fig. 1. Segment-level error structure for the Tuned Deterministic pipeline. False negatives predominate in clusters 4 + 6.

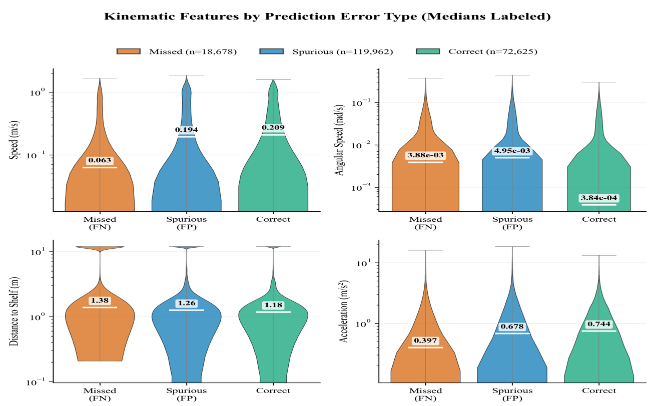

Fig. 2. Trajectory primitive distributions by outcome under the deterministic indicator-first pipeline. False negatives exhibit distinct signatures.

0.127), and show alignment with evaluative outcomes (cluster purity = 0.704, average entropy = 0.584 This suggests that indicator-first pipelines systematically fail to reach parts of the behavioral space, and suggest these parts of the space are predictable using indicator-first primitives.

C. *Prediction of Browsing False Negatives*

We then verify the omitted regimes are predictable from trajectory primitives. If false negatives can be identified from the same feature space used by the indicator-first specification, then the issue cannot be that these regimes are invisible in the underlying trajectory data. Instead, it would suggest that the indicator-first order of operations fails to assemble those signals into browsing classifications. Fig. 2 illustrates this pattern for the deterministic indicator-first pipeline. False-negative browsing segments exhibit different kinematic signatures, most notably lower speeds, than both true-positive and false-positive segments.

To quantify this structure, we train a diagnostic classifier to distinguish false negatives from true positives using only the same features used by the indicator-first specification. Under 5-fold cross-validation, this diagnostic classifier achieves an F1 score of approximately 0.7 on held-out data. Comparable predictive performance is observed when repeating the analysis across deterministic indicator-first pipelines. The diagnostic classifier is not intended as a new behavioral model. It is used only to test whether the omitted regimes are detectable from the primitive feature space itself. The same features used by the indicator-first specifications contain sufficient information to identify where those specifications fail, suggesting that these false negatives form a coherent part of the structure-first browsing space which indicator-first methods reliably fail to reach.

VI. Discussion

A. *Temporal Mechanism of Undercoverage*

The empirical pattern suggests that mismatches arise from the temporal structure of the classification procedure. In indicator-first pipelines, semantic constraints are introduced at the level of local evidence: frame-level primitives must satisfy browsing-related conditions before they can participate in a segment during aggregation. If portions of a trajectory do not meet these criteria, they may not contribute to a positive segment even when the broader temporal unit would be interpretable as browsing. Structure-first specifications reverse this order. Trajectories are first partitioned into temporally coherent units, and semantic interpretation is then applied to the segment as a whole. Because temporal coherence is established prior to classification, a different set of trajectory realizations can enter the candidate set for interpretation. The difference lies in when semantic commitments are imposed during inference, which affects which behavioral episodes are reachable as positive instances.

This interpretation is reinforced by the observed error structure. Indicator-first failures are not diffuse residuals, but concentrated in recurring motion regimes and predictable from the same primitive feature space. This suggests that omitted browsing regimes are not absent from the data, but excluded by the ordering through which local evidence becomes a behavioral segment. The relevant distinction is therefore not simply between stronger and weaker models, but between temporal specifications that induce different feasible behavioral manifolds.

B. *Temporal Specification as a Measurement Problem*

More broadly, these results suggest that temporal specification in trajectory-based inference should be understood as part of the measurement model. Behavioral constructs such as browsing, intent, or exploration are not directly observable in trajectories. They must be inferred by relating observable motion primitives to latent behavioral abstractions [5]. It is therefore already well understood that changing the formal definition of a construct alters what can be represented as an instance of that construct. Our results extend this logic by showing that temporal ordering choices can have a similar effect even when the behavioral target is held fixed.

In this, temporal specification is not merely an operational detail or a performance choice. It embeds assumptions about how a behavior is physically manifested across time and about what kinds of temporal evidence are sufficient for classification. Changing those assumptions changes which parts of behavioral space become visible to the classifier. Temporal specification should therefore be treated as a substantive design decision within the measurement process, not only as a technical choice about pipeline architecture.

C. Future Work

This paper represents a preliminary and narrow attempt to specify and define the problem of temporal specification undercoverage. As such, we intend for future work to examine generality: extensions to other diffuse and temporally extended behaviors in fora outside retail, inclusion of additional segmentation methods, alternate approaches to indicator-first behavior classification, and experimentation to determine the degree of undercoverage introduced by reversing the above order of modeling. Finally, indicator-first methods have very real advantages in robotics applications, where online inference is mandatory. Future work will examine the presentation of temporal specification undercoverage under these particular constraints.

VII. CONCLUSION

This work examined whether temporal specification can change the classifiable behavioral space of a trajectory-based inference pipeline even when the behavioral target remains fixed. Using browsing behavior in a retail environment as a case study, we compared indicator-first and structure-first specifications applied to pose-derived trajectories. Indicator-first pipelines, including tuned deterministic rules and supervised models trained against the same segment-level target, did not recover the set of browsing episodes induced by the structure-first specification. Episodes excluded by indicator-first pipelines formed coherent regions in trajectory feature space and were predictable from the same motion primitives used by the specifications, suggesting a representational constraint introduced by temporal order of operations.

We describe this phenomenon as *temporal specification construct undercoverage*: the omission of plausible realizations of a target behavior induced by the temporal structure of the classification pipeline. The central implication is that behavioral classification pipelines do not simply assign semantics to motion patterns; they help define which realizations of a behavior are available to the classifier.

For robotics and pedestrian prediction, this matters because behavioral abstractions are often used to organize datasets, guide forecasting, and support downstream reasoning about human action. If temporal specification excludes valid behavioral regimes, then learning systems built on those abstractions may inherit incomplete conceptions of the behavioral space. A robot trained on one temporal specification may therefore learn a different practical ontology of human behavior than a robot trained on another, even when both nominally target the same construct. For autonomous agents this is a major barrier to socially aware navigation. Temporal specification construct undercoverage may lead to representational blind spots for subtle pedestrian behaviors like browsing, resulting in paths that feel "unnatural" or socially disruptive. More broadly, temporal specification should be treated as part of the measurement problem in behavioral inference: a design choice that shapes not only performance, but what forms of human behavior a system can represent.

REFERENCES

[1] A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras, "Human motion trajectory prediction: A survey," *The International Journal of Robotics Research*, 2020.

[2] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *CVPR*, 2016.

[3] L. C. Shum, R. Faieghi, T. Borsook, T. Faruk, S. Kassam, H. Nabavi, S. Spasojevic, J. Tung, S. S. Khan, and A. Iaboni, "Indoor location data for tracking human behaviours: A scoping review," *Sensors*, vol. 22, no. 3, p. 1220, 2022.

[4] S. Messick, "Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning," *American Psychologist*, vol. 50, no. 9, pp. 741–749, 1995.

[5] D. Borsboom, G. J. Mellenbergh, and J. Van Heerden, "The concept of validity." *Psychological review*, vol. 111, no. 4, p. 1061, 2004.

[6] G. C. Bowker and S. L. Star, *Sorting Things Out: Classification and Its Consequences*. MIT Press, 1999.

[7] A. Z. Jacobs and H. Wallach, "Measurement and fairness," in *Proc. ACM FAccT*, 2021, pp. 375–385.

[8] J. S. Larson, E. T. Bradlow, and P. S. Fader, "An exploratory look at supermarket shopping paths," *Int. Journal of Research in Marketing*, vol. 22, no. 4, pp. 395–414, 2005.

[9] S. K. Hui, E. T. Bradlow, and P. S. Fader, "Testing behavioral hypotheses using an integrated model of grocery store shopping path and purchase behavior," *Journal of Consumer Research*, vol. 36, no. 3, pp. 478–493, 2009.

[10] S. Rallapalli, A. Ganesan, K. K. Chintalapudi, V. N. Padmanabhan, and L. Qiu, "Enabling physical analytics in retail stores using smart glasses," in *Proc. ACM MobiCom*, 2014.

[11] L. Shangguan, Z. Zhou, X. Zheng, L. Yang, Y. Liu, and J. Han, "Shopminer: Mining customer shopping behavior in physical clothing stores with cots rfid devices," in *Proc. ACM SenSys*, 2015.

[12] J.-G. Lee, J. Han, and K.-Y. Whang, "Trajectory clustering: A partition-and-group framework," in *Proc. ACM SIGMOD*, 2007, pp. 593–604.

[13] S. Passi and S. Barocas, "Problem formulation and fairness," in *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, 2019, pp. 39–48.

[14] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data," in *Proc. ECCV*, 2020, pp. 683–700.

[15] G. Hoffman, T. Bhattacharjee, and S. Nikolaidis, "Inferring human intent and predicting human action in human–robot collaboration," *Annu. Rev. Control Robot. Auton. Syst.*, vol. 7, pp. 73–95, 2024.

[16] K. D. Katyal, G. D. Hager, and C.-M. Huang, "Intent-aware pedestrian prediction for adaptive crowd navigation," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2020, pp. 3277–3283.

[17] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. ICML*, 2006, pp. 369–376.

[18] G. Ding, F. Sener, and A. Yao, "Temporal action segmentation: An analysis of modern techniques," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 2, pp. 1011–1030, 2024.

[19] C. Breseman, I. Moiseev, B. Abbate, and D. M. Woollard, "Between: Boundary estimation through time warping, energy, & entropy neutralization," in *Proc. IEEE ICKG*, 2025, pp. 17–26.