

Article

A Two-Stage Diagnostic Framework for Winner-Takes-All Failures Reveals Hollow Diversity in Long-Tail Trajectory Prediction

Zimo Zhuang

School of Management Science and Engineering, Shandong University of Finance and Economics,
Jinan 250014, China; 202306140438@stu.sdufe.edu.cn

Abstract

The Winner-Takes-All (WTA) training paradigm inherently prioritises global trajectory accuracy. For long-tail samples, where terminal prediction is inherently difficult, this prioritisation leads to degraded Final Displacement Error (FDE). This asymmetry is reflected in the considerably poorer performance of FDE than Average Displacement Error (ADE) on such samples, yet its underlying mechanisms remain insufficiently diagnosed in the literature. The core contribution of this paper is a two-stage diagnostic framework that reveals two deeper asymmetries and identifies a previously unrecognised phenomenon, which we term hollow diversity. The framework is designed to be independent of specific model architectures and datasets. We first deployed the framework on the JAAD dataset using a custom-implemented baseline and subsequently validated its effectiveness on the ETH/UCY datasets.

Keywords: winner-takes-all training; hollow diversity; mode collapse; two-stage diagnostic framework; Oracle intervention; long-tail trajectory prediction; terminal error asymmetry; pedestrian trajectory forecasting

1. Introduction

1.1. Research Background

Pedestrian trajectory prediction is a key task in autonomous driving and intelligent transportation systems. In the context of autonomous driving perception, predicting pedestrian trajectories is crucial for ensuring the safety of both pedestrians and vehicle occupants. To capture the multimodal nature of future pedestrian behaviour, mainstream approaches typically employ a Mixture Density Network (MDN) to generate multiple candidate trajectories [1], trained under the Winner-Takes-All (WTA) paradigm [2]. This paradigm propagates gradients only to the candidate trajectory with the lowest overall error, which is typically quantified by the Average Displacement Error (ADE). For long-tail samples, where terminal prediction is inherently difficult, this ADE-driven selection can substantially degrade the Final Displacement Error (FDE), producing a pronounced asymmetry between the two metrics.

1.2. Mode Collapse in the WTA Training Paradigm

Numerous studies have sought to remedy the failures of the WTA training paradigm, with varying degrees of success. The concept of mode collapse was first introduced by Goodfellow et al. in the context of generative adversarial networks [3], describing the



Academic Editor: Roberto Montemanni

Received: 8 June 2026

Revised: 8 July 2026

Accepted: 13 July 2026

Published: 15 July 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

phenomenon in which the diversity of the generator's outputs is lost and multiple modes collapse into a single mode. In trajectory prediction, Makansi et al. were among the first to identify mode collapse arising from the MDN–WTA combination, where the model outputs multiple nearly identical trajectories and thus loses multimodal expressiveness. They proposed a “sampling-and-fitting” framework to overcome this limitation [4]. Rupprecht et al. identified uncertainty modelling as the essence of multimodal prediction and proposed representing ambiguity through multiple hypotheses [2]. This work laid the theoretical foundation for the widespread adoption of the WTA paradigm in multimodal prediction. Narayanan et al. improved WTA initialisation via a divide-and-conquer strategy that assigns modalities to distinct spatial regions to promote diversity [5]. Xu et al. introduced an annealing schedule that adopts a softer selection strategy early in training to mitigate WTA's greedy convergence [6]. Kreutz proposed a commonsense-based pre-training initialisation scheme to improve training stability [7]. Hugenholtz developed an evaluation framework for detecting mode collapse in multi-agent interaction settings [8]. Zhou proposed EMTA as an alternative to WTA, modelling modes as sequences to more explicitly capture inter-modal correlations, and achieved state-of-the-art performance on the Waymo Open Dataset [9]. While this work highlighted issues of limited diversity and confidence misalignment, it remained focused on remediation and did not pursue a fine-grained mechanistic diagnosis.

The above works collectively define the research landscape of mode collapse under the WTA paradigm. However, none has yet provided a mechanism-level diagnosis of the specific failures induced by this paradigm—in particular, the asymmetry between terminal and global prediction accuracy. Our work fills this gap. We establish a diagnostic framework that provides clear directions for subsequent algorithmic repair.

1.3. The Long-Tail Problem

In trajectory prediction, the long-tail problem denotes the phenomenon in which a small fraction of hard-to-predict samples contributes disproportionately to the overall prediction error. Wang et al. tackled long-tail trajectory prediction with a distribution-aware contrastive learning framework that improves the model's capacity to fit rare scenarios [10]. Li et al. proposed hierarchical contrastive learning, reducing the FDE of long-tail samples by 27% to 79% across multiple datasets [11]. While these works have made significant progress in mitigating the long-tail problem, they do not investigate the mechanisms by which the WTA training paradigm itself causes terminal prediction to fail on difficult samples. They frame the long-tail problem as a challenge stemming from imbalanced data distributions, without examining whether the inherent mechanisms of the WTA paradigm systematically mislead the model on difficult samples.

1.4. Diversity Assessment

As multimodal prediction methods have advanced, the rigorous evaluation of modal diversity has emerged as a distinct research direction. Shi et al. provided a systematic survey of the datasets, models, and evaluation metrics in multimodal trajectory prediction [12]. Weng et al. noted that conventional minADE and minFDE metrics fail to capture the quality of joint multi-agent prediction, and they proposed alternative criteria specifically designed for this setting [13]. Barajas et al. proposed a motion-semantics-oriented diversity metric that measures modal coverage from a behavioural rather than purely geometric perspective [14]. These efforts focus on designing more rigorous and effective metrics to quantify the diversity and quality of model outputs. Our focus, by contrast, is on diagnosing hollow diversity: a regime in which diversity appears adequate yet proves functionally hollow, and we aim to quantitatively expose this concealed failure via dedicated metrics.

The three metrics introduced in this paper—Modal Pairwise Distance (MPD), Endpoint Winner Gravity Ratio (G_end), and Cross-Sample Error Premium (CSEP)—are designed as specialised tools for detecting hollow diversity in this context.

1.5. Key Contributions

The asymmetry between terminal and global prediction accuracy exhibited by the WTA paradigm in long-tail scenarios fundamentally reflects a terminal prediction failure induced by the paradigm on difficult samples. We establish a two-stage diagnostic framework to provide a fine-grained attribution of this phenomenon. Our contributions are summarised as follows.

First, we propose a two-stage diagnostic framework. Applying this framework in a progressive manner reveals two deeper asymmetries: an efficacy asymmetry between the decision-making mechanism and the generation mechanism of the WTA paradigm, and a dimensional response asymmetry between longitudinal and lateral errors. The framework is designed to be independent of specific model architectures and datasets. It was deployed for the first time on the JAAD dataset with our self-implemented baseline and subsequently validated on the ETH/UCY datasets.

Second, we identify a failure mode distinct from classical mode collapse, which we term hollow diversity. It is characterised by widely dispersed trajectory endpoints that systematically deviate from the true endpoint. We visualise this phenomenon in Section 3 and analyse its underlying mechanisms in the discussion.

Finally, we close the diagnostic loop: our diagnostic results provide concrete intervention targets that can directly inform subsequent algorithmic repair efforts.

2. Materials and Methods

2.1. Problem Statement

Pedestrian trajectory prediction is typically formulated as a supervised learning problem, where a mapping function learned from past observations is used to predict future trajectories. Formally, let $X = \{x_0, x_1, \dots, x_\tau\}$ denote the sequence of pedestrian observations from the initial time to time τ . Each observation x_i is a tuple (v_i, l_0, m_i, a_i) , where v_i is visual information captured by an egocentric camera; l_0 is the initial pedestrian position given by the pixel-level bounding box in the first frame, x_0 ; m_i is pedestrian motion, computed as the displacement between the position at time i and the initial position; and a_i represents the vehicle action. The goal is to predict the pedestrian position sequence $L_\eta = \{l_{\tau+1}, l_{\tau+2}, \dots, l_{\tau+\eta}\}$ for the next η frames.

In multimodal trajectory prediction, mainstream methods typically introduce a Mixture Density Network (MDN) to output K candidate trajectories and train the model under the WTA paradigm. This paradigm propagates gradients only to the candidate trajectory that best fits the ground truth overall, where overall fit is conventionally measured by the Average Displacement Error (ADE). For long-tail samples, in which terminal prediction is inherently difficult, this ADE-driven selection can degrade the Final Displacement Error (FDE), producing an asymmetry between terminal and global prediction accuracy—ultimately amounting to terminal prediction failure.

To diagnose this failure, we logically decouple the WTA paradigm into two components. The generation mechanism refers to the forward computation process through which the model outputs K candidate trajectories. The decision mechanism refers to the process by which WTA selects what it considers the optimal trajectory from these K candidates based on its internal confidence scores. Throughout this paper, the terms “generation mechanism” and “decision mechanism” refer to these decoupled concepts. The central question of this study is whether the observed asymmetry originates from poor-quality

candidate trajectories produced by the generation mechanism or from biased selection criteria in the decision mechanism—and, further, what spatial and physical structure this bias exhibits.

2.2. Overview of the Diagnostic Framework

We decompose the analysis of the asymmetry between terminal and global prediction accuracy exhibited by the WTA paradigm in long-tail samples into two progressive stages: mechanism attribution and dimension attribution. The mechanism attribution stage determines whether the root cause lies in the decision mechanism or the generation mechanism of the WTA paradigm, while the dimension attribution stage determines whether the resulting error is concentrated in the longitudinal or lateral component.

The first-stage diagnosis is mechanism attribution. Through Oracle intervention experiments, the decision mechanism and generation mechanism of the WTA paradigm are logically decoupled to isolate the root cause of terminal prediction failure. The K predicted trajectories are then visually inspected to characterise the specific failure pattern. If anomalous patterns emerge, quantitative metrics are defined, and mechanism-level hypotheses are formulated based on structural evidence. The second-stage diagnosis is dimension attribution. Once the failure mechanism has been identified, its output error is orthogonally decomposed into longitudinal and lateral components for precise quantification.

In principle, this framework is applicable to any multimodal trajectory prediction model trained under the WTA paradigm. Its deployment on the JAAD dataset represents its first application, and its deployment on the ETH/UCY datasets constitutes the first cross-dataset validation of its diagnostic efficacy. Extending this diagnostic framework to additional datasets and more diverse model architectures is left for future work.

For datasets that do not involve vehicle–pedestrian interaction, the framework can be instantiated in a reduced form. Concretely, when no vehicle action signal is available, the vehicle action input a_i is set to the empty set, and the corresponding ActionEncoder is omitted from the architecture. The framework then operates with only the pedestrian tower, taking as input the visual features, the initial bounding box, and the pedestrian motion sequence. This single-tower instantiation is logically equivalent to the dual-tower version. The two architectures share the same diagnostic logic, and the core experiments—the Oracle intervention, the error decomposition, and the three diagnostic metrics—are applied identically. The only difference is that the vehicle-related input branch is bypassed. We explicitly describe this reduced form here so that the architectural transition in the cross-dataset validation (Section 3.7) follows naturally from the framework definition rather than appearing as an ad hoc modification.

2.3. Mechanism Attribution

The first-stage diagnosis—mechanism attribution—aims to determine whether the root cause of the WTA paradigm’s terminal prediction failure on long-tail samples lies in the generation mechanism or the decision mechanism. We designed an Oracle intervention experiment to logically decouple the generation and decision mechanisms and to identify which of the two is responsible for the failure. We then analysed the specific failure pattern of the identified mechanism and defined quantitative metrics to formulate a mechanistic hypothesis.

In the multimodal trajectory prediction task under the WTA paradigm, terminal prediction failure can originate from only two logical sources: either the generation mechanism fails to produce a sufficiently high-quality set of candidate trajectories, or the decision mechanism does not select the truly optimal trajectory from the candidate set. The Oracle intervention experiment introduces an external Oracle criterion during inference to replace

the original WTA selection rule, thereby decoupling the two mechanisms and enabling mechanism-level attribution. This experiment is performed only at the step of selecting the optimal trajectory from the K candidates; it neither modifies any model parameters nor participates in training. For any given input sample, the baseline model first completes a full forward pass, outputting K candidate prediction trajectories and their corresponding confidence scores. In the standard WTA inference process, the model directly selects the prediction trajectory with the highest confidence score as the final prediction output and uses this to calculate performance metrics. In our experiment, however, we replace the confidence-based selection mechanism with an external Oracle criterion: for the K candidate trajectories of a sample, we compute the FDE between each candidate and the ground-truth future trajectory and simply select the trajectory with the lowest FDE as the final prediction. Importantly, the Oracle criterion is applied only in the decision mechanism, i.e., the selection among the K candidates; the generation mechanism remains completely unaffected. The coordinate values, number, and geometric distribution of the K candidate position sequences in the output space are identical to those in the standard WTA inference process.

The Oracle intervention is strictly confined to the evaluation stage and does not influence training in any way. For each test sample, the baseline model performs a complete forward pass and outputs $K = 5$ candidate trajectories together with their associated confidence scores. In the standard WTA protocol, the trajectory with the highest confidence score is selected as the final prediction. Under the Oracle protocol, we instead compute the FDE between each of the K candidate endpoints and the ground-truth endpoint, and select the trajectory that yields the smallest FDE. Concretely, let e_k denote the endpoint of the k -th candidate and e^{gt} denote the ground-truth endpoint. The index of the Oracle-selected trajectory is $k' = \underset{k}{\operatorname{argmin}} \|e_k - e^{gt}\|_2$. Crucially, the set of K candidate trajectories—their coordinates, count, and geometric distribution in the output space—remains identical to that produced under standard WTA inference. No model parameters are modified; the generation mechanism is entirely untouched. This design guarantees that any discrepancy between Top-1 FDE and Oracle FDE is attributable solely to the selection criterion of the decision mechanism.

We denote the FDE under the WTA decision mechanism as Top-1 FDE, and the FDE under the Oracle criterion as Oracle FDE. Their difference is defined as $\Delta = \text{Top-1 FDE} - \text{Oracle FDE}$.

If Δ is significantly positive, it means that replacing the selection criterion yields substantially lower FDE, indicating that trajectories with high terminal prediction accuracy do exist among the candidates produced by the generation mechanism. The WTA confidence mechanism, however, mistakenly selects trajectories with poor terminal accuracy as the final output. Therefore, the terminal prediction failure should be attributed to flawed selection criteria in the decision mechanism rather than to deficient generative capacity.

If Δ approaches zero, the replacement of the original WTA criterion with the Oracle criterion yields no meaningful improvement in terminal accuracy. This would imply that no high-quality terminal prediction exists within the candidate set produced by the generation mechanism, and the failure should therefore be attributed to deficient generative capacity.

We define three quantitative metrics to further characterise the specific forms of terminal prediction failure: Modal Pairwise Distance (MPD), Endpoint Winner Gravity Ratio (G_{end}), and Cross-Sample Error Premium (CSEP).

Modal Pairwise Distance (MPD) quantifies the spatial dispersion among the endpoints of the K candidate trajectories output by the multimodal generator. For a single test sample, let the endpoint coordinates of the K predicted trajectories be denoted as $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_K \in \mathbb{R}^2$.

The modal pairwise distance for this sample is defined as the average Euclidean distance between the endpoints of all pairs of predicted trajectories, as shown in Equation (1):

$$MPD = \frac{1}{K(K-1)} \sum_{i=1}^K \sum_{j \neq i} \| \mathbf{e}_i - \mathbf{e}_j \|_2, \quad (1)$$

where $\| \cdot \|_2$ denotes the L_2 norm of a vector. A larger MPD indicates greater spatial dispersion among the endpoints of the trajectories produced by the generation mechanism.

The Endpoint Winner Gravity Ratio (G_{end}) quantifies the extent to which the Top-1 trajectory selected by the WTA decision mechanism exerts a gravitational pull on the spatial distribution of the modes produced by the generation mechanism. We define all prediction trajectories other than the Top-1 prediction trajectory as non-winning prediction trajectories. The Endpoint Winner Gravity Ratio measures, for a given test sample, whether the spatial distribution of non-winning predicted trajectories tends to cluster around the Top-1 endpoint or around the Oracle endpoint. For a single test sample, let the endpoint coordinates of the k -th predicted trajectory be denoted as $\mathbf{e}_k$. We define two key indices:

k^* : the index of the trajectory selected by the WTA decision mechanism according to the highest confidence score, i.e., $k^* = \underset{k}{\operatorname{argmax}} k$.

k' : the index of the trajectory whose endpoint is closest to the ground-truth endpoint, $\mathbf{e}^{gt}$, i.e., $k' = \underset{k}{\operatorname{argmin}} \| \mathbf{e}_k - \mathbf{e}^{gt} \|_2$. The Endpoint Winner Gravity Ratio is defined in Equation (2):

$$G_{end} = \frac{\frac{1}{K-1} \sum_{k \neq k^*} \| \mathbf{e}_k - \mathbf{e}_{k^*} \|_2}{\frac{1}{K-1} \sum_{k \neq k'} \| \mathbf{e}_k - \mathbf{e}_{k'} \|_2}, \quad (2)$$

The numerator measures the mean distance from all non-winning trajectories to the Top-1 endpoint, while the denominator measures their mean distance to the Oracle endpoint. A value of $G_{end} < 1$ therefore indicates that the non-winning trajectories are, on average, spatially closer to the WTA-selected winner than to the Oracle endpoint.

The Cross-Sample Error Premium (CSEP) quantifies the consistency and cross-scenario generalisability of the selection criteria employed by the WTA decision mechanism across different samples. To eliminate the influence of differences in absolute position and motion orientation between samples, all endpoint coordinates must be transformed into a local coordinate system before calculating the CSEP. For sample A, let the absolute coordinates of the last point in the observation sequence be $\mathbf{o}_A^{last} \in \mathbb{R}^2$, and those of the penultimate point be $\mathbf{o}_A^{prev} \in \mathbb{R}^2$. The origin of the local coordinate system is defined as $\mathbf{o}_A^{last}$, and the positive x-axis direction is the direction of the observed motion $\mathbf{d}_A = \mathbf{o}_A^{last} - \mathbf{o}_A^{prev}$. For any endpoint coordinate $\mathbf{e} \in \mathbb{R}^2$, its standardised coordinates in the local coordinate system of sample A are given by Equation (3):

$$\tilde{\mathbf{e}} = \left(\frac{(\mathbf{e} - \mathbf{o}_A^{last}) \cdot \mathbf{d}_A}{\| \mathbf{d}_A \|_2}, \frac{(\mathbf{e} - \mathbf{o}_A^{last}) \cdot \mathbf{d}_A^\perp}{\| \mathbf{d}_A \|_2} \right), \quad (3)$$

where $\mathbf{d}_A^\perp$ is the orthogonal direction vector obtained by rotating $\mathbf{d}_A$ 90° anticlockwise. Let the standardisation operator be denoted by $\mathcal{S}_A(\cdot)$, such that $\tilde{\mathbf{e}}_A^{(\cdot)} = \mathcal{S}_A(\mathbf{e}_A^{(\cdot)})$. For a pair of samples (A, B), the Cross-Sample Error Premium is defined in Equation (4):

$$CSEP(A \rightarrow B) = \| \tilde{\mathbf{e}}_A^{top1} - \tilde{\mathbf{e}}_B^{gt} \|_2 - \| \tilde{\mathbf{e}}_A^{top1} - \tilde{\mathbf{e}}_A^{gt} \|_2, \quad (4)$$

where $\tilde{\mathbf{e}}_A^{top1} = \mathcal{S}_A(\mathbf{e}_A^{top1})$ is the Top-1 predicted endpoint of sample A in A's local frame; $\tilde{\mathbf{e}}_A^{gt} = \mathcal{S}_A(\mathbf{e}_A^{gt})$ is the ground-truth endpoint of sample A in A's local frame; and

$\tilde{\mathbf{e}}_B^{gt} = \mathcal{S}_A(\mathbf{e}_B^{gt})$ is the ground-truth endpoint of sample B mapped into A's local frame. We report the average CSEP for each sample across multiple sets of random pairs and then aggregate the results by FDE quantile. A higher CSEP value indicates that the selection criteria provided by the WTA decision mechanism across different samples are more contradictory and less generalisable.

2.4. Dimension Attribution

The second-stage diagnosis—dimension attribution—aims to characterise the spatial and physical structure of the errors produced by the identified failure mechanism. For any test sample, let the coordinates of the predicted trajectory endpoint be $\mathbf{p}_{pred} \in \mathbb{R}^2$, and the true endpoint coordinates be $\mathbf{p}_{gt} \in \mathbb{R}^2$. Define the endpoint error vector for this sample as $\mathbf{e} = \mathbf{p}_{gt} - \mathbf{p}_{pred}$. The magnitude of the error vector is the FDE for this sample; its direction points from the predicted endpoint to the true endpoint, intuitively representing the orientation of the prediction bias. To decompose the error vector into longitudinal and lateral components, we take the pedestrian positions $\mathbf{p}_{\tau-1}$ and $\mathbf{p}_{\tau}$ from the last two observed frames and compute the instantaneous motion direction vector $\mathbf{d} = \mathbf{p}_{\tau} - \mathbf{p}_{\tau-1}$. Normalising $\mathbf{d}$ yields the longitudinal unit vector $\mathbf{d}_{forward}$. Rotating $\mathbf{d}_{forward}$ 90° anticlockwise yields the lateral unit vector $\mathbf{d}_{lateral}$. The projections of the error vector onto these two directions are $e_{forward} = \mathbf{e} \cdot \mathbf{d}_{forward}$ and $e_{lateral} = \mathbf{e} \cdot \mathbf{d}_{lateral}$. For convenience, we refer to $e_{forward}$ as the longitudinal error and $e_{lateral}$ as the lateral error in what follows. The longitudinal error quantifies the model's prediction bias regarding motion dynamics, such as velocity and acceleration, whilst the lateral error quantifies the model's prediction bias regarding motion geometry, such as path curvature and changes in direction. We perform a statistical analysis of the absolute error values in both directions over the entire test set, present the results visually, and analyse their physical interpretation.

2.5. Experimental Setup

We used the widely adopted JAAD benchmark dataset [15–23] to experimentally evaluate the failure mechanisms of the WTA paradigm on a baseline model. JAAD is an egocentric pedestrian trajectory prediction dataset built from in-vehicle dashcam video clips. We followed the official data split provided by the dataset for training and testing.

To further validate the diagnostic framework and its core findings across datasets, we employed ETH/UCY, the most widely used benchmark in pedestrian trajectory prediction [24,25]. It comprises five sub-scenarios across the ETH and UCY datasets, namely ETH, Hotel, Univ, Zara1 and Zara2. Unlike the ego-view crossing scenarios in JAAD, ETH/UCY features pedestrian interaction scenarios captured from a bird's-eye view. We likewise followed the official data split for training and testing.

To quantify prediction accuracy, we adopted two classic and widely used metrics: Average Displacement Error (ADE) and Final Displacement Error (FDE). ADE measures the average frame-by-frame Euclidean distance between the predicted trajectory and the ground-truth trajectory; a smaller value indicates higher global prediction accuracy. It is defined in Equation (5):

$$ADE = \frac{1}{T_f} \sum_{t=1}^{T_f} \|\mathbf{p}_{pred}^t - \mathbf{p}_{gt}^t\|_2, \quad (5)$$

where T_f is the number of prediction frames, and $\mathbf{p}_{pred}^t$ and $\mathbf{p}_{gt}^t$ are the predicted and ground-truth coordinates at frame t , respectively.

Similarly, FDE measures only the Euclidean distance at the final frame; a smaller value indicates higher terminal prediction accuracy. It is defined in Equation (6):

$$FDE = \|\mathbf{p}_{pred}^{T_f} - \mathbf{p}_{gt}^{T_f}\|_2, \quad (6)$$

where $\mathbf{p}_{pred}^{T_f}$ and $\mathbf{p}_{gt}^{T_f}$ denote the predicted and ground-truth endpoint coordinates, respectively.

To support the core diagnostic analysis, we employed three dedicated metrics: Modal Pairwise Distance (MPD), Endpoint Winner Gravity Ratio (G_end), and Cross-Sample Error Premium (CSEP), as defined in Section 2.3.

The baseline model follows the dual-tower architecture [23], which we implemented from scratch, and predicts pedestrian trajectories for the next 45 frames. The MotionEncoder in the pedestrian tower uses a Transformer encoder [26] with two layers, two attention heads, and an embedding dimension of 16. The ActionEncoder in the vehicle tower also uses a Transformer encoder with two layers, two attention heads, and an embedding dimension of 32. The initial bounding box is encoded by the LocationEncoder into a 16-dimensional feature vector. Visual features are extracted by a pre-trained R3D-18 3D convolutional network, producing 512-dimensional features, which are then projected through a fully connected layer while preserving dimensionality. The model is trained under the WTA paradigm and outputs $K = 5$ candidate trajectories. The loss function comprises three components: trajectory regression loss [27], confidence loss, and a diversity regularisation term. The model is optimised using AdamW with an initial learning rate of 2×10^{-5} and a cosine annealing schedule that decays the learning rate to a minimum of 1×10^{-6} . A linear learning rate warm-up is applied over the first 8 epochs, increasing the learning rate from zero to its initial value. The total number of training epochs is 600. The effective batch size is 128, achieved through gradient accumulation. The gradient clipping threshold is set to 1.0 to prevent gradient explosion. The weight decay coefficient is 1×10^{-4} .

All models were implemented using PyTorch (version 2.9.1, Meta AI, New York, NY, USA). Training was performed on a MacBook Pro equipped with an Apple M2 chip (Apple Inc., Cupertino, CA, USA). Data processing and statistical analyses were conducted using Python (version 3.11.14, Python Software Foundation, Wilmington, DE, USA), NumPy (version 1.24, NumPy Developers, USA), and SciPy (version 1.10, SciPy Developers, USA). All figures were generated using Matplotlib (version 3.7, Matplotlib Development Team, USA). The pre-trained R3D-18 feature extractor was sourced from the torchvision model zoo.

3. Results

3.1. Phenomenon Verification

We first verify the core phenomenon that motivates our diagnosis. The baseline model achieved an ADE of 18.24 pixels, yet its FDE reached 69.34 pixels—approximately four times the ADE—revealing a stark asymmetry between the two metrics.

To further validate the impact of long-tail samples on terminal prediction accuracy, we conducted a long-tail exclusion experiment. As shown in Table 1, the model's FDE dropped from 69.34 to 35.34 after excluding samples with FDE > 100 pixels—a reduction of approximately 49%. This confirms that a small fraction of difficult samples is responsible for a disproportionate degradation in terminal prediction accuracy.

Table 1. ADE and FDE before and after excluding long-tail samples.

Test Set Range	ADE	FDE	Sample Count	Proportion
All Samples	18.23	69.34	11,896	100%
Excluding FDE > 100	13.02	35.34	9196	77.3%

We then stratified all test samples by FDE quintiles, labelling them Q1 through Q5 from lowest to highest FDE, to examine how these difficult samples are distributed. As shown in Table 2, the intra-group FDE spans for Q1 through Q3 range from roughly ten

to twenty-something pixels, whereas the span for Q4 reaches 57.3 pixels, and that for Q5 jumps sharply to 975.2 pixels. This indicates a threshold within Q4 at which the FDE rises discontinuously. We therefore selected an integer value of 100 pixels within Q4 as the cut-off, defining samples with $FDE > 100$ pixels as long-tail samples. This subset accounts for only 22.7% of the entire test set, yet, as already demonstrated, removing it nearly halves the overall FDE.

Table 2. FDE quintile distribution.

Quintile	FDE Interval
Q1 (0–20%)	[0.1, 17.8]
Q2 (20–40%)	[17.8, 32.1]
Q3 (40–60%)	[32.1, 56.4]
Q4 (60–80%)	[56.4, 113.7]
Q5 (80–100%)	[113.7, 1088.9]

To verify the robustness of the chosen long-tail threshold, we conducted a sensitivity analysis across five FDE cut-offs: 80, 90, 100, 110, and 120 pixels. For each threshold, we report the proportion of samples classified as long-tail, the Top-1 FDE, the Oracle FDE, and the Oracle improvement ratio (Table 3). The Oracle improvement ratio remains remarkably stable, consistently within the 58–62% range across all tested thresholds. The proportion of long-tail samples decreases smoothly from 30.9% to 15.1%, in line with the heavy-tailed nature of the FDE distribution. These results confirm that our diagnostic findings are not sensitive to the specific choice of threshold, and we retain $FDE > 100$ pixels as a balanced operational definition of the long-tail group throughout the remainder of this work.

Table 3. Sensitivity of the long-tail threshold.

FDE Threshold (Pixels)	Long-Tail Proportion (%)	Top-1 FDE (Mean)	Oracle FDE (Mean)	Oracle Improvement (%)
80	30.89	148.92	58.51	60.71
90	28.81	172.24	64.62	62.48
100	22.70	199.83	77.77	61.08
110	16.03	217.71	86.98	60.04
120	15.09	236.56	98.50	58.36

For each threshold, the Oracle improvement ratio is calculated as $(\text{Top-1 FDE} - \text{Oracle FDE}) / \text{Top-1 FDE} \times 100\%$. All FDE values in the table are reported as means over the corresponding long-tail subsets.

3.2. First-Stage Diagnosis: Efficacy Asymmetry Between the Decision Mechanism and the Generation Mechanism

We conducted an Oracle intervention experiment to perform the first-stage diagnosis; the experimental design is described in Section 2. To isolate the impact of long-tail samples on terminal prediction accuracy, we adopted the same criterion used in the long-tail exclusion experiment, defining samples with $FDE > 100$ pixels as high-FDE samples and the remainder as low-FDE samples. We then quantified the results separately for the two groups.

The experimental results are shown in Figure 1. Replacing the selection criterion led to a substantial drop in FDE, and this improvement was especially pronounced in the high-FDE group, where the FDE decreased by approximately 61%. This reveals that, among the five candidate trajectories produced by the generation mechanism, a trajectory with high terminal accuracy does exist, yet the WTA decision mechanism, governed by its

own confidence-based criterion, fails to select it. Moreover, this systematic bias is further amplified when the difficulty of endpoint prediction increases in long-tail samples.

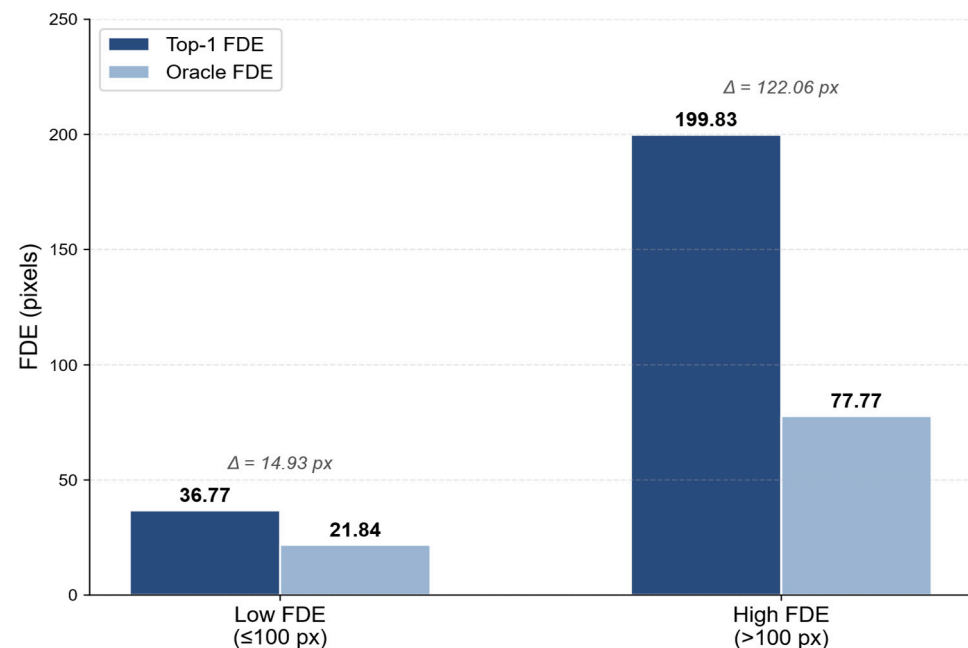


Figure 1. Oracle FDE Analysis—Quantifying Decision Failure. Performance comparison between standard WTA selection and Oracle selection for low-FDE and high-FDE samples. All FDE values are reported as mean $\pm$ SD. Low-FDE (≤ 100 px): Top-1 FDE = 36.77 ± 18.21 px, Oracle FDE = 21.84 ± 13.93 px, $\Delta = 14.93$ px. High-FDE (> 100 px): Top-1 FDE = 199.83 ± 113.74 px, Oracle FDE = 77.77 ± 52.96 px, $\Delta = 122.06$ px.

It can therefore be concluded that, in long-tail samples, the generation mechanism remains capable of producing trajectories with a certain level of terminal prediction accuracy. However, constrained by the selection criteria of the decision mechanism, these higher-accuracy trajectories are not correctly identified or selected. This reveals an efficacy asymmetry between the decision mechanism and the generation mechanism—a form of symmetry breaking: the decision mechanism’s ability to select the optimal trajectory cannot match the generation mechanism’s ability to produce it; its internal selection criteria fail to align with the legitimate outputs of the generation mechanism. This asymmetry degrades the model’s terminal prediction accuracy, and because long-tail samples are inherently more difficult than typical samples, the resulting degradation is even more pronounced.

3.3. Visualisation of Endpoint Prediction Failure: Characterisation of Hollow Diversity

To further analyse the specific form of endpoint prediction failure induced by the decision mechanism, we randomly selected one representative sample from each of the high-FDE and low-FDE groups. Rather than examining only the spatial distribution of the predicted endpoints, we visualise the full predicted trajectories alongside the ground-truth future trajectory. This allows us to observe not only where the predictions terminate but also how they evolve over the prediction horizon, providing a more complete picture of the failure mode. The results are shown in Figure 2, where the ground-truth future trajectory is drawn as a black dashed line and the five predicted trajectories are drawn as coloured dashed lines, with their endpoints highlighted by filled circles of the corresponding colours.

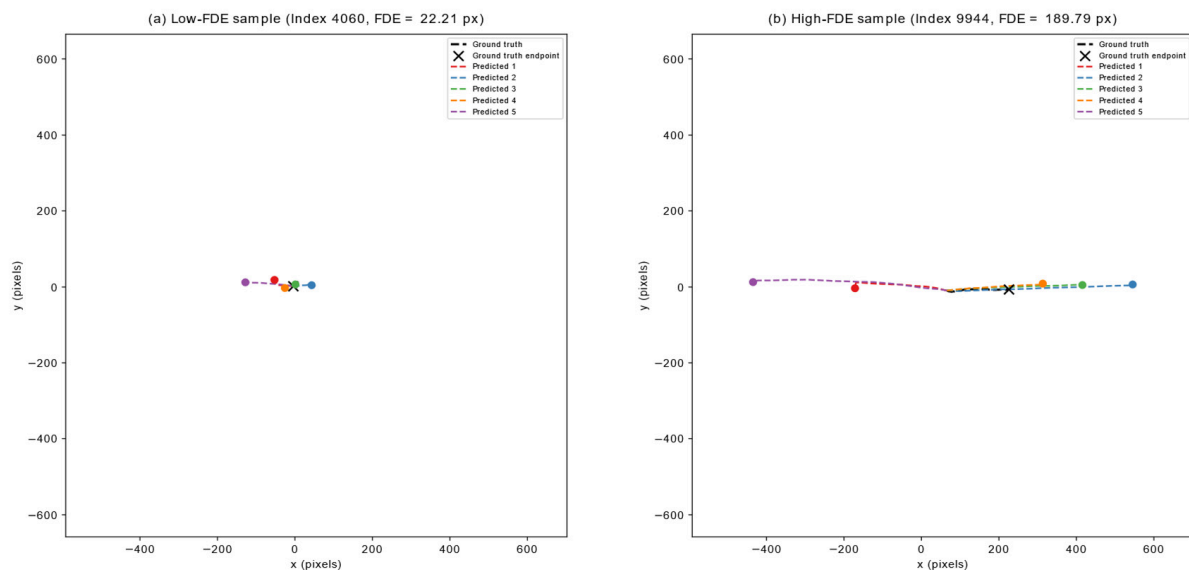


Figure 2. Visualization of hollow diversity: predicted trajectories vs. ground truth. Predicted trajectories are shown as coloured dashed lines with endpoints highlighted; the ground-truth future trajectory is shown as a black dashed line. (a) A low-FDE sample (Index 4060, FDE = 22.21 px). The predictions remain tightly clustered around the ground truth, which is partially overlaid and therefore less visible. (b) A high-FDE sample (Index 9944, FDE = 189.79 px). The predictions initially track the ground truth but progressively diverge, leaving the black dashed line clearly visible as endpoints scatter and systematically deviate from the true endpoint, characteristic of hollow diversity.

As Figure 2 shows, the two samples exhibit markedly different behaviours. In the low-FDE sample, the five predicted trajectories remain closely aligned with the ground-truth future trajectory across all frames. Their endpoints are tightly clustered, and at least one trajectory terminates very close to the true endpoint. In the high-FDE sample, by contrast, the predicted trajectories initially track the ground truth but gradually fan out as the prediction horizon extends. By the final frame, the endpoints are widely dispersed, and even the closest among them deviates substantially farther from the true endpoint than in the low-FDE case. This pattern reveals a specific form of terminal prediction failure. The generation mechanism is capable of producing diverse candidate trajectories that capture short-term dynamics adequately, yet none of them maintains accuracy through to the end. Crucially, the diversity among the predicted trajectories is preserved in appearance—the five modes are clearly separated in space—but rendered functionally meaningless because all of them systematically miss the true endpoint. This distinguishes the phenomenon from classical mode collapse, in which multiple modes collapse into a single cluster and diversity is lost entirely. We term this phenomenon hollow diversity to capture the coexistence of apparent modal variety and uniform functional failure at the trajectory terminus.

To quantitatively characterise this morphological feature, we apply the three metrics defined in Section 2.3, analysing test samples grouped by FDE quintiles.

Figure 3 compares the Modal Pairwise Distance (MPD) between the high-FDE and low-FDE groups. In the high-FDE group, the mean MPD was 359.77 ± 255.38 , and the median reached 450.18. In the low-FDE group, the mean MPD was 125.71 ± 72.26 , with a median of 133.46. A Mann–Whitney U test confirmed that MPD is significantly higher in the high-FDE group than in the low-FDE group ($U = 31,954,057.50$, $p < 0.001$). The substantially larger mean and median MPD values, together with the wider spread indicated by the standard deviation, demonstrate that in long-tail samples, the endpoints of the five predicted trajectories are not only more dispersed on average but also exhibit far greater sample-to-sample variability, consistent with the hollow diversity phenomenon.

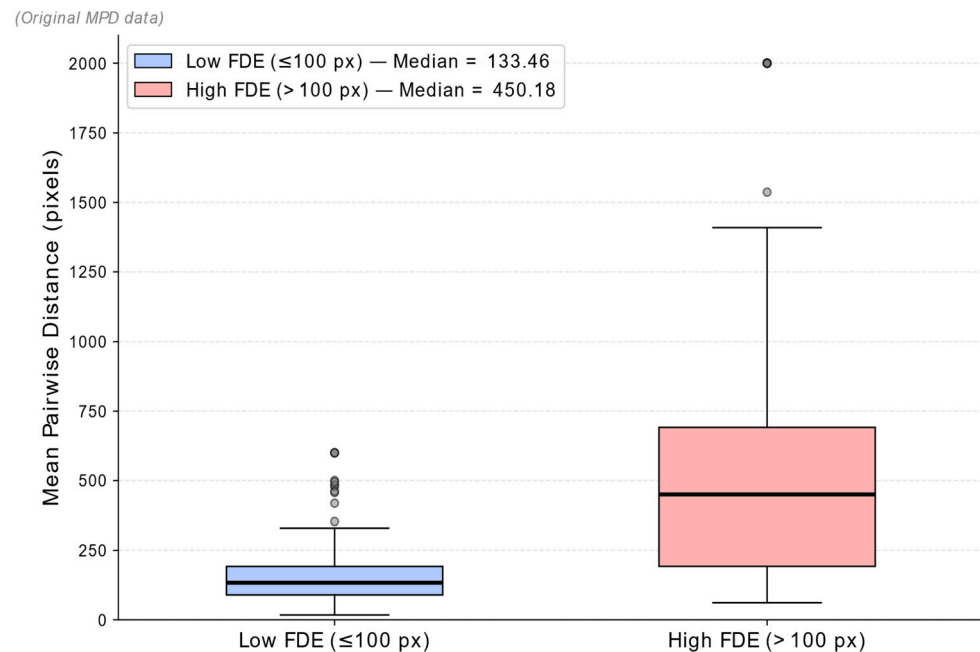


Figure 3. Comparison of Modal Diversity(MPD). Modal pairwise distance (MPD) for the low-FDE and high-FDE groups. Boxes show the median (centre line), interquartile range (IQR; box limits), and $1.5 \times$ IQR whiskers. Mean MPD: 125.71 ± 72.26 (low-FDE) vs. 359.77 ± 255.38 (high-FDE). A Mann–Whitney U test confirms a significant difference between the two groups ($p < 0.001$).

Figure 4 shows how G_{end} and CSEP vary across FDE quintiles. G_{end} remains consistently below 1 and decreases significantly with increasing FDE quintile (Spearman $\rho = -0.24$, $p < 0.001$), indicating that the gravitational pull of the Top-1 winner over the non-winning trajectories intensifies as prediction difficulty grows. Meanwhile, CSEP increases strongly with FDE quintile (Spearman $\rho = 0.83$, $p < 0.001$), confirming that the selection criteria of the WTA decision mechanism become increasingly contradictory and scenario-specific in long-tail samples.

Table 4 provides the quintile-stratified descriptive statistics for G_{end} and CSEP. G_{end} decreases from 0.999 ± 0.096 in Q1 to 0.854 ± 0.188 in Q5 (Q5 vs. Q1 t -test: $p < 0.001$), confirming that the gravitational dominance of the Top-1 winner over non-winning trajectories intensifies as prediction difficulty increases. Meanwhile, CSEP rises sharply from 96.4 ± 130.5 in Q1 to 282.4 ± 366.4 in Q5 ($p < 0.001$). The standard deviation of both metrics widens substantially across quintiles, indicating that the decision mechanism’s behaviour becomes not only more biased on average but also increasingly unstable from sample to sample in long-tail scenarios. These stratified results complement the Spearman trend tests reported in Figure 4 and provide a fuller picture of the variability underlying the aggregate trends.

Table 4. G_{end} and CSEP stratified by FDE quintile (mean $\pm$ SD). Q5 vs. Q1 t -test: G_{end} $p < 0.001$, CSEP $p < 0.001$.

Quintile	FDE Range (Pixels)	G_{end} (Mean $\pm$ SD)	CSEP (Mean $\pm$ SD)
Q1	[0.1, 17.8]	0.999 ± 0.096	96.4 ± 130.5
Q2	[17.8, 32.1]	0.969 ± 0.135	111.2 ± 141.2
Q3	[32.1, 56.4]	0.931 ± 0.183	171.3 ± 194.8
Q4	[56.4, 113.7]	0.866 ± 0.210	256.3 ± 280.9
Q5	[113.7, 1088.9]	0.854 ± 0.188	282.4 ± 366.4

Decision-Level Failure Indicators Across FDE Quintiles

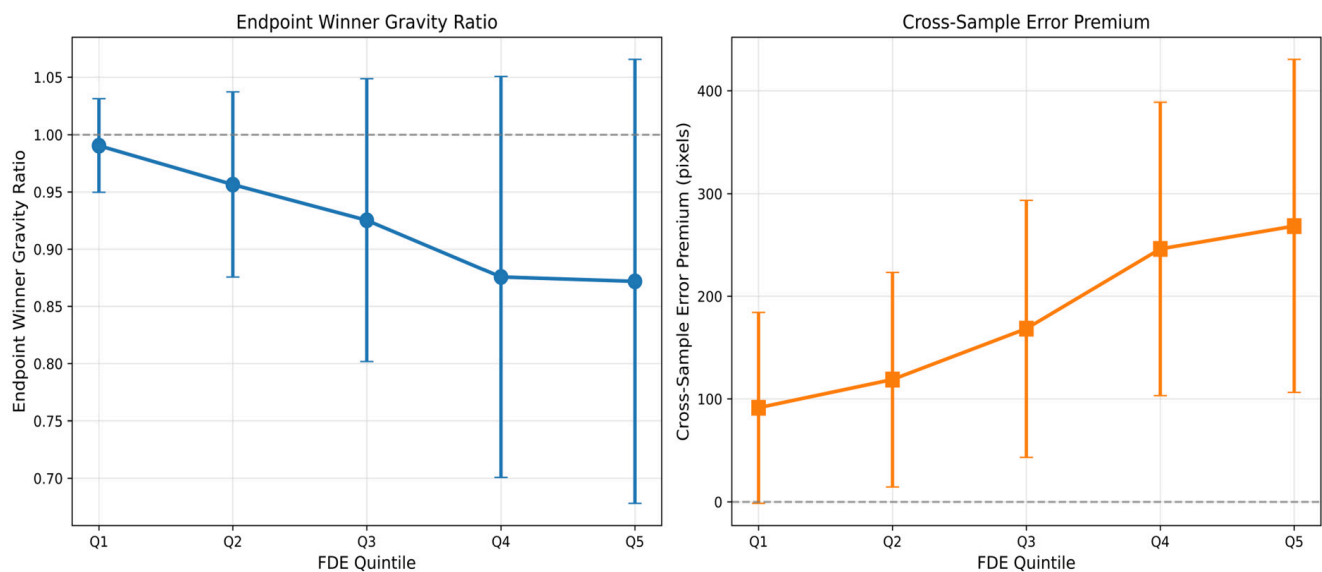


Figure 4. Trends of Endpoint Winner Gravity Ratio (G_{end}) and Cross-Sample Error Premium (CSEP) across FDE quintiles. G_{end} decreases significantly with increasing FDE quintile (Spearman $\rho = -0.24$, $p < 0.001$). The horizontal dashed line at $G_{end} = 1.00$ indicates the threshold where the gravitational pull of the Top-1 winner and the Oracle endpoint on the non-winning trajectories are equal. CSEP increases strongly with FDE quintile (Spearman $\rho = 0.83$, $p < 0.001$). The horizontal dashed line at CSEP = 0 serves as the baseline representing zero cross-sample error premium.

3.4. Second-Stage Diagnosis: Dimensional Response Asymmetry Between Longitudinal and Lateral Errors

As described in Section 2.4, we orthogonally decompose the endpoint error vector of each test sample into longitudinal and lateral components. Samples with FDE > 100 pixels are classified as the high-FDE group, following the same criterion used in the previous sections. The results are visualised through endpoint offset scatter plots and error decomposition boxplots.

Figure 5 compares, for the two orthogonal directions, the offsets between the true endpoint and the predicted endpoint selected by the WTA decision mechanism and those selected by the Oracle criterion. Each scatter point represents the error of a predicted endpoint relative to the true endpoint; the black asterisk at the origin marks the true endpoint itself, i.e., zero error. In the low-FDE group, the errors of the WTA-selected endpoints are small and distributed relatively uniformly in both the longitudinal and lateral directions, and their performance is close to that of the Oracle-selected endpoints. In the high-FDE group, however, the errors of the WTA-selected endpoints are substantially larger than those under the Oracle criterion, and they are concentrated predominantly in the longitudinal direction. Lateral errors exist but are clearly not dominant. This indicates that, in long-tail samples, the endpoint prediction bias of the WTA decision mechanism is predominantly aligned with the longitudinal dimension.

Figure 6 presents the error decomposition boxplots, providing quantitative evidence for the high-FDE group. The orange boxes represent the error between the WTA-selected endpoint and the true endpoint, while the grey boxes represent the error between the Oracle-selected endpoint and the true endpoint, with the longitudinal and lateral directions shown separately. Under the WTA decision mechanism, the longitudinal error is substantially larger than the lateral error. This confirms that the endpoint prediction failure of the decision mechanism is particularly pronounced in the longitudinal dimension and that

the error magnitudes in the two dimensions exhibit a response asymmetry, with the longitudinal dimension being disproportionately affected. Notably, even under the Oracle criterion, the prediction error still exhibits a slight longitudinal bias. This may reflect an inherent property of the JAAD dataset: as an ego-view dataset capturing pedestrian–vehicle interaction scenarios, pedestrian trajectories are primarily driven by collision avoidance rather than by free wandering, so that kinematic variability is naturally concentrated in the longitudinal dimension, making prediction inherently more difficult along that axis. Crucially, however, under the WTA decision mechanism, this longitudinal–lateral asymmetry in error responsiveness is dramatically amplified.

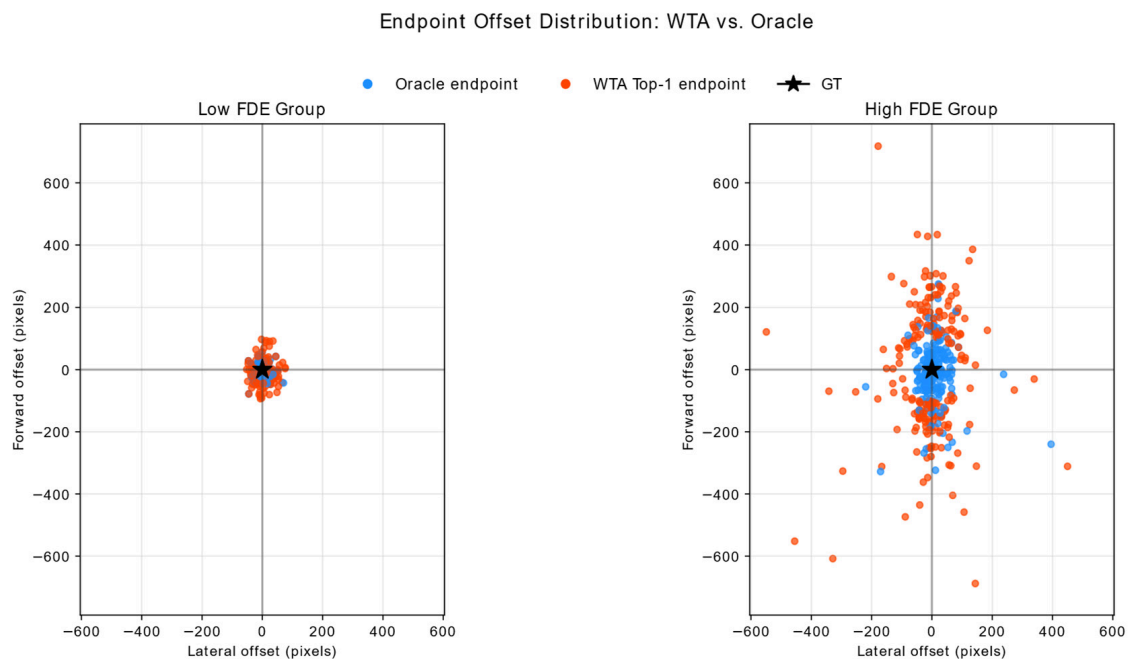


Figure 5. Scatter plot of endpoint error decomposed into longitudinal and lateral components under WTA and Oracle selection: low-FDE group and high-FDE group. The black asterisk at the origin marks the true endpoint, i.e., zero error.

To further elucidate the physical nature of the longitudinal error, we analysed the frame-by-frame velocity distributions of the trajectories selected by the WTA decision mechanism, those selected by the Oracle criterion, and the ground-truth trajectories within the high-FDE group. Figure 7 presents a comparison of the three velocity distributions. The horizontal axis represents the frame-by-frame velocity, i.e., the magnitude of the displacement vector between adjacent frames, measured in pixels per frame; the vertical axis represents the density, normalised so that the total area of the histogram equals 1. As shown in the figure, the velocity distributions of the Oracle-selected trajectories and the ground-truth trajectories largely coincide, both exhibiting a unimodal shape that is consistent with the natural rhythm of pedestrian crossing—accelerating through the crossing and then decelerating smoothly. This indicates that the generation mechanism is capable of producing candidate trajectories that are physically consistent with real pedestrian motion. In contrast, the velocity distribution of the WTA-selected trajectories differs markedly, displaying a bimodal structure. A trajectory exhibiting such a bimodal velocity profile would correspond to a motion pattern of a rapid start followed by abrupt deceleration and then a slight re-acceleration. Regardless of whether such a motion pattern is physically plausible, it is clearly inconsistent with the motion pattern of the ground-truth trajectories.

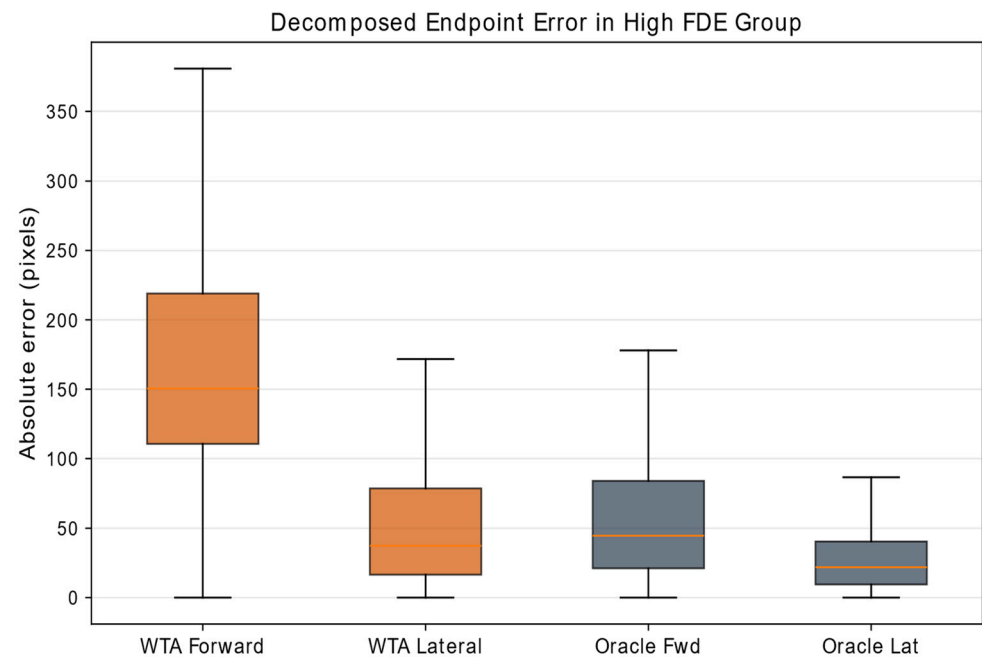


Figure 6. Box plots of longitudinal and lateral endpoint errors under WTA and Oracle selection for the high-FDE group.

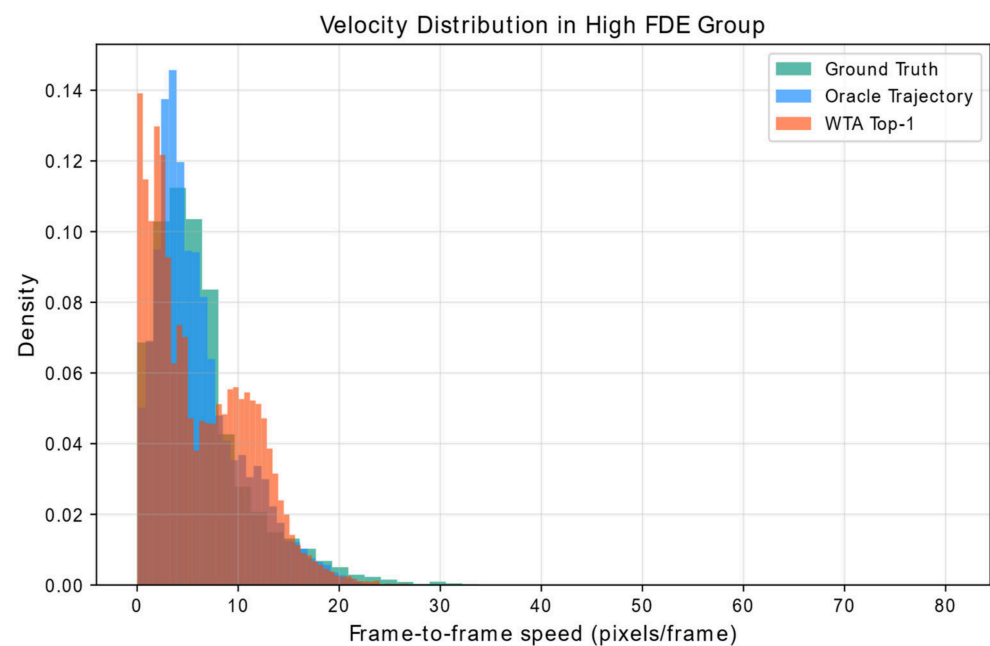


Figure 7. Per-frame speed distributions of the ground-truth trajectory, the WTA-selected trajectory, and the Oracle-selected trajectory for the high-FDE group. Speed is computed as the magnitude of the per-frame displacement vector (in pixels per frame). All per-frame speed values from all samples in the high-FDE group are pooled together. Densities are estimated using 30 equally spaced bins spanning the full range of observed speeds and are normalised so that the total area of each histogram equals 1.

3.5. Physical Verification and Structural Analysis of Decision Mechanism Bias

To verify whether the Oracle-selected trajectories are indeed closer to real pedestrian motion at the kinematic level, we compared the root mean square errors (RMSE) of velocity, acceleration, and jerk (the rate of change in acceleration) between the trajectories selected by the WTA decision mechanism and those selected by the Oracle criterion within the high-FDE group. A paired Wilcoxon test was performed.

As shown in Table 5, the Oracle-selected trajectories outperform the WTA-selected trajectories on all three metrics. The improvement is most pronounced for velocity, with a large effect size ($r = 0.827$); for acceleration and jerk, the effect sizes are moderate, but the improvements remain statistically significant. This confirms that the generation mechanism is indeed capable of producing physically reasonable trajectories, yet the decision mechanism fails to correctly identify them—an efficacy asymmetry between the two mechanisms.

Table 5. Kinematic RMSE comparison between the WTA-selected and Oracle-selected trajectories (paired Wilcoxon test). For reference, the mean $\pm$ SD values are as follows: Velocity RMSE—WTA: 5.19 ± 4.66 , Oracle: 2.14 ± 1.46 ; Acceleration RMSE—WTA: 2.55 ± 1.34 , Oracle: 1.87 ± 1.18 ; Jerk RMSE—WTA: 3.45 ± 2.01 , Oracle: 2.93 ± 1.98 .

Metric	Median WTA Error	Median Oracle Error	Z	p	Effect Size r
Velocity RMSE	4.185	2.734	43.565	<0.001	0.827
Acceleration RMSE	1.601	1.421	13.198	<0.001	0.251
Jerk RMSE	2.320	2.217	11.442	<0.001	0.217

To further examine the structural characteristics of the decision mechanism’s selection bias, we performed K-means clustering on the endpoints of the WTA-selected trajectories within the high-FDE group.

As shown in Table 6, the two cluster centres exhibit negligible lateral offset but a large longitudinal separation. This further corroborates the dimensional response asymmetry between longitudinal and lateral errors. Moreover, the two cluster centres have opposite signs and are nearly symmetric about the origin, indicating that the selection criteria of the WTA decision mechanism are internally contradictory. This provides structural support for the “training signal conflict” hypothesis: during training, the generation mechanism is highly likely to receive alternating gradient signals from two opposite directions, tearing the output space and giving rise to hollow diversity.

Table 6. K-means clustering results of WTA-selected trajectory endpoints (high-FDE samples).

Cluster	Longitudinal Centre (px)	Lateral Centre (px)	Proportion
0	−179.7	−5.1	51.50%
1	+175.6	+2.6	48.50%

3.6. Physical Verification of Difficult Samples

To confirm the practical relevance of our diagnostic findings, we must verify that the small number of long-tail samples in the dataset are genuinely difficult cases rather than artefacts of annotation noise or data corruption. We systematically examine these samples along two dimensions: physical motion characteristics and model epistemic uncertainty. Following the same FDE quintile grouping used in the previous sections, we compute, for each quintile, a set of physical motion metrics and a model uncertainty metric, capturing both the objective difficulty of the samples and the model’s own perception of this difficulty.

Figure 8 shows how three physical motion characteristics and model epistemic uncertainty vary across FDE quintiles, from the easiest samples in Q1 to the most difficult samples in Q5. Velocity variance, which captures the frame-to-frame fluctuation in walking speed, increases monotonically with FDE quintile. This indicates that pedestrians in higher-FDE samples do not maintain a steady pace but instead undergo repeated acceleration and deceleration throughout the crossing. Maximum acceleration, defined as the peak instantaneous rate of change in velocity over the prediction horizon, likewise increases

monotonically. The rising trend in maximum acceleration suggests that the most forceful speed adjustments—sudden bursts or sharp braking manoeuvres—become progressively more extreme as prediction difficulty grows. Path curvature, which quantifies how sharply the trajectory deviates from a straight line, decreases monotonically across quintiles. Rather than indicating simpler motion, this decreasing trend reveals that the difficult samples in JAAD are not characterised by complex turning or winding paths. The difficulty instead arises from pedestrians walking along relatively straight routes but with highly irregular velocity profiles. Finally, model epistemic uncertainty, estimated from the variance of the model's own predictions, also increases monotonically with FDE quintile, indicating that the model itself recognises the growing difficulty and becomes less confident in its outputs for these samples.

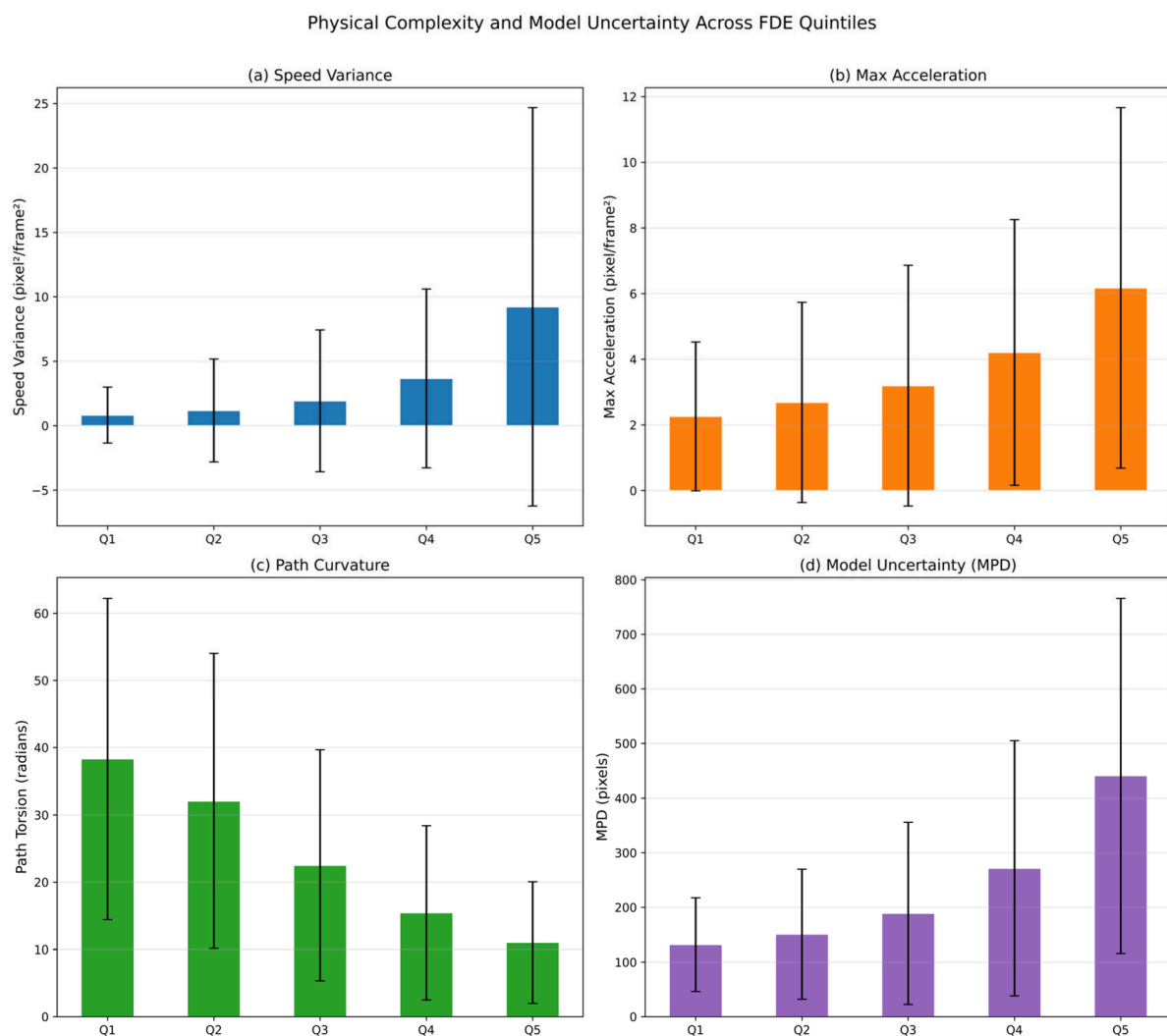


Figure 8. Trends of physical motion characteristics and model cognitive uncertainty across FDE quintiles: (a) speed variance; (b) maximum acceleration; (c) path curvature; (d) model cognitive uncertainty.

These results demonstrate that pedestrians in high-FDE samples exhibit larger velocity fluctuations, indicating less stable motion patterns and inherently greater prediction difficulty. The rate of change in velocity is also larger, showing that the dynamic characteristics are substantially more complex than those of simpler samples. Moreover, the difficult samples in the JAAD dataset arise predominantly from abrupt velocity changes during straight-line walking rather than from complex turning or curving trajectories. That is,

the difficulty originates not from the geometric complexity of the path but from the turbulence of motion rhythm. This indirectly corroborates our earlier finding that the decision mechanism's prediction error is concentrated in the longitudinal dimension rather than the lateral dimension. The consistent trends across these three physical metrics confirm that high-FDE samples are indeed kinematically more complex and constitute genuinely difficult cases, not annotation noise or corrupted data. The fact that model uncertainty also rises monotonically with FDE quintile indicates that the model is aware of the increasing prediction difficulty, which aligns with and reinforces the physical feature analysis. Taken together, these results confirm that long-tail samples differ significantly from simpler samples in terms of velocity variability and motion rhythm turbulence and that their elevated prediction difficulty has a clear physical basis. This provides a solid reliability foundation for all our diagnostic conclusions.

3.7. Cross-Dataset Validation

To validate the diagnostic efficacy of our framework, we deployed it on the ETH/UCY dataset following the single-tower instantiation described in Section 2.2. ETH/UCY differs substantially from JAAD in both camera perspective and scenario type, which avoids potential interference from dataset-specific commonalities and thereby more rigorously establishes the validity of the diagnostic findings. Moreover, since ETH/UCY does not involve vehicle–pedestrian interaction, we disabled the vehicle tower when deploying the baseline model, reducing the architecture from a dual-tower to a single-tower configuration. This also eliminates the potential confound that the diagnostic results might be specific to a dual-tower design.

After deploying the framework and the baseline model on ETH/UCY, the full-sample FDE under the WTA decision mechanism was 3.41 ± 2.78 m, compared with 0.85 ± 0.73 m under the Oracle criterion. Replacing the selection criterion thus improved terminal prediction accuracy by 75%, even exceeding the improvement observed on JAAD. This provides cross-dataset validation for the conclusion that the root cause of terminal prediction failure under the WTA paradigm lies in the biased selection criteria of the decision mechanism—i.e., an efficacy asymmetry between the generation mechanism and the decision mechanism. Furthermore, the terminal prediction error in the longitudinal dimension was 2.90 ± 2.59 m, compared with only 1.33 ± 1.15 m in the lateral dimension. This confirms that terminal prediction error is concentrated in the longitudinal dimension and supports the conclusion of a dimensional response asymmetry between longitudinal and lateral errors. In addition, the average Modal Pairwise Distance (MPD) across the full ETH/UCY sample was 6.25 ± 3.86 m, indicating substantial dispersion among modes and clearly distinguishing the phenomenon from classical mode collapse. The characterisation of hollow diversity is therefore replicated across datasets. However, the full-sample Endpoint Winner Gravity Ratio (G_{end}) was 1.01 ± 0.12 , indicating that the gravitational pull of the WTA-selected endpoint on the spatial distribution of the predicted trajectories was comparable to that of the Oracle-selected endpoint. This differs from the pattern observed on JAAD and may suggest that the causes of hollow diversity are scenario-specific.

3.8. Robustness to the Number of Output Modes

All preceding analyses were conducted with $K = 5$. To verify that the core diagnostic findings are not an artefact of this specific choice, we repeated the Oracle intervention experiment and the MPD analysis with $K = 10$ and $K = 20$. As K increases, the overall FDE distribution shifts toward lower values, and the absolute threshold of 100 pixels used in the earlier sections no longer provides a meaningful separation. We therefore report the

overall Oracle improvement and quintile-stratified results to assess whether the efficacy asymmetry and hollow diversity persist.

The overall Oracle improvement is 64.42% for $K = 10$ and 72.40% for $K = 20$, consistent with the ~61% observed at $K = 5$. In all three settings, the decision mechanism leaves a substantial margin of terminal accuracy unexploited. Tables 7 and 8 report the quintile-stratified Top-1 FDE, Oracle FDE, and MPD for $K = 10$ and $K = 20$, respectively. In both cases, the Oracle FDE is considerably lower than the Top-1 FDE in the higher quintiles, and MPD rises monotonically with prediction difficulty. For $K = 10$, the Oracle improvement reaches 74.97% in Q5, while MPD increases from 151.64 ± 110.21 in Q1 to 419.41 ± 273.13 in Q5 (Mann–Whitney U test, $p < 0.001$). For $K = 20$, the Oracle improvement in Q5 is 82.70%, with MPD rising from 179.27 ± 94.77 to 444.68 ± 275.91 ($p < 0.001$). The hollow diversity pattern is therefore replicated across all three values of K , confirming that the efficacy asymmetry is a structural feature of the WTA paradigm and not a consequence of limited modal capacity.

Table 7. Quintile-stratified Oracle intervention and MPD results for $K = 10$ (mean $\pm$ SD).

Quintile	Top-1 FDE (px)	Oracle FDE (px)	MPD (px)
Q1	11.06 ± 4.57	11.06 ± 4.57	151.64 ± 110.21
Q2	26.73 ± 4.93	21.19 ± 5.99	184.76 ± 149.22
Q3	47.00 ± 7.10	22.62 ± 13.58	230.31 ± 187.67
Q4	83.38 ± 15.67	27.17 ± 16.88	296.27 ± 230.16
Q5	210.46 ± 107.24	52.68 ± 48.01	419.41 ± 273.13

MPD Q5 vs. Q1: Mann–Whitney U = 5,107,085, $p < 0.001$.

Table 8. Quintile-stratified Oracle intervention and MPD results for $K = 20$ (mean $\pm$ SD).

Quintile	Top-1 FDE (px)	Oracle FDE (px)	MPD (px)
Q1	12.97 ± 4.42	12.93 ± 4.37	179.27 ± 94.77
Q2	27.98 ± 4.92	18.93 ± 6.56	207.31 ± 139.18
Q3	48.81 ± 7.65	20.90 ± 10.91	242.85 ± 177.33
Q4	90.52 ± 18.00	23.23 ± 12.73	319.58 ± 225.0
Q5	235.10 ± 111.33	40.66 ± 36.67	444.68 ± 275.91

MPD Q5 vs. Q1: Mann–Whitney U = 5,086,062, $p < 0.001$.

4. Discussion

4.1. Cross-Dataset Validation of the Diagnostic Framework and General Findings

Following the initial deployment of our two-stage diagnostic framework on the JAAD dataset, we conducted a cross-dataset validation of its diagnostic efficacy on ETH/UCY. We chose ETH/UCY for two main reasons. First, it is the most widely used benchmark in pedestrian trajectory prediction, and its broad acceptance helps rule out misleading results caused by low-level issues such as data corruption, noise, or annotation inconsistencies. Second, and more importantly, ETH/UCY provides a bird’s-eye view and captures only pedestrian–pedestrian interactions, without vehicle involvement. This constitutes a substantial difference in physical scenario from JAAD, reducing the risk that our findings are specific to a particular type of dataset. Moreover, because ETH/UCY lacks vehicle–pedestrian interaction, we disabled the vehicle tower during cross-dataset deployment, reducing the architecture from a dual-tower to a single-tower. This change in model structure further mitigates the concern that the diagnostic results might be tied to a specific architecture.

Under these safeguards, the conclusions drawn from JAAD are validated across scenarios on ETH/UCY. The two asymmetries revealed by the two-stage diagnosis—the efficacy asymmetry between the decision mechanism and the generation mechanism under

the WTA paradigm, and the dimensional response asymmetry between longitudinal and lateral errors—as well as the phenomenon of hollow diversity identified in this process, all receive cross-dataset support. It should be noted that our Oracle intervention experiment and physical plausibility tests confirm that the decision mechanism's selection efficacy lags behind the generation mechanism's output capability; they do not demonstrate that the trajectories produced by the generation mechanism are themselves flawless. Our work achieves the diagnostic goal of decoupling the two mechanisms to determine which one introduces the efficacy bottleneck that degrades terminal prediction accuracy. Assessing whether the generation mechanism's outputs are already optimal lies beyond the scope of this study.

4.2. Hypotheses on the Causes of Hollow Diversity and Cross-Scenario Specificity

The experimental results for the specific failure modes show that the Endpoint Winner Gravity Ratio (G_{end}) decreases monotonically with increasing FDE quintile, while the Cross-Sample Error Premium (CSEP) increases monotonically. This indicates that as prediction difficulty increases, the spatial distribution of predicted trajectories becomes increasingly concentrated around the Top-1 endpoint selected by the WTA decision mechanism rather than around the Oracle endpoint. At the same time, the selection criteria of the WTA decision mechanism are contradictory across samples, exhibiting high scenario-specificity and low cross-sample generalisability. This implies that the gravitational pull of the winner trajectory selected by the decision mechanism on the spatial distribution of predicted trajectories intensifies as prediction difficulty increases, and that the WTA decision mechanism introduces a degree of training signal conflict, which also intensifies with prediction difficulty. Furthermore, the cluster analysis of the winning predicted endpoints reveals the specific structure of this conflict: the decision mechanism's selection criteria split into two nearly opposite patterns, providing evidence for cross-sample signal conflict.

Based on this structural evidence, we propose a mechanistic hypothesis for the origin of hollow diversity. The monotonically increasing CSEP with prediction difficulty, together with the large divergence between the two cluster centres, indicates that cross-sample signal conflict during training also gradually intensifies as prediction difficulty increases, producing a direct tearing effect on the output space of the generation mechanism. Under the WTA training paradigm, the generation mechanism receives gradient supervision only from the winning trajectories of each sample, and the winner features directly represent the selection criteria of the WTA decision mechanism. When these selection criteria are highly scenario-specific and exhibit low cross-sample generalisability, the winning features of different samples become mutually contradictory. Consequently, the generation mechanism receives conflicting gradient signals, forcing the output space to split so that different modalities adapt separately to these conflicting signals. This yields dispersed modalities that are fundamentally distinct from classical mode collapse. Concurrently, G_{end} decreases monotonically with prediction difficulty, indicating that the gravitational pull exerted by the WTA-selected winner trajectory on the trajectories produced by the generation mechanism strengthens as endpoint prediction becomes more difficult. The spatial distribution of the generated trajectories thus deviates from the true trajectory and instead clusters around the local winner, which acts as a spurious gravitational centre. Together, these two effects cause the generation mechanism's output in long-tail samples to undergo modal dispersion in the wrong direction—i.e., hollow diversity—manifesting as numerically larger inter-modal distances in long-tail samples than in simpler samples, while actual endpoint prediction performance remains poor. In summary, CSEP reveals the root of the conflict, G_{end} captures the shift in the gravitational centre, and MPD quantifies the final output consequence.

It must be emphasised that the above analysis is based on structural evidence from behavioural metrics. While logically self-consistent, it does not yet constitute definitive proof. Rigorous causal confirmation requires interventional training experiments. Furthermore, the cross-dataset results on ETH/UCY indicate that the gravitational pull of the Oracle endpoint on the generated trajectory space does not fully yield to that of the Top-1 endpoint but rather competes with it. This suggests that the causes of hollow diversity are highly likely to exhibit cross-scenario specificity, with potential subtypes such as excessive gravitational dominance and gravitational competition. The precise origin of hollow diversity remains a significant open question and is left for future work.

4.3. Limitations and Future Work

Our proposed framework has so far been deployed only on the JAAD and ETH/UCY datasets. Although its diagnostic efficacy has been validated for the first time, its generalisability across additional models and datasets remains to be further assessed. We encourage future work to apply this diagnostic framework to other trajectory prediction datasets, such as nuScenes, and to a wider range of multimodal models to evaluate its broader diagnostic utility and to potentially uncover further error asymmetries and failure modes, thereby providing more meaningful targets for algorithmic repair in this field. Regarding the causes of hollow diversity, hypotheses have thus far been formulated only on the JAAD dataset; a strict causal relationship has not yet been established, and whether these findings are specific to particular scenarios remains to be determined. This is a significant and meaningful open question, and future work will pursue it accordingly. Furthermore, the Oracle intervention employed in this study relies on ground-truth information and is thus purely diagnostic. A natural extension of this work is to examine whether the structural evidence uncovered here can inform a learnable confidence module that better aligns the decision mechanism with the generation mechanism, thereby closing the efficacy gap identified in our diagnosis.

We also wish to clarify that the primary goal of this work is diagnosis rather than the immediate proposal of a new algorithm. The absence of a concrete remedy or interventional causal proof is a deliberate scoping choice, not a shortcoming. A precise diagnosis is a prerequisite for effective repair. Without knowing whether the bottleneck lies in the generation mechanism or the decision mechanism, any attempt at algorithmic intervention risks being misdirected. Our framework narrows down the failure to the decision mechanism and, within it, to a misalignment between confidence scores and endpoint accuracy. This provides a well-defined target for subsequent work, which can now focus on aligning the selection criterion with terminal prediction quality instead of globally retraining the entire model. We believe that such a targeted direction, grounded in mechanistic evidence, offers a more efficient path to mitigation than undirected exploration. The interventional training experiments required to establish causality are a natural next step, and the diagnostic foundation laid here is intended to make those experiments both more meaningful and more interpretable.

5. Conclusions

This paper addresses the problem of terminal prediction failure under the WTA training paradigm in multimodal trajectory prediction, focusing on long-tail samples. We established a two-stage diagnostic framework, which was first applied to the JAAD dataset with a self-implemented baseline and subsequently validated on the ETH/UCY dataset. In the process, we identified and named the phenomenon of hollow diversity. The first-stage diagnosis, through Oracle intervention experiments, precisely attributes the terminal prediction failure to selection bias in the decision mechanism. During the analysis of the

specific failure mechanism, this previously unrecognised failure mode was characterised, and a mechanistic hypothesis was proposed based on structural evidence. The second-stage diagnosis, employing orthogonal error decomposition and velocity distribution analysis, reveals the spatial and physical structure of the decision mechanism's failure. This further confirms that the decision mechanism's failure is fundamentally a systematic misalignment between its confidence scores and the true physical state of the trajectory. By precisely attributing the failure to a specific component of the training paradigm and characterising its spatial structure, the proposed framework demonstrates that rigorous diagnosis, independent of immediate algorithmic repair, constitutes a meaningful and necessary contribution to the understanding of WTA-induced prediction failures.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are derived from publicly available benchmark datasets (JAAD and ETH/UCY). No new data were created in this study.

Acknowledgments: During the preparation of this manuscript, the author used DeepLTranslator and AI-assisted language editing tools for the purposes of translation and language polishing. The author have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Bishop, C.M. *Mixture Density Networks*; Technical Report NCRG/94/004; Aston University: Birmingham, UK, 1994.
2. Rupprecht, C.; Laina, I.; DiPietro, R.; Baust, M.; Tombari, F.; Navab, N.; Hager, G.D. Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 3591–3600.
3. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the 27th Conference on Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
4. Makansi, O.; Ilg, E.; Cicek, O.; Brox, T. Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 7144–7153.
5. Narayanan, S.; Moslemi, R.; Pittaluga, F.; Liu, B.; Chandraker, M. Divide-and-conquer for lane-aware diverse trajectory prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 15779–15788.
6. Xu, Y.; Letzelter, V.; Chen, M.; Zablocki, É.; Cord, M. Annealed winner-takes-all for motion forecasting. *arXiv* **2024**, arXiv:2409.11172.
7. Kreutz, T.; Mühlhäuser, M.; Guinea, A.S. Common sense initialization of mixture density networks for motion planning with overestimated number of components. In Proceedings of the 2nd Tiny Papers Track at the 12th International Conference on Learning Representations (ICLR), Vienna, Austria, 11 May 2024.
8. Hugenholtz, M.D. Mode Collapse Happens: Evaluating Critical Interactions in Joint Trajectory Prediction Models. Master's Thesis, Delft University of Technology, Delft, The Netherlands, 2024.
9. Zhou, Z.; Zhou, H.; Hu, H.; Wen, Z.; Wang, J.; Li, Y.-H.; Huang, Y.-K. ModeSeq: Taming sparse multimodal motion prediction with sequential mode modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–17 June 2025; pp. 1612–1621.
10. Wang, Y.; Zhang, P.; Bai, L.; Xue, J. FEND: A future enhanced distribution-aware contrastive learning framework for long-tail trajectory prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 1400–1409.
11. Li, Z.; Wang, X.; Liu, Y. Hi-SCL: Fighting long-tailed challenges in trajectory prediction with hierarchical wave-semantic contrastive learning. *Transp. Res. Part C Emerg. Technol.* **2024**, *159*, 104486. [[CrossRef](#)]
12. Shi, S.; Jiang, L.; Dai, D.; Schiele, B. Multimodal trajectory prediction: A survey. *arXiv* **2023**, arXiv:2302.10463.
13. Weng, E.; Hoshino, H.; Ramanan, D.; Kitani, K. Joint metrics matter: A better standard for trajectory forecasting. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 9876–9886.

14. Barajas, C.; Giampieri, G.; Sabatini, S.; Poerio, N. Addressing mode collapse in trajectory prediction: A maneuver-oriented metric and approach. In Proceedings of the IEEE Intelligent Vehicles Symposium (IV), Jeju Island, Republic of Korea, 2–5 June 2024; pp. 3363–3370.
15. Rasouli, A.; Kotseruba, I.; Tsotsos, J.K. Are they going to cross? A benchmark dataset and baseline for pedestrian cross-walk behavior. In Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 22–29 October 2017; pp. 206–213.
16. Yao, Y.; Xu, M.; Choi, C.; Crandall, D.J.; Atkins, E.M.; Dariush, B. Egocentric vision-based future vehicle localization for intelligent driving assistance systems. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 9711–9717.
17. Yagi, T.; Mangalam, K.; Yonetani, R.; Sato, Y. Future person localization in first-person videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7593–7602.
18. Bhattacharyya, A.; Fritz, M.; Schiele, B. Long-term on-board prediction of people in traffic scenes under uncertainty. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 4194–4202.
19. Rasouli, A.; Kotseruba, I.; Kunic, T.; Tsotsos, J.K. PIE: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6262–6271.
20. Neumann, L.; Vedaldi, A. Pedestrian and ego-vehicle trajectory prediction from monocular camera. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021; pp. 10199–10207.
21. Yin, Z.; Liu, R.; Xiong, Z.; Yuan, Z. Multimodal transformer networks for pedestrian trajectory prediction. In Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI), Montreal, QC, Canada, 19–27 August 2021; pp. 1259–1265.
22. Rasouli, A.; Rohani, M.; Luo, J. Bifold and semantic reasoning for pedestrian behavior prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 15580–15590.
23. Zhang, Z.; Ding, Z.; Tian, R. Decouple ego-view motions for predicting pedestrian trajectory and intention. *IEEE Trans. Image Process.* **2024**, *33*, 4716–4728. [[CrossRef](#)] [[PubMed](#)]
24. Pellegrini, S.; Ess, A.; Schindler, K.; Van Gool, L. You’ll never walk alone: Modeling social behavior for multi-target tracking. In Proceedings of the IEEE 12th International Conference on Computer Vision (ICCV), Kyoto, Japan, 27 September–4 October 2009; pp. 261–268.
25. Lerner, A.; Chrysanthou, Y.; Lischinski, D. Crowds by Example. *Comput. Graph. Forum* **2007**, *26*, 655–664. [[CrossRef](#)]
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; Volume 30, pp. 5998–6008.
27. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.